UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

LUCAS HOLTZ WAMSER

EXPLORANDO ABORDAGENS PARA LEGENDAGEM AUTOMÁTICA DE IMAGENS PARA A IDENTIFICAÇÃO DO SUJEITO PRINCIPAL EM FOTOGRAFIAS DE EVENTOS DE FORMATURAS

DISSERTAÇÃO

CURITIBA
2023

LUCAS HOLTZ WAMSER

EXPLORANDO ABORDAGENS PARA LEGENDAGEM AUTOMÁTICA DE IMAGENS PARA A IDENTIFICAÇÃO DO SUJEITO PRINCIPAL EM FOTOGRAFIAS DE EVENTOS DE FORMATURAS

Exploring Automatic Image Captioning Approaches for Main Subject Identification in Graduation Event Photos

Dissertação apresentada como requisito para obtenção do grau de Mestre em Computação Aplicada da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador: Prof. Dr. Bogdan Tomoyuki Nassu

CURITIBA

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Curitiba

LUCAS HOLTZ WAMSER

EXPLORANDO ABORDAGENS PARA LEGENDAGEM AUTOMÁTICA DE IMAGENS PARA A IDENTIFICAÇÃO DO SUJEITO PRINCIPAL EM FOTOGRAFIAS DE EVENTOS DE FORMATURAS

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Computação Aplicada da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia De Sistemas Computacionais.

Data de aprovação: 28 de Novembro de 2023

Dr. Bogdan Tomoyuki Nassu, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Pedro Luis Kantek Garcia Navarro, Doutorado - Universidade Federal do Paraná (Ufpr)

Dr. Ricardo Dutra Da Silva, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 28/11/2023.

RESUMO

HOLTZ WAMSER, Lucas. **Explorando Abordagens para Legendagem Automática de Imagens para a Identificação do Sujeito Principal em Fotografias de Eventos de Formaturas**. 2023. 83 f. Dissertação (Mestrado em Computação Aplicada) – Universidade Tecnológica Federal do Paraná. Curitiba, 2023.

A seleção de fotografias de um evento de formatura é uma tarefa essencial para empresas que organizam tais eventos, pois a venda de álbuns personalizados é parte importante da sua arrecadação. Tendo isso em mente, o presente trabalho se concentra na aplicação de redes neurais pré-treinadas para auxiliar a organização de álbuns de formatura, explorando a legendagem automática de imagens e a identificação do sujeito principal. O estudo começa com uma análise comparativa de três abordagens baseadas em *deep learning* para a legendagem automática de imagens no contexto de formaturas. O modelo One For All (OFA), baseado em *transformers*, destaca-se como uma escolha promissora. O OFA é pré-treinado em uma ampla variedade de dados, e foi especializado para a tarefa específica de legendagem de imagens de formatura. Além disso, o trabalho investiga a utilização do conhecimento implícito nos modelos de legendagem automática para identificar o sujeito principal em uma imagem. Isso é essencial para a organização eficaz de álbuns de formatura, onde é crucial destacar os principais protagonistas. O modelo OFA foi especializado para gerar caixas delimitadoras para esta tarefa, o que resultou em melhorias notáveis, com medidas de *Intersection over Union* médias de 0,47, em comparação com 0,17 sem especialização. Também exploramos a utilização das legendas geradas pelo modelo para a geração de uma nuvem de palavras, a qual pode ser útil para a filtragem das fotografias. As contribuições deste trabalho são diretamente relevantes para a organização de álbuns de formatura, incluindo a geração de legendas e caixas delimitadoras para o sujeito principal em fotografias, bem como a criação de nuvens de palavras para a organização eficiente dos álbuns. Em síntese, este estudo destaca a eficácia das redes neurais pré-treinadas na legendagem automática de imagens e na identificação do sujeito principal, proporcionando benefícios significativos na automatização da separação de álbuns de formatura, uma tarefa valiosa para as empresas e formandos.

Palavras-chave: Agrupamento de Imagens. Legendagem Automática de Imagens. Aprendizado Profundo. Redes Transformer. Identificação de Sujeito Principal.

ABSTRACT

HOLTZ WAMSER, Lucas. **Exploring Automatic Image Captioning Approaches for Main Subject Identification in Graduation Event Photos**. 2023. 83 p. Dissertation (Master's Degree in Applied Computing) – Universidade Tecnológica Federal do Paraná. Curitiba, 2023.

The selection of photographs from a graduation event is an essential task for companies that organize such events, as the sale of personalized albums is a significant part of their revenue. With that in mind, this work focuses on the application of pre-trained neural networks to assist in organizing graduation albums, exploring techniques for automatic image captioning and main subject identification. The study begins with a comparative analysis of three deep learning-based approaches to automatic image captioning in the context of graduations. The One For All (OFA) model, based on transformers, stands out as a promising choice. OFA is pre-trained on a wide variety of data and has been specialized for the specific task of captioning graduation images. Additionally, the work investigates the use of the implicit knowledge in automatic captioning models to identify the main subject in an image. This is essential for the effective organization of graduation albums, where it is crucial to highlight the main protagonists. The OFA model has been specialized to generate bounding boxes for this task, resulting in notable improvements, with average Intersection over Union measures of 0.47, compared to 0.17 without specialization. We also explore the use of the captions generated by the model to generate a word cloud, which can be useful for filtering photographs. The contributions of this work are directly relevant to the organization of graduation albums, including the generation of captions and bounding boxes for the main subject in photographs, as well as the creation of word clouds for efficient album organization. In summary, this study highlights the effectiveness of pre-trained neural networks in automatic image captioning and subject identification, providing significant benefits in automating the organization of graduation albums, a valuable task for companies and graduates.

Keywords: Image Clustering. Automatic Image Captioning. Deep Learning. Transformer Networks. Main Subject Identification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Quando olhamos uma imagem qualquer, nosso cérebro consegue de maneira automática contextualizá-la. Por exemplo, na Figura 1, que pode ser descrita por: “Um homem com uma beca de formatura entregando um diploma para uma estudante”. Não conseguimos perceber, mas isso é uma tarefa complexa, ainda mais quando tratamos de generalizar a contextualização, ou seja, analisar imagens de diferentes categorias e conseguir descrevê-las ou rotulá-las. Vale ressaltar que um dos primeiros passos é identificar os sujeitos principais da cena, o que pode ser algo subjetivo. Poder utilizar uma rede neural artificial para descrever imagens em linguagem natural de forma automática é um dos avanços na área de inteligência artificial dos últimos tempos (HSU *et al.*, 2021; AUNG *et al.*, 2020; LI *et al.*, 2020b). Um modelo computacional que possa descrever uma fotografia corretamente pode ter várias utilizações práticas, como criar descrições de imagem para pessoas cegas, ou o agrupamento de imagens por contextos parecidos, ou identificar os elementos que tenham maior destaque ou importância.

Figura 1 – Exemplo de descrição de uma imagem: Um homem com uma beca de formatura entregando um diploma para uma aluna [Fonte: Dataset Used].

A criação de álbuns de fotografias de formatura é uma tarefa que atualmente é realizada por humanos, que precisam observar o contexto e o conteúdo de milhares de fotografias para preparar álbuns personalizados. Embora esse trabalho seja árduo e consuma muito tempo, é

importante destacar que ele é extremamente valioso para os formandos e suas famílias, que buscam manter registros daqueles momentos. A automatização de parte desta tarefa pode ser interessante para acelerar o trabalho e reduzir custos. Entretanto, em testes iniciais, foi observado que simplesmente tentar criar álbuns de forma totalmente automática se mostrou desafiador, entre outras razões pela grande quantidade de cenários e pessoas em cada imagem, o que dificulta, por exemplo, o uso de técnicas para identificação automática de rostos. Por outro lado, a primeira etapa desse processo: uma pré-filtragem mais simples de fotografias, já seria útil para empresas que organizam eventos de formatura — isto foi constatado em conversas com profissionais do ramo. Uma das formas de se realizar esta pré-filtragem seria através da identificação das pessoas mais em destaque na fotografia, o "sujeito principal", o que ajudaria a direcionar etapas posteriores na criação dos álbuns.

Em explorações iniciais que visavam encontrar outras formas de se separar as fotografias, consideramos a possibilidade de selecioná-las de acordo com o contexto de cada uma: colação de grau, festa, churrasco, etc. Entretanto, nesses testes iniciais, observamos que a atribuição de um rótulo único a cada fotografia era uma abordagem excessivamente limitada, incapaz de agrupar a imagens de uma maneira que se mostrasse útil para os usuários humanos, isso por conta da grande quantidade de rótulos que seriam necessários para esse processo e a amostragem limitada para eles. Passamos então a considerar a identificação de sujeitos principais. Entretanto, as técnicas encontradas na literatura se mostraram ineficazes para o cenário proposto, visto que se baseiam na localização de saliências visuais, enquanto o problema que abordamos exigiria uma interpretação de mais alto nível da imagem. Partimos então da hipótese de que o conhecimento necessário para a identificação do sujeito principal em uma imagem pode estar implicitamente codificado em modelos capazes de realizar a legendagem automática de fotografias, já que estas abordagens precisam identificar quais são os elementos que serão descritos. Também em conversas com profissionais do ramo, mencionou-se que a construção de uma "nuvem de palavras" (como aquela mostrada na Figura 2) seria uma ferramenta de grande valor para eles, pois, por exemplo, se uma imagem contém palavras-chave como "vestido", "gravata" e "formatura", "grupo de amigos", isso sugere que a imagem pode ser relevante para um álbum de formatura. Embora a identificação do sujeito principal e a construção de nuvens de palavras sejam problemas distintos, identificamos que é possível resolver ambos os problemas a partir de redes treinadas para legendagem automática, pois esta tarefa pressupõe tanto a geração de um texto contendo palavras quanto a identificação de quais elementos serão descritos.

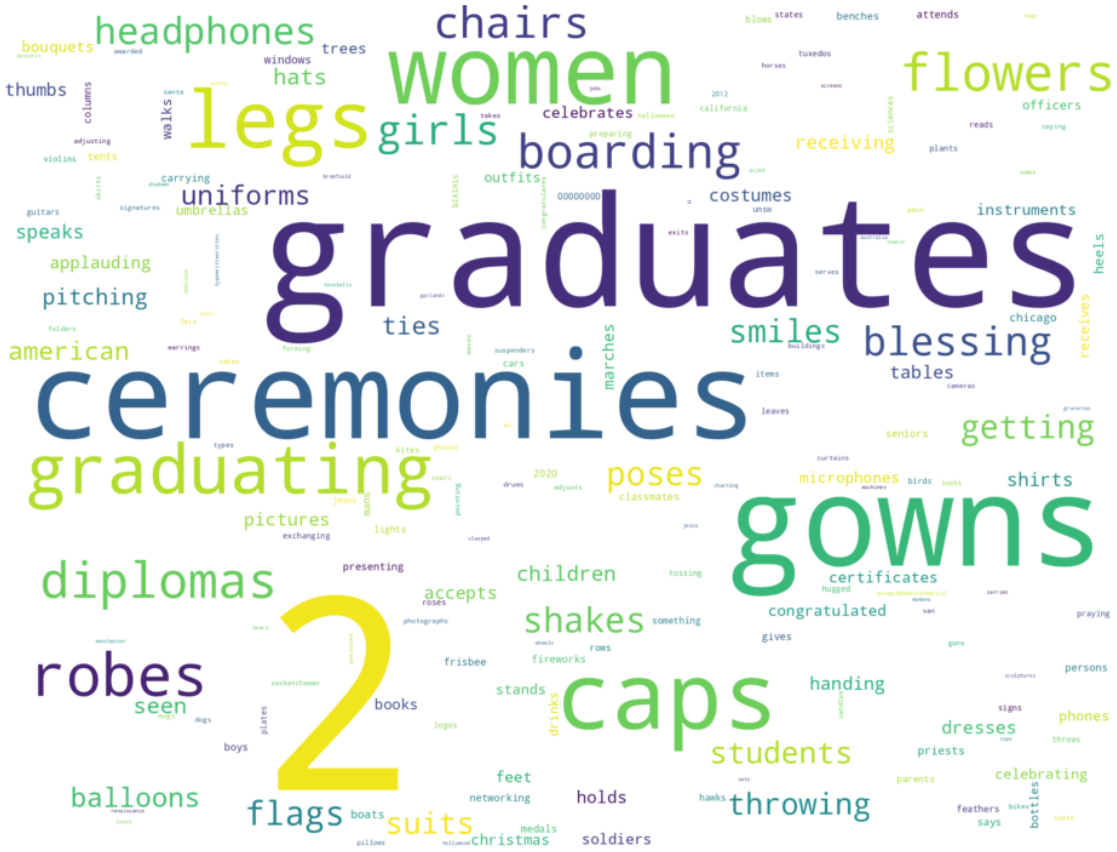

Figura 2 – Exemplo de nuvem a qual seria util para a filtragem de fotos de formaturas [Fonte: Autoria Própria

Neste trabalho, avaliamos a possibilidade de se utilizar uma mesma arquitetura de rede neural pré-treinada para legendagem automática de imagens tanto para a produção de nuvens de palavras quanto para, após um estágio de especialização, a identificação do sujeito principal da imagem. Em particular, buscamos inicialmente alternativas que fossem capazes de produzir boas descrições para imagens de formatura, e que pudessem posteriormente ser exploradas para a identificação do sujeito principal, por exemplo, extraindo informações de uma camada de atenção.

Referente às abordagens de legendagem automática, destacamos técnicas que atingem bons resultados em *datasets* como COCO (*Common Objects in COntext*) (LIN *et al.*, 2015) e ImageNet (DENG *et al.*, 2009), *datasets* publicamente disponíveis e que são amplamente utilizados para testar e avaliar modelos para anotação automática de imagens. Entre as arquiteturas que produzem os melhores resultados nestes testes padronizados, encontram-se aquelas baseadas em *transformers*, mais especificamente a rede One For All (OFA), a qual é uma arquitetura complexa que trata não somente da tarefa de legendagem automática de imagens, mas também inclui outras tarefas como *visual grounding*, *grounded captioning*, preenchimento automático de

regiões e texto, e geração de imagens. O OFA é pré-treinado sobre uma grande quantidade de imagens e textos de diferentes datasets, com posterior especialização (*fine tuning*) para tarefas específicas. Uma versão do OFA foi testada na tarefa de legendagem automática sobre um dataset contendo diversas imagens de eventos e obteve um desempenho melhor do que duas outras soluções testadas, ambas baseadas em LSTM/GRU, com uma delas utilizando diretamente um mecanismo de atenção. Um modelo OFA pré-treinado foi especializado para o domínio de fotografias de formaturas e utilizado para gerar legendas que servem de base para a criação de nuvens de palavras. O mesmo modelo foi então especializado para produzir a "caixa delimitadora" (*boundig box*) do sujeito principal que foi descrito, obtendo a média de *Intersection over Union* (IoU) de 0,47 – utilizando como base um dataset de 80 Imagens sendo 30 delas o conjunto de validação. Embora algumas imagens ainda tenham tido resultados abaixo de 0,5 — valor considerado em alguns desafios como sendo indicativo de uma detecção bem-sucedida — trata-se de uma melhora considerável se comparada ao valor obtido pela rede sem a especialização, que foi 0,17.

As principais contribuições deste trabalho são:

- Uma comparação do desempenho de três alternativas baseadas em *deep learning* para legendagem automática de imagens.
- Criação de legendas e caixas delimitadoras para o sujeito principal sobre imagens do dataset USED, que serão disponibilizadas para outros trabalhos futuros.
- A geração de nuvens de palavras para a organização de álbuns de fotografais de formatura com base nas legendas produzidas automaticamente por uma rede neural especializada para este cenário.
- Demonstração de que é possível explorar o conhecimento implícito de uma rede para legendagem automática de imagens para identificar o sujeito principal de uma fotografia.

O restante desta dissertação é dividido da seguinte forma: o Capítulo 2 apresenta trabalhos relacionados à identificação do sujeito principal e à legendagem automática de imagens. O Capítulo 3 apresenta os materiais e métodos utilizados no trabalho, incluindo os datasets, as técnicas para legendagem automática empregadas, as abordagens para identificação do sujeito principal e criação de nuvens de palavras, e por fim as métricas que serão utilizadas para avaliar

o desempenho dessas técnicas. O Capítulo 4 descreve os experimentos, resultados obtidos e a discussão dos resultados. O Capítulo 5 contém as conclusões e trabalhos futuros.

2 REVISÃO DA LITERATURA

A descrição de imagens é uma das principais capacidades da visão computacional, podendo ser aplicada em uma gama de serviços, pois permite que o usuário possa acessar informações importantes em uma imagem. Para que uma imagem possa ser descrita de forma precisa, é necessário compreender a relação entre os objetos e as ações que ela apresenta, e pode ser útil sumarizar e apresentar esta informação em linguagem natural. Por isso, neste trabalho consideramos a hipótese de que o conhecimento necessário para que um modelo descreva uma cena em linguagem natural também envolve implicitamente a identificação do sujeito principal em uma fotografia contendo pessoas.

O trabalho de Gupta e Davis (2008) buscam uma maneira de diminuir a ambiguidade durante o treinamento de classificadores para reconhecimento de objetos baseados em dados fracamente rotulados. O trabalho aponta que manter a relação provável entre objetos e a cena na qual eles estão inseridos melhora tanto a anotação feita por humanos quanto a inferência de modelos treinados. Já Li e Fei-Fei (2007) mostram que o reconhecimento de eventos é aprimorado por inferência em um modelo generativo que relaciona a cena em que o evento ocorre com os objetos na imagem. Os mesmos autores usam o fato de que objetos e poses humanas são acoplados, e mostram que reconhecer uma ajuda no reconhecimento da outra (LI *et al.*, 2009). Nesses trabalhos é demonstrado que todos os objetos de uma cena estão relacionados, e trazer essas relações para as anotações e para os modelos de inferência faz com que o contexto da imagem ajude a reconhecer os objetos nela contidos efetivamente. Por exemplo, um surfista surfando no mar: reconhecendo o mar com uma onda, é possível que haja um surfista lá, usando como base as imagens de treinamento. Desta forma, foram propostos diversos trabalhos que buscam estabelecer as relações entre elementos de uma cena, descrevendo-os em linguagem natural.

Neste capítulo, comentamos na Seção 2.1 sobre trabalhos que visam a identificação do sujeito principal em uma cena. Em seguida, serão discutidos os principais trabalhos da literatura que abordam o tema da legendagem automática de fotografias em um contexto geral, visto que a abordagem que propomos para a identificação do sujeito principal se baseia em técnicas criadas para legendagem automática. Na Seção 2.2 são discutidos métodos mais simples, que se baseiam em estruturas de frase fixas. Métodos baseados em redes convolucionais e redes recorrentes são descritos na Seção 2.3. Na Seção 2.4, são apresentados métodos baseados em redes com

mecanismos de atenção e *transformers*.

2.1 MÉTODOS PARA IDENTIFICAÇÃO DO SUJEITO PRINCIPAL

A detecção do sujeito principal em imagens é uma questão fundamental na análise e interpretação de conteúdo visual. Várias abordagens têm sido investigadas para automatizar esse processo e compreender melhor os elementos de destaque em uma imagem.

Uma abordagem inicial para identificar o assunto principal se baseia na premissa de que as partes menos importantes de uma imagem tendem a ser menos nítidas e possuir menos detalhes de borda. O trabalho de Vaioulis *et al.* (2005) adota essa ideia e emprega a detecção de bordas para identificar áreas mais nítidas na imagem, presumindo que essas regiões representem o assunto principal. Embora sensível à qualidade da imagem e à intensidade do desfoque, essa técnica oferece uma abordagem direta e intuitiva para detectar elementos relevantes. Por outro lado, a associação entre quantidade de bordas e importância dos objetos nem sempre é verificada na prática.

Quando o assunto principal envolve pessoas, a detecção da posição de partes do corpo, como tronco e, principalmente, a cabeça, pode ser crucial para determinar o conteúdo relevante da imagem. Por exemplo, o trabalho de Shen *et al.* (2009) utiliza o clássico detector Viola-Jones (VIOLA; JONES, 2001) para identificar as regiões da cabeça e do tronco nas imagens. O objetivo central desse método é identificar a presença de linhas que possam "cortar" a cabeça das pessoas na imagem, uma prática fotográfica indesejada. Embora aplicado a um contexto específico, esse método demonstra a viabilidade de utilizar características anatômicas para discernir o sujeito principal. Embora a detecção de partes do corpo, como cabeça e tronco, seja valiosa para identificar o sujeito principal em imagens, isso não é sempre infalível. O contexto cultural, poses incomuns e ambiguidades podem limitar sua eficácia. Além disso, avanços recentes em algoritmos, como CNNs, oferecem melhor desempenho.

Uma abordagem mais avançada para a detecção do assunto principal baseia-se na identificação de saliências visuais, ou seja, regiões que capturam a maior atenção visual. O trabalho de L. *et al.* (2021) empregou a biblioteca de visão da Apple[1] para detectar saliências visuais e auxiliar usuários na captura de fotografias com menos "entulho", ou seja, nas quais somente o elemento desejado está em destaque. Este processo envolve mapas de calor gerados por uma rede neural profunda, treinada com base no rastreamento ocular de observadores humanos.

[1] https://developer.apple.com/documentation/vision/cropping_images_using_saliency

Essa abordagem incorpora a atenção visual humana para atribuir maior importância às áreas de maior interesse, destacando o potencial das abordagens de aprendizado profundo na detecção do assunto principal. Entretanto, treinar uma rede neural para reproduzir a atenção humana a regiões específicas da imagem é um desafio. Montar um conjunto de dados que capture de maneira precisa e abrangente as preferências de atenção visual dos observadores humanos é complexo. Diferentes pessoas podem ter diferentes padrões de atenção, dependendo de fatores individuais e contextuais. Além disso, as informações sobre a atenção visual em diversas situações podem ser difíceis de coletar e representar de maneira adequada para o treinamento da rede neural.

O trabalho acima é um exemplo do avanço trazido para a detecção de objetos salientes com o uso de redes neurais convolucionais (CNNs). As CNNs, especialmente as redes totalmente convolucionais (FCNs), têm sido aplicadas com sucesso para prever a saliência de pixels. Outra abordagem para a detecção de objetos salientes é o trabalho de Zhang *et al.* (2023), que propõe uma rede dividida em duas partes: na primeira, um módulo de convolução dilatada em trisseção extrai características das camadas de uma rede principal. Supervisões intermediárias na forma de mapas de calor e mapas de borda são usadas para capturar informações de posição principal e detalhes de borda, respectivamente. Na segunda parte, essas características intermediárias são somadas a cada camada, e mapas de saliência previstos são obtidos em cada nível. Esses mapas de saliência são combinados para produzir a saída final. Essa abordagem visa aprimorar a extração de informações de localização global e detalhes de borda, levando a uma detecção de objetos salientes mais precisa. Observa-se que, mesmo em uma abordagem mais sofisticada, que envolve o contexto de uma cena, permanece a ideia de que há uma relação entre as bordas e a importância dos objetos na imagem.

Outra abordagem, proposta por Hasan *et al.* (2021), utiliza um método para detectar automaticamente transeuntes, ou seja, pessoas que estão em segundo plano em fotos. O objetivo principal do trabalho era usar esta informação para proteger a privacidade dos transeuntes em fotos que são compartilhadas online, mas este problema pode servir também para identificar os não transeuntes como sendo o sujeito principal. O método dos autores é baseado em um conjunto de conceitos intuitivos que os humanos usam para classificar os assuntos e transeuntes nas fotos. Esses conceitos incluem:

- A pose do transeunte: eles estão virados para a câmera? Estão olhando para o fotógrafo?
- A proximidade do transeunte com o fotógrafo: quão perto eles estão do fotógrafo?

- O tamanho do transeunte na foto: quão grandes eles aparecem na foto?
- A atividade do transeunte: eles estão envolvidos em uma atividade relevante para a foto?

Os autores utilizaram uma sequência métodos baseados em aprendizado profundo para detectar pessoas e suas poses, em um conjunto de dados de 20.000 fotos. Seu método obteve em torno de 93% de sucesso na detecção de transeuntes no conjunto de dados de teste.

Em síntese, essas investigações ilustram a evolução das estratégias para a detecção do assunto principal em imagens, abrangendo desde abordagens intuitivas até métodos baseados em aprendizado profundo. Cada abordagem possui seus próprios pontos fortes e limitações, ressaltando a importância de escolher a técnica apropriada de acordo com as características da imagem e os objetivos analíticos.

No presente trabalho, expandimos a premissa fundamental da detecção do sujeito principal nas imagens, enriquecendo-a com a dimensão semântica capturada pelas legendas. Diante da complexidade da identificação do sujeito principal, que envolve desde características anatômicas até saliências visuais, nossa abordagem capitaliza o conhecimento obtido pelas redes neurais treinadas para legendagem automática. Partimos da hipótese que as legendas não apenas descrevem o conteúdo visual, mas também revelam indiretamente a relevância dos elementos presentes na imagem. Ao estabelecer essa ligação intrínseca, buscamos não apenas aprimorar a precisão da identificação do sujeito principal, mas também inaugurar um novo paradigma que une a interpretação visual ao processamento linguístico. O resultado é uma abordagem integrativa que promove a compreensão holística das imagens e, ao mesmo tempo, destaca o potencial de investigações futuras na interseção da visão computacional e processamento de linguagem natural. Além disso, a abordagem que propomos permite explorar a grande quantidade de datasets e modelos generativos pré-treinados que já se encontram disponíveis, reduzindo o esforço necessário para obter uma rede capaz de identificar o sujeito principal em imagens.

2.2 LEGENDAGEM AUTOMÁTICA BASEADA EM ESTRUTURAS FIXAS

Os primeiros trabalhos para a descrição automática de imagens se baseavam em estruturas rígidas de frase, com a identificação de elementos que pressupõe-se estarem presentes na imagem. Por exemplo, no trabalho de Farhadi *et al.* (2010), uma metodologia que busca ligar a imagem a um trio composto por (objeto, ação, cena) é proposta, como, por exemplo, “um homem andando a cavalo em um campo”. Esse tipo de abordagem ainda possui significativas limitações,

pois trata o problema como classificação e recuperação em uma base lexical, como é o caso Wordnet (MILLER, 1995), buscando a melhor frase que se encaixa à imagem, e assim não pode gerar sentenças de tamanho variável, compostas e fora do escopo do *dataset* de treinamento. Além disso, estas abordagens não são gerais o suficiente para permitir a identificação do sujeito principal. Por conta destas limitações, no presente trabalho, serão exploradas técnicas mais sofisticadas.

2.3 LEGENDAGEM AUTOMÁTICA BASEADA EM REDES CONVOLUCIONAIS E REDES RECORRENTES

Os trabalhos de Krizhevsky *et al.* (2012) e Simonyan e Zisserman (2015) concluem que as redes neurais convolucionais profundas e hierárquicas são mais eficientes na aprendizagem e na extração de características que CNNs rasas, utilizadas em algumas das abordagens anteriores. Inspirado nisso, o trabalho de Elliott e Keller (2013) cita as limitações previamente apresentadas e propõe outra metodologia, que busca resolver em partes essas limitações, utilizando dois modelos de CNNs profundas, o primeiro para segmentar diferentes tipos de objetos em uma imagem, e o segundo que utiliza uma correlação geométrica entre o que foi segmentado na imagem para buscar uma conexão linguística, por exemplo: dados três elementos identificados, pessoa, rua e bicicleta, a conexão mais lógica é: uma pessoa andando de bicicleta na rua. Nesse trabalho é notado que a grande limitação enfrentada pelos autores é gerar conexões eficientes no último modelo, o qual tem como objetivo criar a conexão linguística entre palavras.

Em 2014 foram propostos alguns trabalhos bem-sucedidos que utilizavam a abordagem sequencia a sequencia (BAHDANAU *et al.*, 2014) – a qual utiliza um codificador e um decodificador. A razão para que o resultado desses trabalhos fosse bem-sucedido é que a geração de legenda pode ser considerada análoga à "tradução de uma imagem" em uma sentença de linguagem natural — desta forma, seria possível explorar uma estratégia normalmente usada por tradutores, nos quais um texto em um idioma (ou, no caso, uma imagem) é projetado em um espaço que codifica o seu conteúdo, e que pode ser decodificado para um texto em outro idioma. A partir desses resultados, outros trabalhos surgiram utilizando como base as arquiteturas *encoder-decoder*, como Kiros *et al.* (2014) e Mao *et al.* (2014), que adotaram uma abordagem semelhante à geração, mas substituem uma rede neural *feed-forward* por uma rede recorrente (RNN, de *Recurrent Neural Network*). Seguindo o mesmo princípio, Vinyals *et al.* (2014b) e Donahue *et al.* (2014) utilizaram RNNs LSTM *(Long Short-Term Memory)* para seus modelos.

Ao contrário de Kiros *et al.* (2014) e Mao *et al.* (2014), cujos modelos veem a imagem a cada passo de tempo da sequência de palavras de saída, em (VINYALS *et al.*, 2014b). o modelo olha apenas para as *features* extraídas da imagem no início.

Outros autores continuaram propondo modelos para a descrição automática de imagem utilizando LSTMs. É o caso do trabalho de Wang *et al.* (2018). Conforme é citado pelos autores, com a utilização de LSTM foram resolvidas algumas limitações das RNNs tradicionais propostas em outros trabalhos: explosões e problemas de desaparecimento durante otimização de retro-propagação, as quais causam a perda de conexão entre contexto e palavras muito distantes.

Alguns trabalhos relevantes (HSU *et al.*, 2021; AUNG *et al.*, 2020; RAMPAL; MOHANTY, 2020) utilizam uma estrutura com dois modelos, um do tipo convolucional para extrair características gerais de uma imagem (por exemplo, Inception4 (SZEGEDY *et al.*, 2017) ou ResNet (HE *et al.*, 2016)), e outro LSTM (HOCHREITER; SCHMIDHUBER, 1997) para conectar a saída com uma estrutura linguística especifica, gerando assim uma descrição.

Além da LTSM existem outras arquiteturas exploradas em várias abordagens, é o caso das *Gated Recurrent Units* (GRUs) (CHUNG *et al.*, 2014), as quais são similares às redes LSTM (HOCHREITER; SCHMIDHUBER, 1997) — arquiteturas que resolvem o problema de dissipação de gradiente, mas com uma diferença: não possuem o estado da célula interna, ao invés disso, utilizam o estado oculto para transferir informações.

2.4 LEGENDAGEM AUTOMÁTICA BASEADA EM ATENÇÃO E TRANSFORMERS

O processamento sequencial, utilizado nas redes neurais recorrentes, possui alguns problemas. O principal deles é que são necessários muitos recursos para relacionar dados sequencialmente muito distantes um do outro. Alguns trabalhos como o ConvS2S (KAMEOKA *et al.*, 2018) e ByteNet (DENG *et al.*, 2009) utilizam redes neurais convolucionais e tentam reduzir o processamento sequencial através do paralelismo fornecido pela GPU, mas nesses modelos, o número de operações necessárias para relacionar sinais de duas posições arbitrárias de entrada ou saída cresce conforme cresce a distância entre posições, linearmente para ConvS2S e logaritmicamente para ByteNet. Isto torna mais difícil gerar conexões linguísticas entre blocos distantes. Visando resolver esse problema, foi proposta em 2017 a rede *Transformer* por Vaswani *et al.* (2017). Ela foi proposta por conta dos bons resultados que trabalhos anteriores que utilizavam o mecanismo de atenção trouxeram para a área (DEVLIN *et al.*, 2018; XU *et al.*, 2015b). Ela foi projetada especificamente para melhorar a eficiência do processamento

de linguagem natural em tarefas como tradução automática e resumo de texto. Ela é baseada no conceito de *self-attention*, o qual permite que o modelo considere todas as palavras de uma entrada de texto ao mesmo tempo ao tomar uma decisão. Ele funciona calculando uma pontuação de atenção para cada bloco da sequência, que indica o quão importante é cada palavra/bloco/conceito em relação aos demais para determinar o significado do texto. Essa arquitetura também foi a primeira de processamento de linguagem natural a não usar recorrência, o que a tornou mais eficiente e capaz de lidar com entradas de comprimento variável.

O pré-treinamento de linguagem tem auxiliado a pesquisa de processamento de linguagem natural, com o surgimento de modelos como BERT (DEVLIN *et al.*, 2018) e GPT (RADFORD *et al.*, 2018a) os quais utilizam tranformers. Vários estudos estão avançando o pré-treinamento (RADFORD *et al.*, 2018a; LIU *et al.*, 2019; YANG *et al.*, 2019), melhorando as tarefas e projetando arquiteturas de modelos mais sofisticadas. Além disso, o aprendizado auto-supervisionado também está sendo aplicado na visão computacional. O pré-treinamento multimodal também está ganhando força, com os pesquisadores aplicando estratégias de mascaramento e arquiteturas *encoder-decoder* para adaptar os modelos para tarefas de geração de imagens e textos (SU *et al.*, 2019). Existem vários esforços para criar modelos unificados, que buscam representar tarefas de maneira uniforme e processar informações de múltiplas modalidades de forma consistente (TAN; BANSAL, 2019; LI *et al.*, 2019; LI *et al.*, 2020a). No entanto, esses modelos ainda enfrentam desafios para manter o mesmo desempenho em várias tarefas simultaneamente, e para adicionar a capacidade de geração de imagens.

Seguindo nessa linha foi proposta a rede One For All (OFA) (WANG *et al.*, 2022). O OFA é um *framework* de pré-treinamento multimodal que busca unificar diferentes tarefas, como geração de imagens, fixação visual, descrição de imagens, classificação de imagens, modelagem de linguagem, etc., em uma abstração simples de sequência-para-sequência. O OFA foi treinado com 20 milhões de pares de imagem-texto disponíveis publicamente, mas mesmo assim alcança desempenho altamente competitivo, e tem capacidade de transferir para tarefas e domínios não vistos. O *framework* utiliza um paradigma de aprendizado de sequência-para-sequência unificado para pré-treinamento, *fine-tuning* e inferência em todas as tarefas. Tanto as tarefas de pré-treinamento quanto as tarefas de entendimento e geração são todas formadas como geração sequência-para-sequência. Ele usa um esquema compartilhado em todas as tarefas e especifica instruções elaboradas para discriminação.

3 MATERIAIS E MÉTODOS

Neste trabalho, apresentamos uma abordagem que aproveita o conhecimento implícito adquirido por uma rede neural de legendagem automática para enriquecer o contexto de fotografias de eventos de formatura, simplificando assim o processo de categorização e identificação do sujeito principal. Mais especificamente, investigamos três arquiteturas promissoras para executar essa tarefa: uma que combina uma Convolutional Neural Network (CNN) com unidades Long Short-Term Memory (LSTM), outra que incorpora um mecanismo de atenção e uma arquitetura mais elaborada conhecida como "One For All" (OFA). Além disso, exploramos o papel adicional das legendas como um suporte secundário no agrupamento de fotografias de formatura, mostrando que as próprias legendas podem ser transformadas em representações visuais na forma de nuvens de palavras para conjuntos de fotos.

Nas próximas seções, detalhamos os conjuntos de dados utilizados neste estudo (Seção 3.1), discutimos as arquiteturas de legendagem avaliadas em profundidade (Seção 3.2). Apresentamos a adaptação da arquitetura OFA para o propósito específico de identificação do sujeito principal (Seção 3.3), e descrevemos uma técnica que utiliza as legendas para criar representações visuais por meio de nuvens de palavras (Seção 3.4). Por fim, abordamos as métricas adotadas (Seção 3.5) para a avaliação detalhada das abordagens.

3.1 DATASETS

Um dos grandes empecilhos para o desenvolvimento deste trabalho foi conseguir uma massa de dados que possuísse fotografias de formatura descritas de forma correta e padronizada. Esta é uma tarefa que demanda tempo de anotadores especializados, e atualmente não existe nenhum dataset público disponível para este propósito. Para resolver este problema, buscamos criar um dataset utilizando a legendagem gerada por uma rede neural genérica, pré-treinada sobre um conjunto de imagens rotuladas por anotadores humanos. Esta rede foi utilizada como base para auxiliar o processo de rotulagem manual das imagens em um segundo conjunto, que possui imagens de diversas classes de eventos, mas que ainda não foram descritas (ou seja, imagens ainda não rotuladas). Nesta seção, descrevemos os datasets mais relevantes que utilizamos neste trabalho, além do procedimento utilizado para obter um dataset contendo imagens de eventos de formatura a partir deles.

3.1.1 Common Object In Context (COCO)

O dataset *Microsoft Common Objects in Context* (MS-COCO) (LIN *et al.*, 2015), inicialmente proposto em 2014, é um dos mais empregados por diversos trabalhos em aplicações como detecção, reconhecimento e segmentação de objetos, além da legendagem de imagens (LI *et al.*, 2020b; HSU *et al.*, 2021). O dataset é mantido por profissionais de diversas empresas como Google e Facebook, e serviu como base para vários desafios em anos recentes, estabelecendo um *benchmark* para comparação entre diferentes algoritmos e modelos para análise de imagens. O dataset possui aproximadamente 330.000 imagens, todas segmentadas e descritas pelo menos três vezes por anotadores distintos. A Figura 3 mostra um exemplo de imagem do dataset COCO, assim como suas legendas associadas. O dataset possui fotos de formatura, mas em diversos casos descritas de maneira genérica, como mostra a Figura 4.

Figura 3 – Exemplo de uma imagem segmentada e descrita retirada do dataset COCO. Nela, é visto um vendedor em uma feira vendendo diversas frutas. No caso desta fotografia, 5 legendas diferentes estão disponíveis. [Fonte: https://cocodataset.org]

No ano de 2015, foi publicado o artigo *Microsoft COCO Captions: Data Collection and Evaluation Server* (CHEN *et al.*, 2015), o qual traz uma metodologia atualmente adotada pelos mantenedores do dataset para o aceite das descrições feitas por anotadores, que visa padronizar a

Figura 4 – Exemplo de uma imagem rotulada de maneira genérica no Dataset COCO: *A woman holding on to a graduate who is using his smart phone*. [Fonte: https://cocodataset.org]

descrição e a contextualização de uma imagem no tempo presente com foco na ação do sujeito e não no que há na cena, evitando vícios de um anotador específico. Para isso algumas instruções básicas foram propostas, sendo elas:

- Descrever todas as partes importantes da cena;
- Não iniciar sentenças com "*There is*";
- Não descrever detalhes irrelevantes;
- Não descrever algo que possa ter ocorrido no passado ou ocorrer no futuro;
- Não descrever o que uma pessoa poderia ter dito;
- Não fornecer nomes próprios de pessoas; e
- As sentenças precisam conter ao menos 8 palavras.

Por exemplo, uma legenda como "tem pessoas na praia" é considerada incorreta, enquanto uma legenda como "muitas pessoas na praia em um dia ensolarado." é aceita como sendo válida.

O mesmo artigo também propõe o uso de algumas métricas para avaliação do trabalho dos anotadores, para aproximar ao máximo a descrição de diversas pessoas do padrão esperado: BLEU (PAPINENI *et al.*, 2002), ROUGE (LIN, 2004), METEOR (DENKOWSKI; LAVIE,

2014) e CIDEr (VEDANTAM *et al.*, 2015). Neste trabalho utilizaremos alguns desses métodos. Mesmo com métricas e instruções para descrever as imagens corretamente, alguns vícios e erros humanos de anotadores podem acontecer, por esse motivo cada uma das imagens do dataset possui ao menos 3 descrições de anotadores diferentes.

Para possibilitar a execução de vários testes comparando diferentes modelos, inicialmente utilizamos uma amostragem de 30.000 fotografias selecionadas aleatoriamente do dataset COCO. Estas imagens rotuladas foram utilizadas em conjunto com o dataset USED, que será descrito na Seção 3.1.3. O uso de imagens rotuladas do dataset COCO pode ajudar o modelo a entender o contexto de imagens genéricas, tarefa que pode ser útil mesmo quando queremos gerar legendas para um tipo específico de fotografia.

3.1.2 ImageNet

O dataset ImageNet (DENG *et al.*, 2009) é amplamente utilizado por diversos trabalhos, tanto em testes quanto como base para técnicas como *transfer learning*. Inicialmente apresentado em 2009, ele possui mais de 14 milhões de imagens dividas de maneira hierárquica em milhares de categorias, alguns exemplos são dados na Tabela 1. As imagens são rotuladas de maneira mais simples do que é feito no dataset MS-COCO, utilizando termos do Wordnet (MILLER, 1995), o qual é um grande banco de dados lexical de inglês, que contém substantivos, verbos, adjetivos e advérbios, agrupados em conjuntos de sinônimos cognitivos. Este dataset possui um grande nível de detalhe em suas rotulagens. Um exemplo é ilustrado na Figura 5, onde um cachorro é rotulado tanto por um substantivo (*dog*) quanto por um adjetivo (*Maltese*).

Tabela 1 – Exemplos de 8 classes do *dataset* Imagenet.

0	Maltese dog, Maltese terrier, Maltese
1	pizza, pizza pie
2	toilet tissue, toilet paper, bathroom tissue
3	ballplayer, baseball player
4	chocolate sauce, chocolate syrup
5	Tibetan terrier, chrysanthemum dog
6	curly-coated retriever
7	Old English sheepdog, bobtail
8	malamute, malemute, Alaskan malamute

Para resolver o problema de erros humanos e padronizar a rotulagem, a abordagem do ImageNet é mais simples do que aquela usada no dataset MS-COCO; a mesma imagem é

Figura 5 – Exemplo de uma imagem rotulada do *dataset* ImageNet: Maltese Dog [Fonte: Dataset ImageNet].

classificada por mais de uma pessoa e só faz parte do dataset quando a maioria das rotulagens converge para o mesmo rótulo, conforme é citado em Deng *et al.* (2009). Cada classe possui um número diferente de convergências mínimas, que varia de acordo com o grau de concordância entre os anotadores. Por exemplo, uma foto de um gato birmanês pode precisar de vários anotadores distintos para confirmar o rótulo; já apenas o substantivo gato precisa de bem menos convergências para ser confirmado.

Em nosso trabalho, o dataset ImageNet foi empregado somente na extração de *features* das imagens, pois nas arquiteturas que testamos, foram utilizadas redes pré-treinadas sobre este dataset.

3.1.3 USED

o USED é um dataset [1] que possui 525.000 imagens de diferentes tipos de eventos sociais, como festas, shows, dia no parque, e casamentos. Desse total, 35.000 se referem a fotografias de eventos de graduação. As imagens são rotuladas, mas somente quanto ao tipo de evento no qual elas foram capturadas.

Para produzir um dataset de imagens legendadas, tomamos um subconjunto de 6.800 imagens de eventos de graduação, e as descrevemos manualmente, em um processo de ajuste de resultados, os quais foram inicialmente obtidos por um modelo genérico. Utilizamos a arquitetura OFA (que será descrita na Seção 3.2.3), que é a mais sofisticada das arquiteturas que testamos para legendagem automática. Como teste inicial, foi enviado um conjunto de dez fotografias de formatura para uma das variantes da arquitetura OFA (OFALarge), pré-treinada com o dataset MS-COCO completo. O modelo produziu um resultado aceitável para algumas imagens como por exemplo o mostrado na Figura 6, mas também ocorreram alguns erros, ou ao menos descrições vagas demais para o contexto de formaturas, o que era esperado, dado que o dataset utilizado é muito genérico. Por exemplo, fotos de uma família posando em uma colação sendo descritas como “um grupo de pessoas”, ou uma foto de comemoração sendo descrita como “um grupo de pessoas comemorando”.

Para obter legendas corretas para as 6.800 imagens, aplicamos a arquitetura OFA pré-treinada sobre elas, e todas as legendas geradas foram então ajustadas por anotadores humanos de forma a especializá-las para o contexto de formaturas. Esta estratégia foi empregada por facilitar o processo de rotulagem manual. Além disso, modificar todas as legendas ajuda a evitar que elas apresentem um viés que beneficie a arquitetura OFA nas comparações posteriores. As Figuras 7 e 8 mostram exemplos de imagens do dataset USED legendadas por este processo, junto com a descrição que o OFA gerou e a versão final, a qual foi ajustada por um anotador humano. Este conjunto de 6.800 imagens do dataset USED que rotulamos será o dataset principal utilizado em nossos testes.

Um subconjunto contendo 600 imagens foi selecionado dentre o total de 6.800 disponíveis para marcação manual da região contendo o sujeito principal de cada imagem. Esse processo de seleção considerou critérios relevantes para a representatividade e diversidade do conjunto de dados, como: fotos em auditório, igreja; fotos externas, etc. Posteriormente, as imagens selecionadas foram associadas a marcações contendo retângulos envolventes. A Figura

[1] http://loki.disi.unitn.it/~used/

Figura 6 – Exemplo de uma imagem que produziu um resultado aceitável: ***three people in graduation gowns posing for a picture*** **[Fonte: Dataset USED].**

Figura 7 – Legenda OFA: ***The students are walking on a field*****. Legenda ajustada:** ***The graduates are walking on field. They are wearing a black gown, a black mortarboard.***

9 ilustra um exemplo de uma imagem selecionada juntamente com sua respectiva marcação de retângulo envolvente, destacando o sujeito principal.

3.2 ARQUITETURAS PARA A LEGENDAGEM AUTOMÁTICA DE IMAGENS

Nesta seção, descrevemos três arquiteturas que serão avaliadas para a legendagem automática de fotografias de formatura, foco do presente trabalho. Na Subseção 3.2.1 é abordada

Figura 8 – Legenda OFA: ***Three people are wearing academic attire*****. Legenda Ajustada:** ***Two men and a woman are dressed in graduation robes.***

uma arquitetura mais simples, baseada em uma combinação de uma CNN e uma rede LSTM. Na subseção 3.2.2, abordamos uma arquitetura que adiciona ao conceito de CNN com rede recorrente um mecanismo de atenção. Já na Subseção 3.2.3 é abordado o modelo OFA, um modelo mais sofisticado, que representa o estado da arte no que se refere à legendagem automática.

3.2.1 Modelo CNN+LSTM para Legendagem Automática

Diversos trabalhos sobre legendagem automática de imagens utilizam abordagens que combinam redes neurais convolucionais (CNNs — *Convolutional Neural Networks*) e unidades LSTM (*Long Short-Term Memory*) — um tipo de rede neural recorrente (HOCHREITER; SCHMIDHUBER, 1997). Trataremos esta combinação por CNN+LSTM (HSU *et al.*, 2021; AUNG *et al.*, 2020; RAMPAL; MOHANTY, 2020). Nesses trabalhos, a CNN é usada para extrair as principais características de uma imagem, e uma segunda parte da rede, que inclui uma ou mais camadas LSTM, utiliza essas características para gerar uma sequência de palavras, que formam a legenda. Este conceito é ilustrado na Figura 10. A abordagem baseada em CNN+LSTM utilizada neste trabalho é a proposta no trabalho Show and Tell (VINYALS *et al.*, 2014a) a qual é uma instância de um modelo conhecido como *encoder-decoder* (HSU *et al.*, 2021). A arquitetura que utilizamos é ilustrada na Figura 11.

Figura 9 – Imagem do dataset USED com o sujeito principal selecionado através de uma caixa delimitadora.

Na arquitetura utilizada, a CNN faz o papel de *encoder*: recebendo como entrada uma imagem, ela produz um vetor de características que codifica o seu conteúdo. Normalmente, implementações deste conceito empregam arquiteturas de CNNs populares, que já obtiveram sucesso em tarefas de classificação de larga escala, como VGG (SIMONYAN; ZISSERMAN, 2015) ou InceptionV4 (SZEGEDY *et al.*, 2017). Isto permite que se aplique a técnica de *transfer learning*, com o uso de redes pré-treinadas sobre grandes conjuntos de dados, mesmo que para outras aplicações — os pesos são mantidos todos fixos, sem serem ajustados durante o treinamento; ou isso é feito para os pesos das camadas iniciais da rede, com as restantes sendo ajustadas por um processo de ajuste fino. A implementação que usamos neste trabalho utiliza uma rede InceptionV3, pré-treinada sobre o conjunto ImageNet (DENG *et al.*, 2009). Mantemos todos os pesos fixos e tomamos os valores produzidos pela penúltima camada da rede como um descritor, ou seja, o *encoder* transforma a imagem de entrada em um vetor contendo 2048 valores. Para todo o restante da tarefa de geração de legendas, as imagens não são consideradas, somente os vetores extraídos delas.

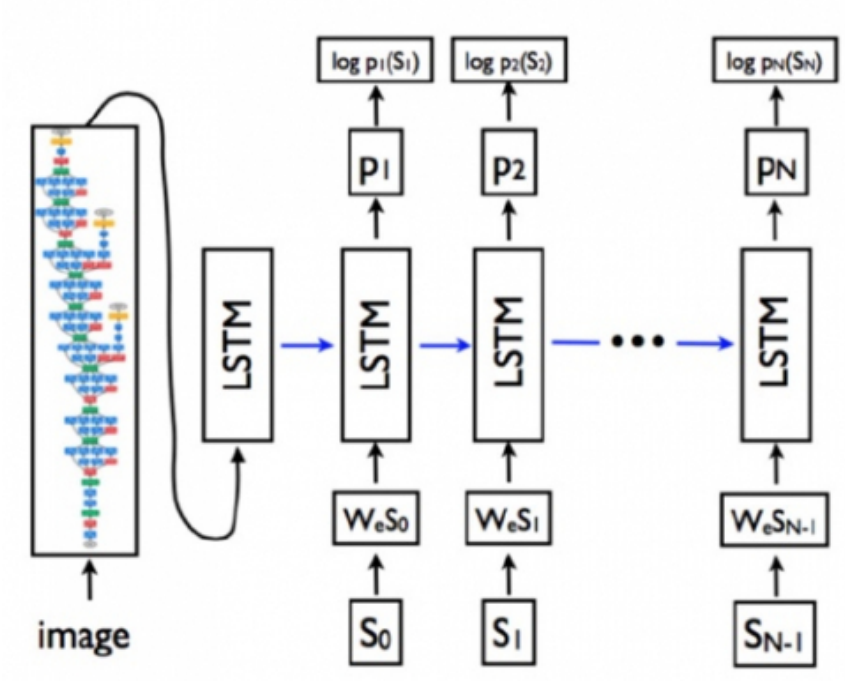

Figura 10 – Exemplo de processamento de uma imagem de entrada, com a utilização de uma CNN em conjunto com um modelo LSTM, para a descrição automática de imagens. Os blocos LSTM são mostrados de forma "desenrolada" no tempo: as setas azuis denotam a recorrência, com S_t sendo a sequência gerada até um dado momento t, e p_t sendo a saída do momento t [Fonte: (VINYALS *et al.*, 2014a)].

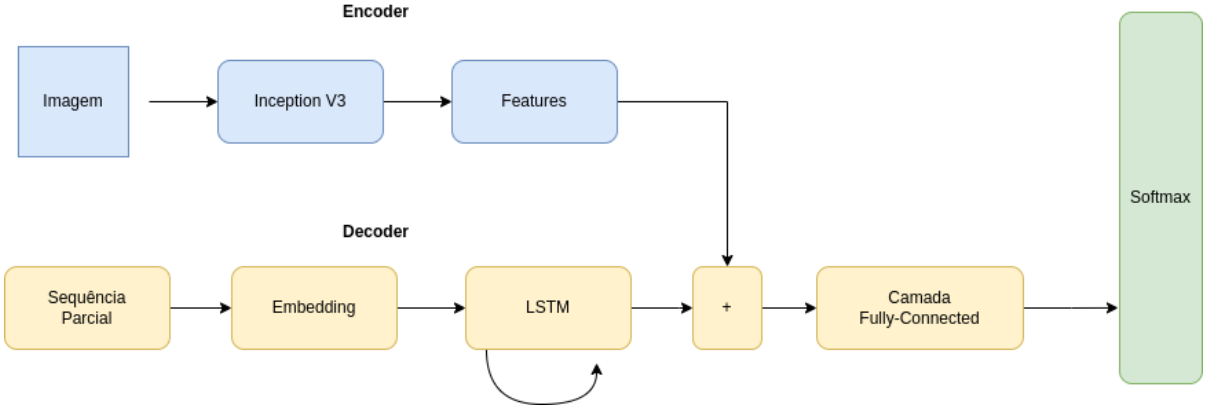

Figura 11 – Arquitetura geral da implementação de CNN+LSTM utilizada. O *encoder* recebe a imagem e extrai suas *features*. Uma camada LSTM recebe como entrada uma versão codificada de uma sequência parcial de palavras. A saída da camada LSTM é combinada com as *features* produzidas pelo *encoder*, e as camadas finais do *decoder* estimam qual seria a próxima palavra da sequência [Fonte: autoria própria].

Em uma abordagem para legendagem com CNN+LSTM, camadas LSTM são incluídas como parte do *decoder*, que é usado para gerar a sequência de palavras que formam a legenda. As redes LSTM são um tipo de rede neural recorrente RNN (*recurrent neural network*), que resolvem alguns problemas comuns nesse tipo de rede, como *vanishing/exploding gradients*. RNNs são utilizadas quando se geram saídas em sequência, com a saída produzida em um momento dependendo não somente das entradas naquele ponto, mas também das saídas produzidas em

momentos anteriores. Um exemplo prático de como esse conceito funciona é como o cérebro humano entende a linguagem natural, pois não seria possível contextualizar uma frase composta se não conectássemos cada palavra com várias palavras anteriores. Cada camada LSTM possui um vetor de estado oculto que armazena a informação retida até o ponto anterior da sequência, criando assim um *loop*, conforme é ilustrado na Figura 12. Nas redes LSTM, a saída em um momento pode ser afetada não somente pelas saídas do passo imediatamente anterior, mas também por saídas produzidas em vários passos anteriores, contando com um mecanismo para "esquecer" dados que se tornem irrelevantes, implementado por meio de múltiplas funções de ativação, chamadas *gates*.

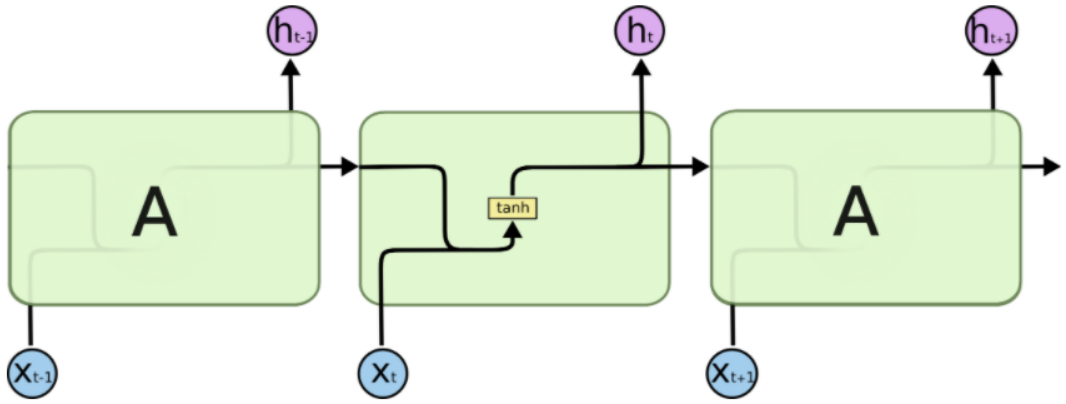

Figura 12 – Esquema do funcionamento das redes LSTM, nele é possivel observar as estruturas de *looping* e *feedback*, as quais retro-alimentam o modelo computacional. Aqui, X e h referem-se respectivamente às entradas e estados ocultos, os subscritos $_{t-1}$, $_{t}$ e $_{t+1}$ indicam diferentes instantes da sequência [Fonte: https://medium.com/turing-talks/turing-talks-27-modelos-de-predicao-lstm-df85d87ad210].

Para gerar uma legenda, o processo de inferência da rede é realizado várias vezes em sequência, com uma palavra sendo adicionada à legenda a cada inferência. A decisão sobre a próxima palavra é tomada pelo *decoder* com base na sequência gerada até o momento e nas *features* extraídas da imagem. Para mapear valores numéricos produzidos pela rede para palavras e vice-versa, utiliza-se o conceito de *tokens*: cada palavra é associada a um identificador numérico único (um *token*), que é mapeado para um espaço multidimensional por meio de um mecanismo de *embedding*, que normalmente é pré-treinado de forma que palavras de significados ou contextos similares sejam mapeadas para pontos próximos no espaço. Em nosso trabalho, utilizamos uma camada de *embedding* inicializada com pesos pré-treinados no dataset GloVe *word embeddings* (PENNINGTON *et al.*, 2014). A arquitetura que utilizamos possui uma única camada de unidades LSTM, que recebe como entrada uma sequência parcial de palavras, convertida por uma camada de *embedding*. A saída da camada LSTM e as *features* extraídas

da imagem são combinadas, com as camadas finais do *decoder* produzindo a próxima palavra da sequência, codificada por meio de uma camada *softmax*, com o número de valores de saída sendo igual ao número de possíveis palavras no dicionário.

De maneira prática, a geração de uma legenda acontece da seguinte maneira:

1. Primeiramente, processa a imagem utilizando o *encoder*, no caso uma rede InceptionV3;
2. Inicializa o *decoder* com o *token* especial `<start>` e o vetor produzido pelo *encoder*.
3. Enquanto o *decoder* não gera o *token* especial `<end>`:
    a) Invoca a camada LSTM, que recebe a sequência gerada até o momento, codificada por meio de uma camada de *embedding*.
    b) Combina as *features* da imagem com a saída da camada LSTM.
    c) Passa pelo restante do *decoder*. A saída da rede é a nova palavra da sequência.
    d) Acrescenta a nova palavra à legenda.
    e) A sequência atualizada, incluindo a nova palavra, "retorna" para a camada LSTM.
4. Converte o vetor de saída em palavras.
5. Fim

Para o treinamento do modelo com CNN+LSTM, primeiramente cada legenda do conjunto de treino (o "alvo" do aprendizado) é transformada em uma sequência de *tokens*. Depois disso, a sequência-alvo é divida em subsequências, ou seja, para cada amostra, a entrada é um par com os descritores extraídos de uma imagem e uma subsequência (legenda parcial), e o alvo para a inferência é a próxima palavra da sequência. Desta forma, cada legenda disponível no conjunto de treinamento resulta em múltiplas amostras de treinamento, uma para cada palavra. Um exemplo prático é dado na Tabela 2. Durante o treinamento, a camada LSTM não é atualizada pela sequência que ela de fato gerou, e sim pela sequência parcial que seria correta, de acordo com a legenda presente no conjunto de treinamento. Este procedimento é conhecido como *teacher forcing* (LAMB *et al.*, 2016).

A arquitetura completa da rede CNN+LSTM que utilizamos neste trabalho é mostrada na Figura 13. Com todas as camadas do *encoder* e a camada de *embedding* permanecendo com os pesos pré-treinados, a rede possui aproximadamente 1,5 milhões de parâmetros.

i	Image feature vector	Partial Caption	Target word
1	Image-1	<start>	the
2	Image-1	<start> the	black
3	Image-1	<start> the black	cat
4	Image-1	<start> the black cat	sat
5	Image-1	<start> the black cat sat	on
6	Image-1	<start> the black cat sat on	grass
7	Image-1	<start> the black cat sat on grass	<end>

Tabela 2 – Exemplo de obtenção de múltiplas amostras para o treinamento da rede a partir de uma legenda (no caso, *"the black cat sat on the grass"*, onde a estrada em um instante é composta pelo vetor de características da imagem (resultado do *encoder*) e pela sequência parcial gerada até o instante anterior, e a saída é a próxima palavra da sequencia. [Fonte: https://medium.com/@harshmalra/image-captioning-using-lstm-with-flickr-data-21e3086fdabe].

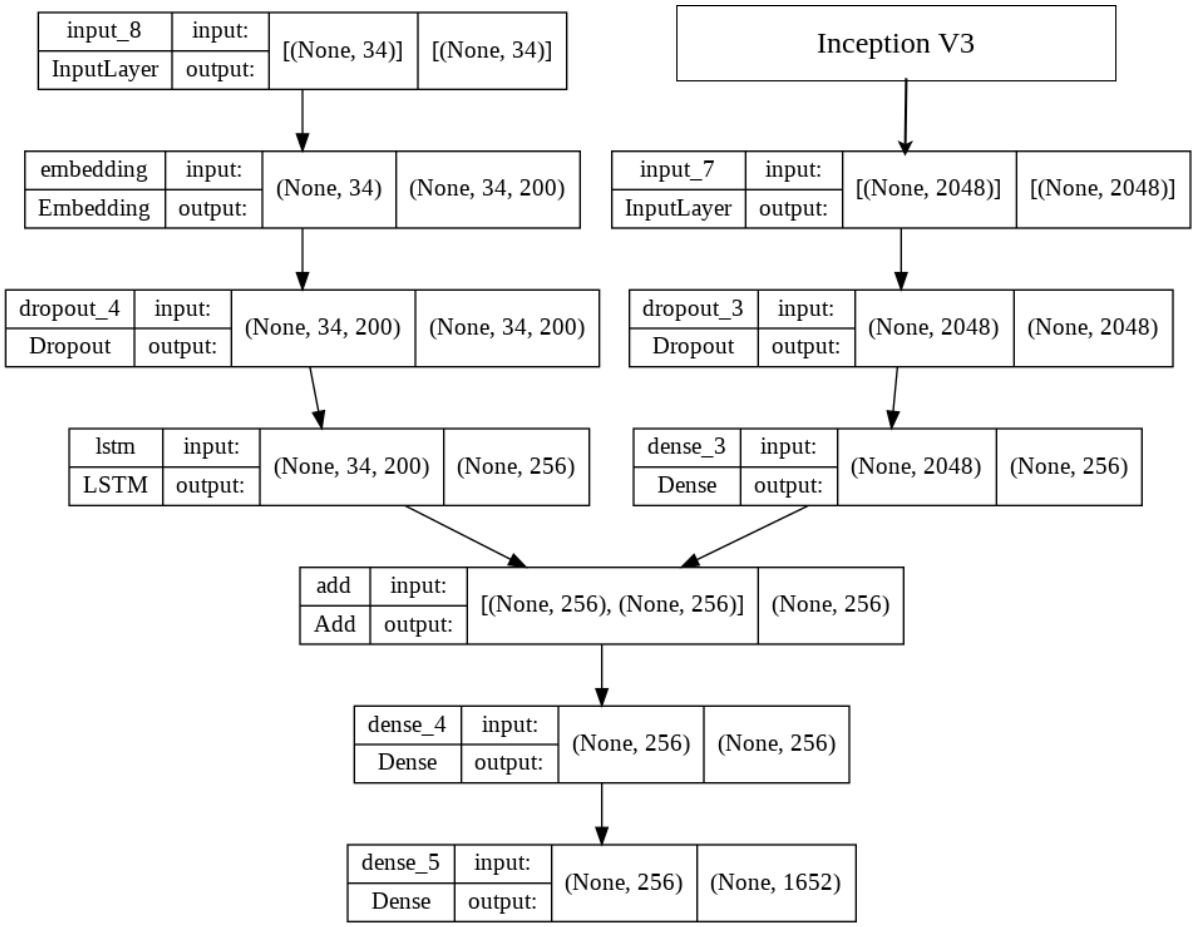

Figura 13 – Estrutura da rede CNN+LSTM utilizada. Pode-se observar a existência de duas entradas distintas: à direita, o *encoder* gera um vetor de 2048 valores; à esquerda, uma sequência parcial de até 34 palavras é codificada e serve de entrada para uma camada LSTM com 256 unidades. A saída da camada LSTM e o vetor de *features* (remapeado para 256 valores por uma camada totalmente conectada) são somados. A camada final da rede possui uma função de ativação softmax e 1652 valores de saída, que é o tamanho do vocabulário utilizado.

3.2.2 Modelo Baseado em Atenção para Legendagem Automática

As arquiteturas baseadas em atenção, mais especificamente aquela proposta no trabalho de Xu *et al.* (2015a), geram legendas para uma imagem de entrada encadeando sequencialmente

a busca de objetos e contextos em uma imagem, havendo uma interação entre o contexto identificado em um instante e o seguinte. Em outras palavras, esta abordagem permite que as diferentes *features* extraídas de uma imagem (e as regiões de onde elas vieram) recebam pesagens que vão mudando conforme a legenda vai sendo gerada, com diferentes *features* recebendo diferentes graus de "atenção" a cada momento. Isto contrasta com a abordagem descrita na Seção 3.2.1, que mantém pesos estáticos para o vetor de *features* da imagem. Por outro lado, existe uma potencial desvantagem que torna o modelo sensível a perder informações, pois ele pode focar a atenção na região errada da imagem, assim gerando descrições insatisfatórias. O mecanismo de atenção é ilustrado na Figura 14.

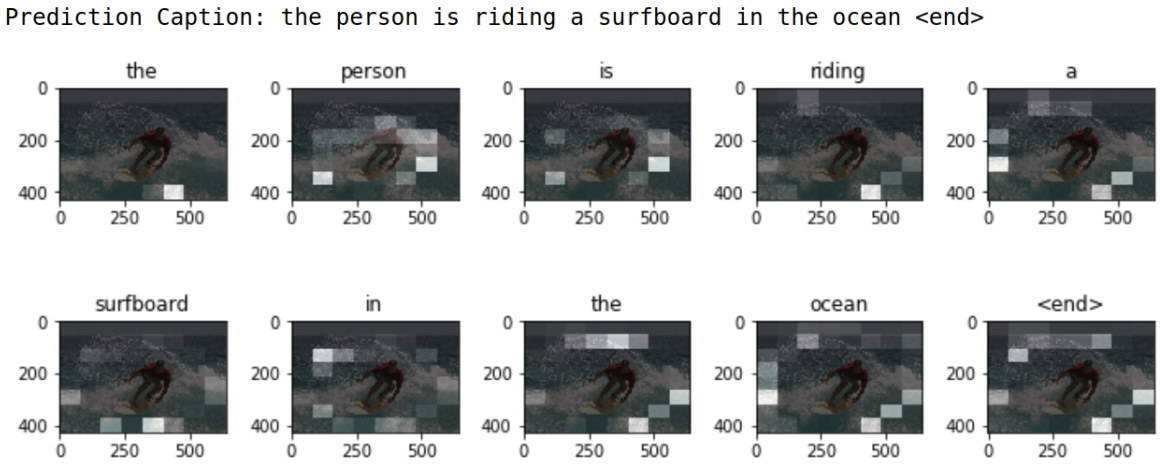

Figura 14 – Exemplo do mecanismo de atenção ao longo do tempo. À medida que o modelo gera cada palavra, sua atenção muda para refletir as partes relevantes da imagem [Fonte: https://www.tensorflow.org/tutorials/text/image_captioning].

Para o objetivo do presente trabalho, a riqueza nos detalhes da descrição é fundamental, pois quando nos referimos a fotos de formaturas alguns detalhes são essenciais e não podem ser descartados. Por exemplo, a Figura 15 mostra uma foto de família: é possível notar isso pelas vestimentas, idades, e distribuição das pessoas na foto. Já na Figura 16, é observado um grupo de formandos/colegas de turma, e cada um deles deve receber esta fotografia em sua versão final de separação das imagens (álbum).

De forma geral, a arquitetura baseada em atenção considerada no presente trabalho é similar àquela apresentada na Seção 3.2.1, com a adição do mecanismo de atenção, como mostrado na Figura 17. A implementação que utilizamos segue o conceito do trabalho de Xu *et al.* (2015a) mas, diferentemente da arquitetura proposta originalmente (ilustrada na Figura 18), na qual o *decoder* utiliza uma camada LSTM, é utilizada uma camada de GRUs (*Gated Recurrent Unit*) (CHUNG *et al.*, 2014), um tipo de RNN que pode ser considerado uma alternativa às LSTM,

Figura 15 – Exemplo de uma fotografia onde é observada uma família [Fonte: Dataset USED].

Figura 16 – Exemplo de uma fotografia, onde é observado um grupo de companheiros de turma [Fonte: Dataset USED].

que possui vantagens quanto ao consumo de memória e número de parâmetros, mas que pode ter o desempenho prejudicado quando trata de sequências muito longas. O uso das GRUs facilita também a aplicação do mecanismo de atenção sobre o estado interno da camada recorrente. Outra diferença entre a proposta original e a implementação que usamos é que o *encoder* utiliza uma rede InceptionV3 (SZEGEDY *et al.*, 2015), no lugar da rede VGG (SIMONYAN; ZISSERMAN, 2015) original — com isso, o poder de extração de *features* será o mesmo daquele utilizado pela arquitetura CNN+LSTM descrita na Seção 3.2.1.

Assim como a arquitetura CNN+LSTM que utilizamos, na implementação do modelo

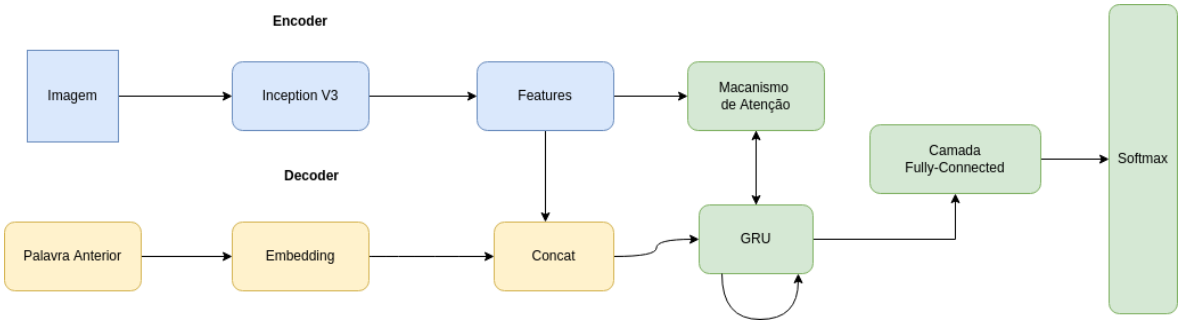

Figura 17 – Arquitetura geral da implementação baseada em atenção utilizada. Assim como na abordagem CNN+LSTM (Seção 3.2.1), o *encoder* recebe a imagem e produz *features*, e o *decoder* contém uma camada recorrente e produz uma palavra da sequência por vez, mas aqui é adicionado o mecanismo de atenção, que permite que as pesagens para as regiões da imagem variem conforme a legenda é gerada. A camada LSTM é substituída por uma camada GRU.

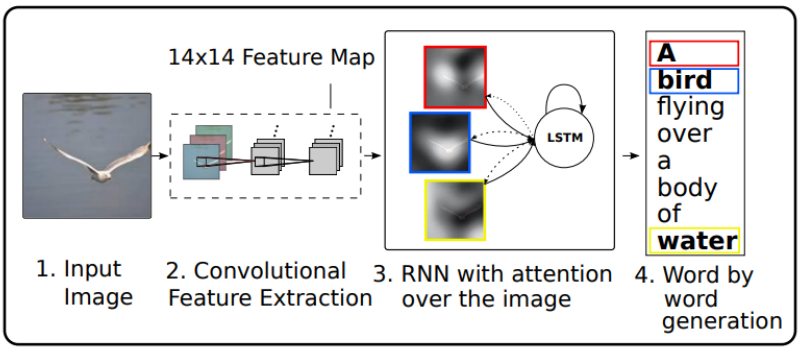

Figura 18 – Arquitetura do modelo do trabalho Show, Attend and Tell (XU *et al.*, 2015a). Nela é observada a imagem de entrada conectada a uma rede convolucional; a RNN do tipo LSTM e por último a saída que é a legendagem da imagem [Fonte: (XU *et al.*, 2015a)].

baseado em atenção, o *encoder* é composto por uma rede InceptionV3 (SZEGEDY *et al.*, 2015). Para não causar divergências quando os modelos forem comparados, tanto o *encoder* quanto a camada de *embedding* foram inicializados com os mesmos pesos utilizados para aquela arquitetura (respectivamente, pesos obtidos a partir de um pré-treinamento sobre o dataset ImageNet e do GloVe *word embeddings* (PENNINGTON *et al.*, 2014). A geração das legendas também segue o mesmo padrão: a rede é inicializada com o *token* `<start>` e invocada várias vezes, com uma palavra sendo gerada a cada chamada, até um tamanho máximo ou a geração do *token* `<end>`. A principal diferença entre as duas arquiteturas, fora o mecanismo de atenção em si, é que aqui a camada recorrente recebe as *features* da imagem na entrada, já ponderadas pelo mecanismo de atenção e concatenadas à saída da camada de *embedding*, que codifica a última palavra da legenda.

O mecanismo de atenção que utilizamos é aquele proposto por Bahdanau *et al.* (2014). Ele é uma versão do esquema que Xu *et al.* (2015a) colocam como *deterministic "soft" attention*. Este mecanismo surgiu a partir dos trabalhos de Cho *et al.* (2014) e Sutskever *et al.* (2014), os

quais codificavam uma frase de tamanho variável usando um vetor de tamanho fixo. Bahdanau *et al.* (2014) exploram o mecanismo de atenção para tratar de casos onde a saída do *encoder* pode ter tamanho variável, mas no nosso trabalho, este tamanho permanece fixo. Neste mecanismo de atenção, tanto as *features* da imagem quanto o estado oculto da camada recorrente passam por camadas totalmente conectadas. As saídas destas camadas são somadas, e uma média ponderada delas é feita, normalizada por uma camada *softmax*. Esta saída é um vetor de pesos, que é aplicado novamente sobre as *features* por uma multiplicação simples elemento-a-elemento. O que permite que os pesos mudem conforme a sequência é gerada é o fato de que o estado oculto da camada recorrente afeta a pesagem das *features*.

A arquitetura completa da rede baseada em atenção que utilizamos neste trabalho é mostrada na Figura 19. Com todas as camadas do *encoder* e a camada de *embedding* permanecendo com os pesos pré-treinados, a rede possui aproximadamente 1,5 milhões de parâmetros.

3.2.3 Modelo *One For All* para Legendagem Automática

A arquitetura *One For All* (OFA) (WANG *et al.*, 2022) é consideravelmente mais complexa do que aquelas apresentadas nas seções anteriores. Ela não trata somente da tarefa de legendagem automática de imagens, tendo sido proposta como um *framework* único que também inclui diversas outras tarefas, tais como *visual grounding* (dizer qual região da imagem contém um elemento buscado), *grounded captioning* (dizer qual o conteúdo de determinada região da imagem), preenchimento automático de regiões e de texto, e geração de imagens. Esta ideia é ilustrada na Figura 20. Além da arquitetura em si, Wang *et al.* (2022) focam no procedimento para pré-treinamento da rede sobre uma grande quantidade de imagens e textos vindos de diversos datasets, com posterior especialização (*fine tuning*) para tarefas específicas. Desta forma, o OFA é pensado não como uma arquitetura a ser treinada "do zero" sobre um dataset, mas como um conjunto de modelos completos, incluindo os pesos, para diferentes especializações e versões da arquitetura — foram propostas 5 versões, com variações no tamanho e número de parâmetros, e os nomes *tiny*, *medium*, *base*, *large* e *huge*.

O OFA foi treinado e validado em um grande conjunto de dados de imagens e texto, incluindo os seguintes conjuntos de dados: COCO (LIN *et al.*, 2015), Visual Genome (KRISHNA *et al.*, 2017), ReferIt (KAZEMZADEH *et al.*, 2014) e Visual Question Answering (VQA) (ANTOL *et al.*, 2015). Foi usando um processo de pré-treinamento e transferência, no qual primeiro é realizado o treino em um conjunto de dados grande e geral, como o COCO, e depois

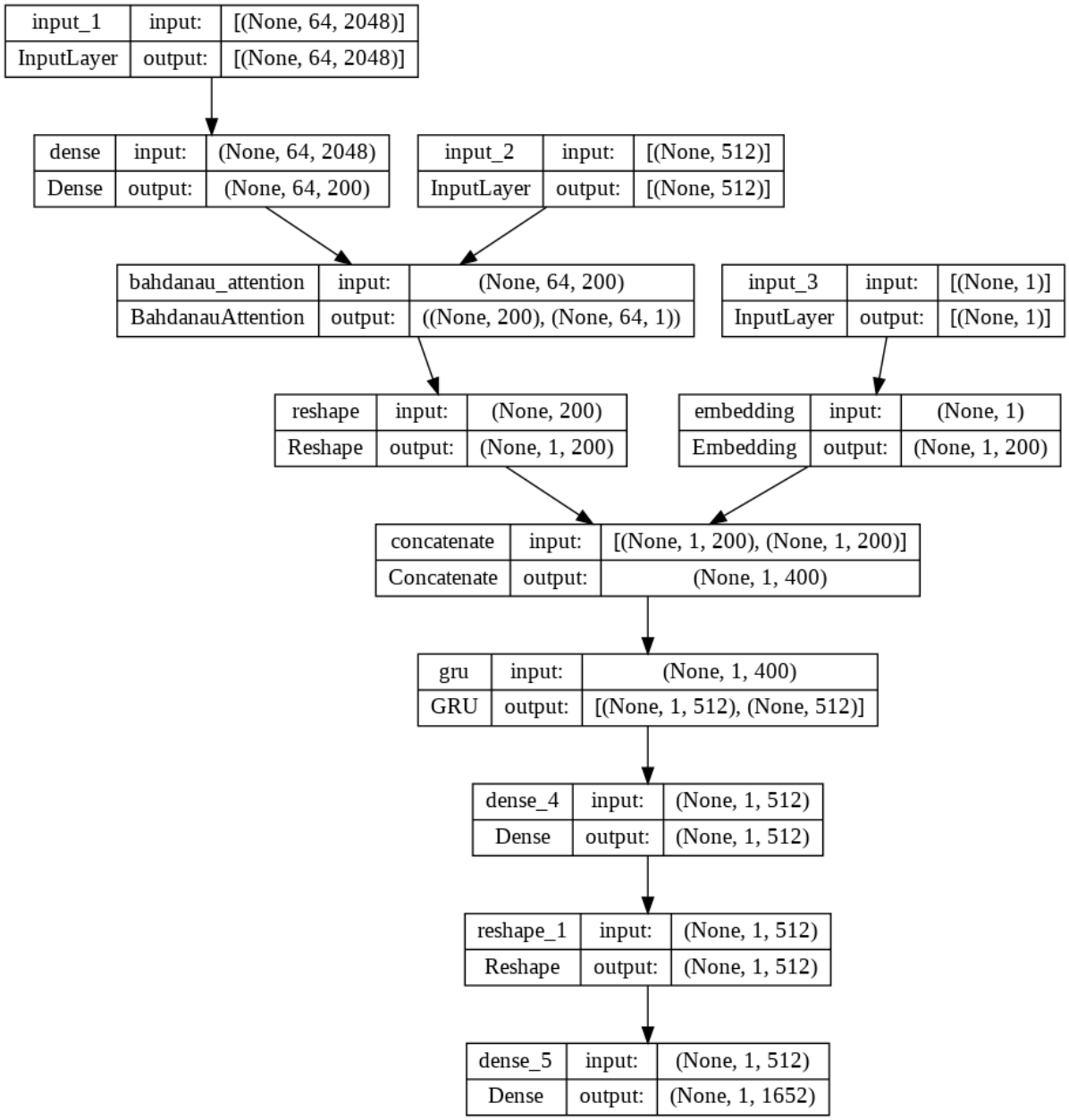

Figura 19 – Estrutura da rede baseada em atenção utilizada.

é feito um ajuste para uma tarefa específica, como *visual grounding* e legendagem automática. O processo de pré-treinamento de transferência permite que o OFA aprenda características gerais que podem ser usadas para realizar várias tarefas.

Para a tarefa de legendagem automática, o modelo foi especializado e testado sobre o dataset COCO, com a versão *huge* do OFA tendo obtido o melhor desempenho entre todas as soluções no *benchmark* na ocasião da sua divulgação. Já para a tarefa de *visual gronding* o modelo foi especializado e testado sobre o dataset COCO.

A arquitetura OFA possui tanto camadas convolucionais quanto camadas do tipo *transformer*. Modelos do tipo *transformer* foram propostos como uma alternativa aos modelos que

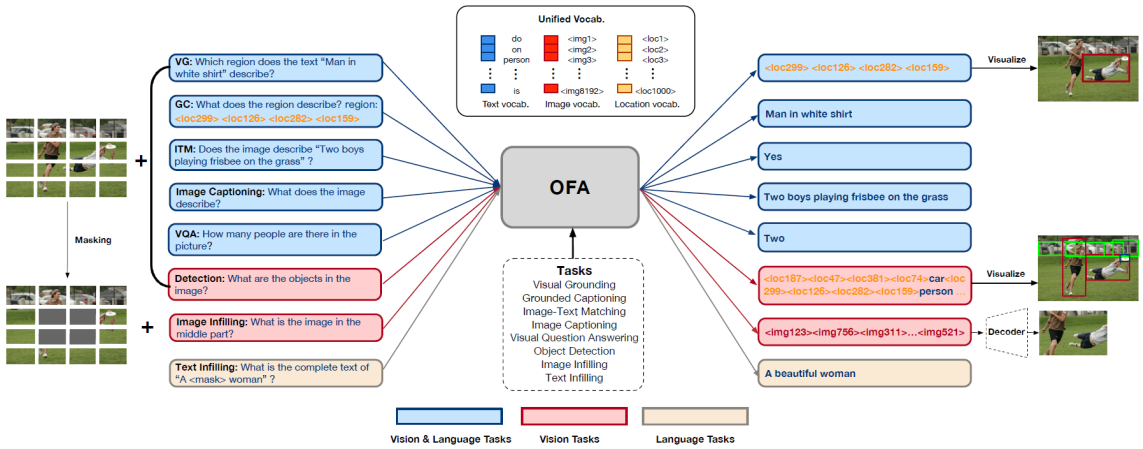

Figura 20 – Conceito geral do OFA. Uma única arquitetura é treinada para diversas tarefas, que são todas unificadas em suas representações e vocabulários. [Fonte: (WANG *et al.*, 2022)]

utilizam redes recorrentes com mecanismos de atenção para processamento de sequências linguísticas, como aquele apresentado na Seção 3.2.2 (VASWANI *et al.*, 2017). Além do OFA, outros modelos linguísticos de grande sucesso, como o GPT (*Generative Pre-trained Transformer*) (RADFORD *et al.*, 2018b), também se baseiam em *transformers*. Assim como em redes LSTM, um *transformer* recebe como entrada um vetor sequencial — por exemplo, uma sentença em linguagem natural — mas, diferente das redes LSTM, o processamento feito pelo modelo não tem que ser obrigatoriamente em ordem. Por exemplo, em uma frase um *transformer* não processa o início antes do final, e sim identifica o contexto que atribui o sentido/significado à frase. Como o processamento não é feito em ordem, o fluxo do modelo computacional pode ser paralelizado, diminuindo assim significativamente o tempo de treinamento, e viabilizando que o mesmo seja treinado com conjuntos de dados maiores. Em um modelo do tipo *transformer*, várias camadas de blocos do tipo *transformer* são colocadas em sequência. A principal técnica utilizada nestas camadas é o mecanismo de auto-atenção, que associa elementos de uma sequência com outros elementos da mesma sequência, de forma cruzada, de modo que a presença de certo elemento em uma posição afeta a pesagem da importância de outros elementos em outras posições (VASWANI *et al.*, 2017), como ilustrado na Figura 21.

A ideia central do OFA é unificar sob uma mesma representação diferentes tarefas e tipos de dados. Para isso, a arquitetura segue o padrão *encoder-decoder*, e todas as entradas e saídas são tratadas como sequências de *tokens* — um esquema conhecido como *seq2seq*. Os dados para todas as tarefas são convertidos para textos ou imagens — por exemplo, as saídas de um classificador são comumente dadas por um vetor contendo um valor para cada classe, mas no OFA, elas são convertidas para o texto associado ao rótulo daquela classe.

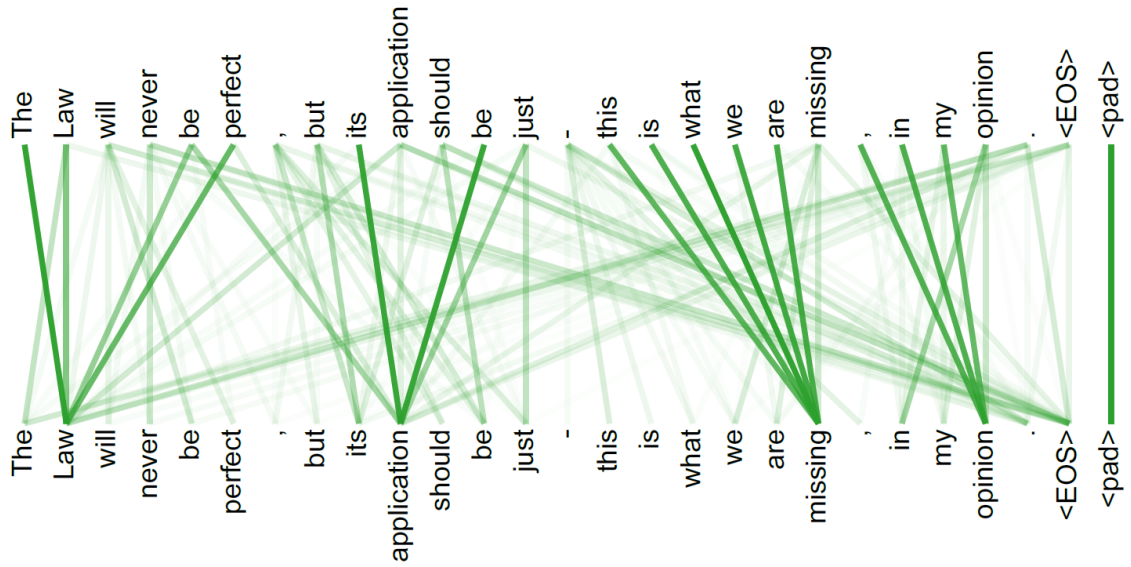

Figura 21 – Ao invés de gerar uma palavra por vez, um modelo do tipo *transformer* gera toda a sequência de forma simultânea, com um mecanismo de auto-atenção servindo para que elementos em cada posição possam afetar as demais posições. Normalmente, os modelos possuem múltiplas "cabeças", com pesagens diferentes para padrões e situações diferentes. Por clareza, neste exemplo são mostradas palavras, mas em uma implementação real, o conceito se aplica não somente a palavras (ou *tokens* assoaciados a palavras), mas também a conceitos mais complexos em sequências de camadas do tipo *transformer*. [Fonte: (VASWANI *et al.*, 2017)]

Para representar um texto como aquele de uma legenda, o OFA utiliza a técnica *byte pair encoding* (BPE), a mesma utilizada pelos modelos de linguagem GPT (RADFORD *et al.*, 2018b). Diferente da técnica mais simples usada para as arquiteturas CNN+LSTM e com o mecanismo de atenção, que consideravam cada palavra como um *token*, na representação BPE cada *token* se refere a uma sequência de caracteres, ou "sub-palavras". Em um esquema similar àquele utilizado para codificação de Huffman, inicia-se considerando um *token* para cada caractere de texto, e novos *tokens* são iterativamente criados a partir do agrupamento de pares de *tokens*, com os pares mais frequentes sendo agrupados a cada iteração. Desta forma, nesta codificação, um *token* pode se referir a uma palavra, a uma parte de uma palavra, ou mesmo a um caractere isolado. Isto permite que o espaço de *embeddings* funcione mesmo diante de palavras desconhecidas, com o aproveitamento de subsequências anteriores para representar e mapear estas palavras com sucesso. Por exemplo, mesmo que a palavra *friendly* não tenha sido observada durante a criação do dicionário, ela pode ser obtida pela combinação da palavra *friend* com a sequência comum *ly*; com o mesmo valendo mesmo para neologismos, com *unfriend* (*un* e *friend*). No OFA, foram considerados *tokens* como sequências de até 256 subsequências.

No OFA, uma imagem também é vista como uma sequência. Uma forma simples de se observar uma imagem como uma sequência, explorada por modelos to tipo *vision transformer*

Version	Params	Backbone	Hidden size	Intemediate size	Number of heads	Encoder layers	Decoder layers
OFATiny	33M	ResNet50	256	1024	4	4	4
OFAMedium	93M	ResNet101	512	2048	8	4	4
OFABase	180M	ResNet101	768	3072	12	6	6
OFALarge	470M	ResNet152	1024	4096	16	12	12
OFAHuge	930M	ResNet152	1280	5120	16	24	12

Tabela 3 – Diferentes versões do OFA, com suas respectivas características
Fonte: https://github.com/OFA-Sys/OFA

(DOSOVITSKIY *et al.*, 2021), é dividir a imagem em sub-regiões (*patches*) em um "grid", e considerar que cada sub-região é um elemento da sequência. No OFA, as sub-regiões não são consideradas diretamente, mas sim *features* extraídas delas por camadas convolucionais. Mais especificamente, as camadas iniciais de um modelo ResNet (HE *et al.*, 2016) são utilizadas para extrair mapas de *features* da imagem, que são subdivididos em um "grid", com as *features* relativas a cada sub-região de 16×16 pixels sendo consideradas como um *token* da sequência. Esta abordagem permite combinar características positivas das redes convolucionais (generalização e velocidade de convergência) com aquelas observadas em redes do tipo *vision transformer* (maior capacidade de representação, incluindo relações espaciais entre sub-regiões da imagem) (DAI *et al.*, 2021). Um ponto importante a se ressaltar é que no OFA, o treinamento para dados em formato de texto e de imagens ocorre simultaneamente, e que os *tokens* de texto e de imagens são mapeados para um espaço unificado.

O modelo explorado no presente trabalho é a implementação original dos criadores fo OFA[2]. Nela estão disponíveis 6 modelos de tamanhos diferentes, conforme mostra a Tabela 3. Escolhemos o modelo OFA Large por ele apresentar um equilíbrio entre o desempenho para a tarefa de legendagem sobre o dataset COCO e o custo computacional para treinamento e inferência.

3.3 IDENTIFICAÇÃO DO SUJEITO PRINCIPAL EM FOTOGRAFIAS

A identificação do sujeito principal em fotografias representa um desafio significativo no âmbito da visão computacional e do processamento de imagens. O objetivo fundamental é desenvolver algoritmos ou modelos que possam de maneira automática detectar e delimitar a área da imagem que contém o sujeito principal ou objeto de interesse. No contexto específico de álbuns de formaturas, esse problema assume uma relevância ainda maior. Fotografias desse tipo frequentemente apresentam várias pessoas, sendo que nem todas desempenham um papel

[2] https://github.com/OFA-Sys/OFA

importante ou são consideradas sujeitos principais. Portanto, a capacidade de filtrar e identificar com precisão quem são os sujeitos principais em tais fotografias torna-se essencial.

A aplicação prática dessa identificação se mostraria altamente valiosa, especialmente integrada ao reconhecimento facial. Ao utilizar a identificação dos sujeitos principais, seria possível direcionar os recursos de reconhecimento facial especificamente para os formandos que desejam receber suas fotos de formatura, otimizando o processo de organização e distribuição das imagens. Isso resultaria em uma experiência mais satisfatória para os formandos e seus familiares, que receberiam as fotografias corretas, contendo-os como sujeitos principais, em vez de meros coadjuvantes nas imagens.

Neste trabalho, nossa abordagem visa solucionar esse problema empregando a saída de atenção de uma rede de legendagem automática. A escolha dessa estratégia se baseia no fato de que uma rede deste tipo deve ser capaz de aprender a concentrar-se nas regiões corretas da imagem antes de gerar sua descrição. Isso possibilita focar apenas na região relevante, identificando com maior precisão o sujeito principal ou objeto de interesse. Em particular, redes com mecanismo de atenção são promissoras, visto que uma camada de atenção funciona como um mecanismo que pondera a importância de diferentes partes da imagem, destacando aquelas que são mais relevantes para a tarefa de legendagem. Essa característica pode ser particularmente útil em fotografias com várias pessoas ou objetos, pois poderia permitir direcionar o foco para a área que contém o sujeito principal, tornando o processo de identificação mais eficaz.

Considerando que como relatado mais adiante, o modelo OFA trouxe os melhores resultados em relação à legendagem automática, nos experimentos que realizamos, foram buscadas maneiras de se extrair a caixa delimitadora de onde o modelo focou mais a atenção ao descrever a imagem. Por exemplo, na Figura 22, onde a legenda gerada foi: *a man and a woman on a podium at a graduation ceremony*, a saída esperada seria a caixa verde delimitando apenas a região em que a descrição focou.

Para explorar a estrutura do OFA para a identificação do sujeito principal em fotografias, foi realizada uma implementação baseada em um estágio de *fine-tuning* da tarefa de *visual grounding*. Esta tarefa é definida como a localização de elementos da imagem a partir de uma *query* de texto. Por exemplo, uma entrada possível para esta tarefa seria: *"show where the red ball is in the image"*, e a saída seriam as coordenadas da caixa, definida por dois pares (x,y). Esta estrutura foi explorada como ponto de partida para localizar o sujeito principal da imagem. No entanto, simplesmente usar uma *query* como *"show where the main subject is in the image"*

Figura 22 – Imagem em que a saída esperada é a caixa delimitadora verde e a legenda gerada foi *a man and a woman on a podium at a graduation ceremony*.

não funcionou adequadamente no modelo pré-treinado para *visual grounding*.

Por isso, foi adotada a estratégia de realizar o *fine-tuning* do modelo previamente treinado para legendagem automática de imagens de formatura, utilizando o dataset de sujeito principal, descrito na Seção 3.1. Neste contexto, a camada de atenção do modelo não foi acessada diretamente. Em vez disso, aproveitou-se o fato de que sua existência permitiu um treinamento mais rápido do *visual grounding*, com um número consideravelmente reduzido de amostras. Isso ocorre porque as informações necessárias para a regressão da caixa delimitadora estavam implícitas na rede treinada para legendagem automática. Após o ajuste fino, o modelo conseguiu capturar características relevantes do sujeito principal nas fotografias, permitindo localizá-lo.

3.4 CRIAÇÃO DE NUVENS DE PALAVRAS A PARTIR DE LEGENDAS

A criação de nuvens de palavras a partir de legendas é uma técnica visual interessante que visa apresentar, de forma resumida e intuitiva, as palavras-chave mais frequentes e relevantes presentes nas legendas de imagens. Essa abordagem é útil para extrair informações importantes contidas nas descrições das imagens e permitir uma rápida compreensão do conteúdo geral do conjunto de dados.

O problema consiste em desenvolver um método para processar as legendas disponíveis, identificar as palavras mais relevantes e suas frequências, e, a partir disso, criar uma nuvem

de palavras visualmente atraente e informativa para o usuário final, no caso o responsável pela montagem dos álbuns de formatura.

Para atacar esse problema, este trabalho seguiu as seguintes etapas:

1. **Pré-processamento das legendas**: Inicialmente, é necessário realizar uma limpeza e normalização das legendas. Isso inclui a remoção de pontuações, caracteres especiais, *stopwords* (palavras comuns sem relevância para a análise), e a conversão de todas as palavras para letras minúsculas.

2. **Tokenização**: Em seguida, as legendas devem ser divididas em palavras individuais ou "tokens", facilitando a contagem das ocorrências de cada termo.

3. **Contagem de frequência**: Com as palavras tokenizadas, é possível criar um dicionário contendo cada palavra única e sua frequência de ocorrência nas legendas.

4. **Seleção de palavras relevantes**: É aplicada a técnica de ponderação TF-IDF (*Term Frequency-Inverse Document Frequency*) (JONES, 1972), para identificar as palavras-chave mais importantes, levando em conta tanto sua frequência na legenda quanto sua frequência global no conjunto de dados.

5. **Utilização de sinônimos**: Durante a seleção de palavras relevantes, é importante considerar que existem muitas palavras com o mesmo significado. Nesses casos, é possível utilizar um dicionário de sinônimos para aprimorar a precisão da nuvem de palavras.

6. **Criação da nuvem de palavras**: Com as palavras-chave selecionadas e suas respectivas frequências, é possível criar a nuvem de palavras. Nesse processo, as palavras mais relevantes recebem maior destaque visual, enquanto as menos relevantes são exibidas com menor tamanho ou opacidade.

A avaliação dessa etapa será totalmente qualitativa, pois não existem métricas objetivas para medir a qualidade da apresentação dos dados em uma nuvem de palavras. A eficácia será julgada com base na clareza e representatividade da nuvem de palavras em transmitir as principais informações das legendas de forma concisa e visualmente atrativa. Essa técnica, embora não apresente uma avaliação numérica, é uma forma valiosa de explorar e comunicar os *insights* contidos nas legendas das imagens de forma intuitiva e acessível.

3.5 MÉTRICAS PARA AVALIAÇÃO DOS MODELOS

Nesta seção, são descritas algumas métricas que podem ser usadas para avaliar as legendas produzidas por um dos modelos testados, assim como a qualidade dos retângulos envolventes para o sujeito principal

A avaliação da qualidade de uma legenda gerada normalmente é medida comparando-a com um conjunto de legendas produzidas manualmente por humanos. As métricas objetivas que são largamente utilizadas se baseiam na comparação direta entre palavras e sequências de palavras, ou n-gramas — um n-grama é uma sequência de palavras, com uma palavra isolada sendo um "1-grama", ou "unigrama". Esta estratégia é utilizada devido à dificuldade de se interpretar automaticamente o sentido de frases, o que torna a própria avaliação dos resultados um problema complexo. Optamos por utilizar métricas que já são amplamente aceitas pela comunidade, e que possuem implementações disponibilizadas livremente. As métricas consideradas são: BLEU (PAPINENI *et al.*, 2002), ROUGE (LIN, 2004) e METEOR (DENKOWSKI; LAVIE, 2014).

Já para os retângulos envolventes, foi considerada a métrica IoU (*Intersection over Union*) (BESL; JAIN, 1988), que é frequentemente utilizada para avaliar tarefas de detecção de regiões e segmentação.

3.5.1 BLEU

A métrica BLEU (*Bilingual Evaluation Understudy)* (PAPINENI *et al.*, 2002) foi proposta para medir a qualidade de um conjunto de traduções, comparando cada tradução com um conjunto de referências. Ela foi adaptada para a geração automática de legendas, sendo baseada na contagem de quantos n-gramas cada legenda gerada para uma imagem tem em comum com as legendas de referência para aquela imagem. Para cada legenda, computa-se a métrica em comparação com todas as referências correspondentes. Normalmente, o valor de n é variado no intervalo $[1,4]$ — esta limitação para sequências de até 4 palavras é adotada porque quanto maior a quantidade de palavras, menor a probabilidade de ocorrer uma subsequência em comum entre a legenda gerada e uma das referências, com valores muito grandes de n sendo problemáticos até mesmo para textos gerados por humanos.

Sejam S e C, respectivamente, o conjunto de legendas de referência e de legendas geradas, com $S_i = \{s_{i1}, \cdots, s_{im}\}$ sendo o subconjunto das m referências para a imagem i e c_i a legenda gerada para a imagem i. Seja também $h_k(s)$ o número de vezes que um certo n-grama k

ocorre em uma sentença s. Computamos, para um dado valor de n, a precisão $CP_n(C,S)$:

$$CP_n(C,S) = \frac{\sum_i \sum_k \min(h_k(c_i), \max_{j \in m}(h_k(s_{ij}))}{\sum_i \sum_k h_k(c_i)}$$

A equação indica, para cada imagem i, a contagem de quantas vezes cada n-grama possível k aparece nas legendas geradas c_i (valor indicado por $h_k(c_i)$), truncado de forma que a contagem não supere o maior número de ocorrências de k entre as referências pertencentes a S_i; em relação ao número total de ocorrências de k nas legendas geradas. Por exemplo, para $n = 1$, para cada legenda candidata, tomaríamos o número de palavras que ocorrem tanto na legenda candidata quanto em alguma das referências, limitado pelo maior número de vezes que aquela palavra aparece em uma das referências.

Como $CP_n(C,S)$ é uma medida de precisão, esta métrica pode favorecer legendas mais curtas — por exemplo, uma solução que gerasse legendas com apenas uma palavra teria pontuações altas se a única palavra de cada legenda aparecesse nas referências. Por isso, é adotada uma penalidade para legendas curtas:

$$b(C,S) = \begin{cases} 1 & \text{se } l_C > l_S, \\ e^{1-l_S/l_C} & \text{do contrário} \end{cases}$$

, onde l_C é o comprimento total de todas as legendas candidatas, e l_S é o comprimento das legendas de referência com o tamanho mais próximo de cada legenda candidata.

O valor final da métrica BLEU para um conjunto de legendas é obtido considerando-se $n = 1,2,3,4$, computando-se $CP_n(C,S)$ para cada valor de n e unindo os valores computados pela equação:

$$BLEU = b(C,S) \cdot \exp\left(\sum_{n=1}^{4} 0.25 \cdot CP_n(CS)\right)$$

3.5.2 ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (LIN, 2004) é um conjunto de métricas baseadas, como indica o seu nome, na revocação (*recall*). Elas foram originalmente propostas para avaliar a qualidade de resumos gerados automaticamente, comparando-os com resumos produzidos por humanos. Neste trabalho, consideramos 3 medidas: ROUGE-1, ROUGE-2 e ROUGE-L, descritas abaixo.

Para um valor n dado, a métrica ROUGE-n mede a revocação considerando a legenda gerada para uma imagem e as suas referências correspondentes, baseada no casamento de n-gramas presentes na legenda e nas referências, de forma similar ao que é feito pela métrica BLEU. Neste trabalho, consideramos as medidas ROUGE-1 e ROUGE-2, que tratam, respectivamente, de unigramas (i.e. palavras isoladas) e bigramas. Seja $h_k(c_i)$ o número de vezes que um certo n-grama k ocorre na legenda c_i gerada para a imagem i, e $h_k(s_{ij})$ o número de vezes que o n-grama k ocorre na referência j da imagem i. A métrica ROUGE-N para uma legenda c_i e o conjunto de m referências $S_i = \{s_{i1}, \cdots, s_{im}\}$ é dada por:

$$\text{ROUGE-}n(c_i, S_i) = \frac{\sum_j \sum_k \min(h_k(c_i), h_k(s_{ij}))}{\sum_j \sum_k h_k(s_{ij})}$$

Por exemplo, a métrica ROUGE-1 observará, para cada referência, a proporção de palavras da referência que estão também presentes na legenda. Desta forma, se a legenda possuir todas as palavras de todas as referências, o valor computado para a métrica será 1 (revocação de 100%), mesmo que a legenda possua muitas outras palavras que não estão presentes nas referências. Para a ROUGE-2, o conceito é o mesmo, mas considerando sequências de 2 palavras. Para obter a métrica ROUGE-N de todas as legendas geradas, computa-se a média da pontuação obtida por elas.

A métrica ROUGE-L é uma medida baseada na subsequência mais longa em comum entre a legenda e as referências LCS (*Longest Common Subsequence*). Uma LCS é um conjunto de palavras que aparecem na mesma ordem tanto na legenda quanto em uma das referências — mas diferente de um n-grama, podem existir outras palavras entre duas palavras pertencentes à LCS. Seja $l(c_i,s_{ij})$ o comprimento da maior subsequência em comum entre a legenda gerada c_i e uma referência s_{ij} para a imagem i, e $|s_{ij}|$ e $|c_i|$ respectivamente o tamanho total de s_{ij} e de c_i. São computados valores de revocação R_l e precisão P_l, e a métrica ROUGE-L é obtida a partir da média harmônica ponderada:

$$R_l = \max_j \frac{l(c_i,S_{ij})}{|s_{ij}|}$$

$$P_l = \max_j \frac{l(c_i,S_{ij})}{|c_i|}$$

$$\text{ROUGE-L}(c_i,S_i) = \frac{2.44 P_l R_l}{1.44 P_l + R_l}$$

Note que a média harmônica é ponderada de forma que uma revocação baixa possui sobre a métrica um impacto maior do que uma precisão igualmente baixa.

3.5.3 METEOR

A métrica METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) (DENKOWSKI; LAVIE, 2014) é baseada no alinhamento entre a legenda gerada para cada imagem e as suas referências correspondentes. Para alinhar uma legenda a uma referência, cada palavra da legenda é associada a uma correspondente na referência, utilizando heurísticas que buscam minimizar a troca de ordem e a manutenção de blocos contíguos de palavras. Para associar palavras, após buscar por correspondências entre palavras idênticas, buscam-se sinônimos, depois palavras que compartilham um mesmo tronco, e por fim, paráfrases. Exceto pelas correspondências exatas, são usadas listas/tabelas predeterminadas.

Dados os alinhamentos entre cada legenda gerada e as suas referências correspondentes, observam-se medidas de precisão e revocação. Seja c_i a legenda gerada para a imagem i, $S_i = \{s_{i1}, \cdots, s_{im}\}$ o seu conjunto de m referências correspondentes, e $a_i \in S_i$ a referência com o melhor alinhamento com relação a c_i, estes valores são calculados por:

$$P = \frac{\sum_i \sum_k min(h_k(c_i), h_k(a_i))}{\sum_i \sum_k h_k(c_i)}$$

$$R = \frac{\sum_i \sum_k min(h_k(c_i), h_k(a_i))}{\sum_i \sum_k h_k(a_i)}$$

Desta forma, a precisão indica a proporção de palavras nas legendas que também estão presentes nas suas referências correspondentes, e a revocação indica a proporção de palavras das referências que estão presentes nas legendas. Note que, diferente das métricas BLEU e ROUGE descritas anteriormente, aqui os valores de P e R se baseiam na comparação de cada legenda gerada com apenas uma das referências, escolhida com base no alinhamento, e não no conjunto completo de referências. Uma legenda com uma única palavra que também está presente na referência teria uma precisão alta mas uma revocação baixa; já uma legenda contendo todas as palavras do dicionário teria uma revocação alta, mas uma precisão baixa. A pontuação combina as duas medidas eu um único valor, a partir da média harmônica ponderada F:

$$F = \frac{PR}{0.9P + 0.1R}$$

Note que a média harmônica é ponderada de forma que uma revocação baixa possui sobre a métrica um impacto maior do que uma precisão igualmente baixa. Para medir o quanto as palavras na legenda e em uma referência mantêm a mesma ordem, é computada também uma penalidade baseada na quantidade de blocos contíguos de palavras alinhadas na legenda e na referência, com a observação de que dois textos similares tendem a possuir uma quantidade menor de blocos, cada um contendo várias palavras correspondentes. Dado ch o número de blocos contíguos nas legendas geradas e m a quantidade total de palavras nelas, a pontuação final da métrica METEOR é dada por:

$$METEOR = 1 - 0.5 \left(\frac{ch}{m} \right)^3 F$$

A métrica METEOR foi proposta para superar algumas limitações da BLEU, como os fatos da BLEU considerar somente correspondências exatas entre palavras; considerar somente palavras individuais, sem levar em conta a sua organização na frase; e se basear em uma medida de precisão, mas sem computar a revocação. A métrica METEOR consegue abordar estas limitações, além das limitações das métricas ROUGE, mas por outro lado ela também é consideravelmente mais complexa que estas alternativas.

3.6 IOU

A métrica IoU, também conhecida como Índice de Interseção sobre a União (do inglês, *Intersection over Union*) (BESL; JAIN, 1988), é amplamente utilizada na área de visão computacional e aprendizado de máquina para avaliar a eficácia de algoritmos de detecção e segmentação de objetos.

O IoU é calculado comparando a sobreposição entre uma região predita e uma região de referência, normalizando essa sobreposição pela área total coberta pelas duas regiões. Para isso, a área de interseção entre as duas regiões é dividida pela área da união das regiões. A fórmula para o cálculo do IoU é:

$$IoU = \frac{\text{Área de Interseção}}{\text{Área de União}}$$

O valor do IoU varia entre 0 e 1, onde 0 indica nenhuma sobreposição entre as regiões e 1 indica uma sobreposição perfeita, ou seja, as regiões são idênticas. Quanto maior o valor do IoU, maior a eficácia da detecção ou segmentação.

Para o cálculo do IoU, as regiões preditas por um modelo são comparadas a regiões de referência anotadas manualmente. Quando é necessário determinar se uma região predita é considerada correta ou incorreta, é comum definir um limiar de IoU (por exemplo, 0,5 ou 0,75).

4 EXPERIMENTOS E DISCUSSÃO

Neste capítulo, descrevemos o processo de escolha do modelo de legendagem automática usado como base para a identificação do sujeito principal em imagens, o que envolveu a avaliação do desempenho das técnicas descritas na Seção 3.2. Também descrevemos um teste inicial para classificação mais geral das imagens, que não realizava a legendagem, no qual foi constatada a limitação de tal abordagem. Então, relatamos os resultados obtidos para a identificação do sujeito principal, considerando somente o modelo pré-treinado e uma versão especializada após um passo de *fine-tuning*. Por fim, ilustramos os resultados obtidos após a geração de uma nuvem de palavras a partir das legendas produzidas pelo modelo.

Todas as implementações foram realizadas na linguagem Python 3.8, com as bibliotecas Keras e PyTorch, com execução na plataforma Colab do Google, por conta dessa oferecer recursos de hardware de alta potência, como GPUs e TPUs.

4.1 TESTES DOS MODELOS PARA LEGENDAGEM AUTOMÁTICA

Nesta seção, apresentamos os experimentos conduzidos para avaliar o desempenho dos modelos de legendagem automática de imagens. O objetivo principal desses experimentos é investigar como os diferentes modelos se comportam ao produzir descrições semânticas para imagens diversas, considerando especialmente o contexto de formaturas.

Os modelos descritos na seção 3.2 foram testados e comparados sobre um conjunto de dados contendo 6.800 fotografias selecionadas do dataset USED (descrito na Seção 3.1.3), mescladas com 30.000 fotografias e descrições já feitas, escolhidas de maneira aleatória do dataset COCO, por conta deste ser um conjunto de dados com contexto mais amplo e com mais amostras. O dataset foi divido da seguinte maneira: 70% para treinamento e 30% para validação/teste, todos os modelos foram treinados e testados com as mesmas divisões de conjuntos.

As métricas extraídas durante os experimentos devem ser interpretadas com relação a valores de referência. Neste trabalho, observamos o desempenho de anotadores humanos sobre o dataset COCO: computamos as métricas considerando, para cada imagem do dataset COCO, uma das legendas disponíveis comparada com as demais. Todas essas legendas foram produzidas por humanos. Os valores obtidos são mostrados na Tabela 4. Estes valores dão uma noção da escala esperada para cada métrica, e serão considerados nas discussões que faremos adiante.

METEOR	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
0.373415	0.059901	0.234218	0.03204	0.20554

Tabela 4 – Métricas calculadas sobre imagens do conjunto de dados COCO, considerando legendas produzidas por humanos.

4.1.1 Motivação para o Uso de Legendagem Automática

Em um primeiro momento, antes de abordar o problema utilizando legendagem automática, foram realizadas tentativas com uma abordagem mais simples, que atribuiria uma única classe a cada imagem. Para isso, rotulamos um subconjunto de 375 imagens do dataset USED conforme as classes fixas da Figura 23, utilizando uma ferramenta criada para esse propósito conforme ilustrado na Figura 24. No processo percebemos que não era possível rotular algumas amostras, pois essas não entram em nenhuma classe anterior pré-definida. O contorno para esse problema seria incrementar o número de rótulos, e assim aumentar o número de classes a serem treinadas. No final o *dataset* ficou mal distribuído, com alguns rótulos com muitas e outros com poucas amostras conforme mostra a Figura 23.

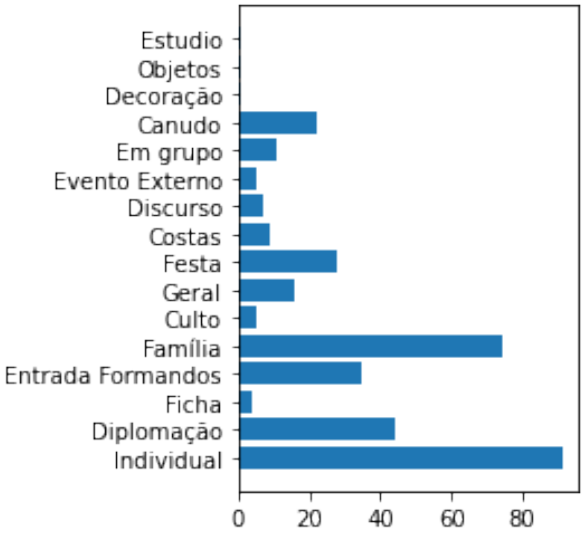

Figura 23 – Distribuição das classes da amostra de treinamento e validação [Fonte: Autoria Própria].

Para verificar como um treinamento com essa amostra se comportaria, o dataset foi divido em 70% para treinamento e 30% para validação/teste. Foi criado um modelo IncepvionV3 (SZEGEDY *et al.*, 2015) usando a técnica *tranfer learning* sobre um modelo pré-treinado com o dataset ImageNet, adicionando uma última camada do tipo softmax, e então treinando por 100 épocas utilizando a função de perda *categorical cross entropy*. O modelo treinado trouxe uma acurácia máxima de validação de 40% conforme é observado na Figura 25. Foi observada também a grande diferença entre a acurácia de treino e de teste, com a provável ocorrência de

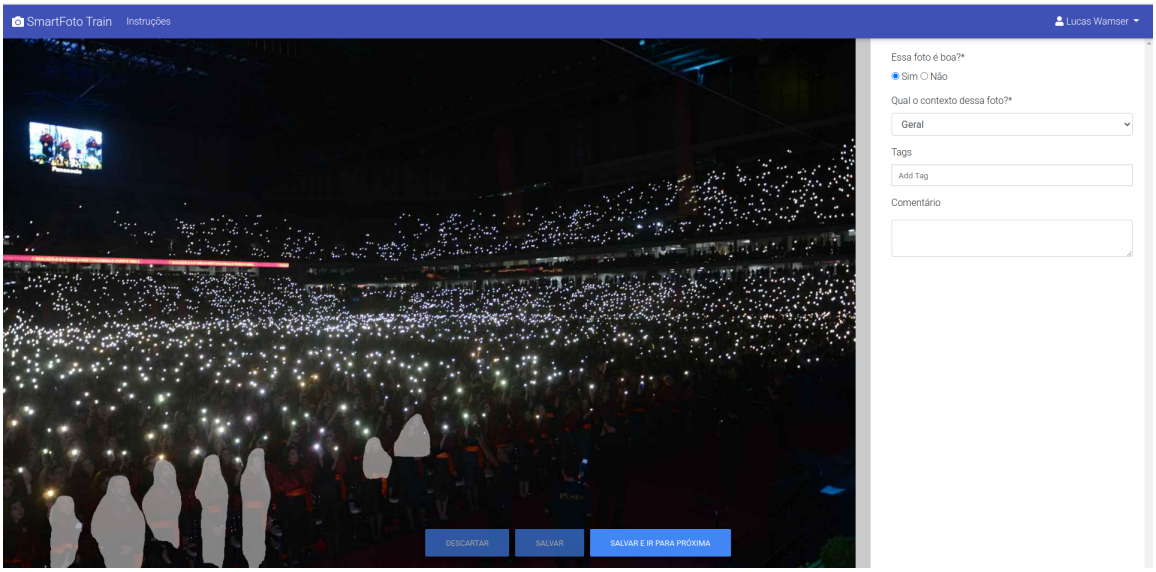

Figura 24 – Ferramenta de classificação utilizada para a rotulagem das imagens em testes preliminares, com apenas uma classe por imagem [Fonte: Autoria Própria].

overfitting. Acreditamos que o motivo disso são fotos e classes isoladas que possuem poucas amostras, o que poderia ser inicialmente resolvido com um incremento no tamanho do conjunto de imagens. Entretanto, observamos que esta solução gerava um outro problema: o aumento no número de classes, com algumas classes permanecendo com poucas amostras; além do aumento no número de casos ambíguos, onde não é possível atribuir claramente uma única classe a uma imagem.

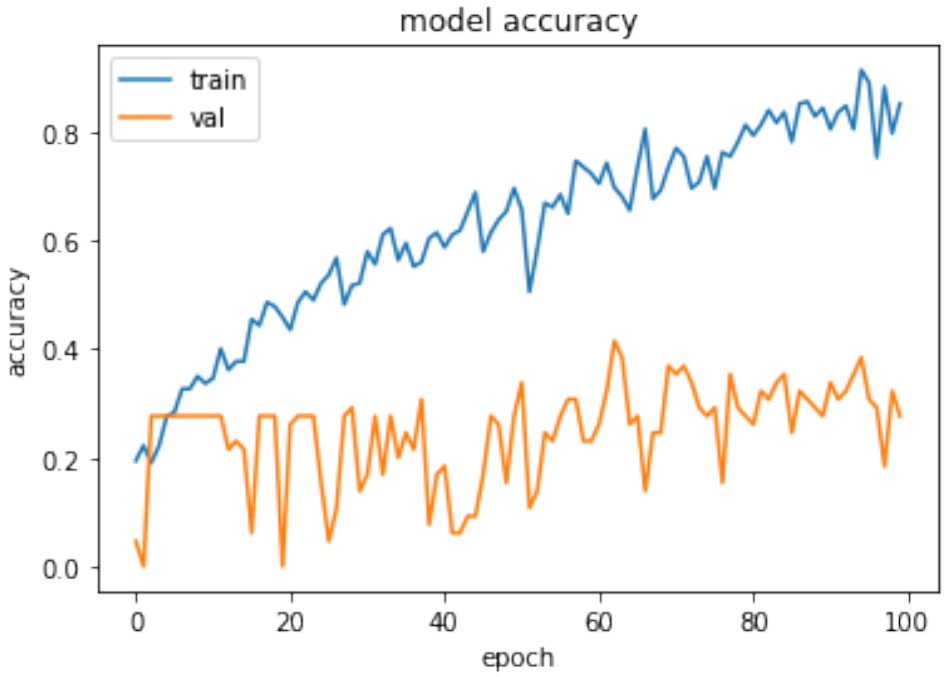

Figura 25 – Acurácia de treinamento × validação usando o otimizador Adam, uma taxa de aprendizado de 0.001, e um subconjunto de 375 imagens do *dataset* USED [Fonte: Autoria Própria].

Quando foi realizada a rotulagem das imagens, os classificadores perceberam que descrever uma imagem é uma tarefa mais fácil do que definir apenas um rótulo para ela. Por exemplo, a Figura 26 se refere a um evento externo, no caso da rotulagem foi necessário criar um rótulo para esse contexto, e quando ela é rotulada como “Evento Externo” podemos estar levando o modelo computacional ao erro, pois outras imagens muito parecidas com essa podem ter, por exemplo, um aluno recebendo o diploma com uma iluminação específica de palco com algum objeto na mão. Mas quando usamos a descrição podemos facilmente descrever a imagem: “um homem em um evento tocando uma guitarra” (embora de fato trate-se de um baixo elétrico, para um observador leigo, poderia ser uma “guitarra”), ou “um homem com beca recebendo um diploma no palco”. Nesses últimos exemplos nenhuma informação do contexto da fotografia é perdida e o nosso cérebro ajuda a contextualizar a fotografia.

Figura 26 – Exemplo de uma imagem que deve ser analisada no contexto de formaturas, no caso da rotulagem: “Evento Externo”, e com uma descrição literal: “Um homem em um evento tocando uma guitarra” [Fonte: Dataset USED].

Desta forma, os resultados nestes experimentos iniciais, com apenas uma classe por imagem, não foram satisfatórios para a generalização da amostra, mais especificamente para fotos de formaturas, então concluímos que apenas atribuir um único rótulo à foto não é o suficiente para contextualizar uma imagem de forma geral quando temos uma amostragem limitada para a classificação, nesse caso 375 imagens. Concluímos também que simplesmente ampliar o tamanho do conjunto não é uma solução eficaz para o problema, e que a legendagem automática, com posterior agrupamento por contexto, poderia ser uma saída mais promissora.

Modelo	METEOR	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
CNN+LSTM	0.30408	0.03306	0.11029	0.00674	0.10599
Humano	0.37342	0.05990	0.23422	0.03204	0.20554

Tabela 5 – Métricas calculadas sobre as imagens do conjunto de dados de teste, usando o modelo baseado em CNN+LSTM utilizando pesos pré-treinados na camada de *embbeddings*.

4.1.2 Testes com o modelo CNN+LSTM

Nesta seção são descritos os processos de treinamento/validação e os resultados para descrição automática de imagens, utilizando exclusivamente o modelo baseado em CNN+LSTM descrito na Seção 3.2.1.

Primeiramente foi realizado um treinamento com o conjunto de dados mesclado COCO+USED, sendo o conjunto dividido em 80% das imagens para treinamento e 20% para testes. Para o treinamento foi utilizado o otimizador ADAM, função de perda entropia cruzada, *batch size* 64, e taxa de aprendizagem de 0.001. Neste primeiro teste, o modelo foi treinado por 1000 épocas. A evolução da perda durante o treinamento é ilustrada na Figura 27, para os conjuntos de treinamento e teste — usamos aqui o conjunto de teste como validação apenas para fins de visualização, ele não afeta a escolha de pesos durante o treinamento. Como observado do treinamento após aproximadamente 400 épocas, o modelo começou apresentar *overfitting*, por conta disso as métricas foram calculadas sobre o *snapshot* do modelo com os pesos apresentados na época 400. A métricas calculadas sobre o conjunto de dados de testes ficou conforme ilustra a Tabela 5.

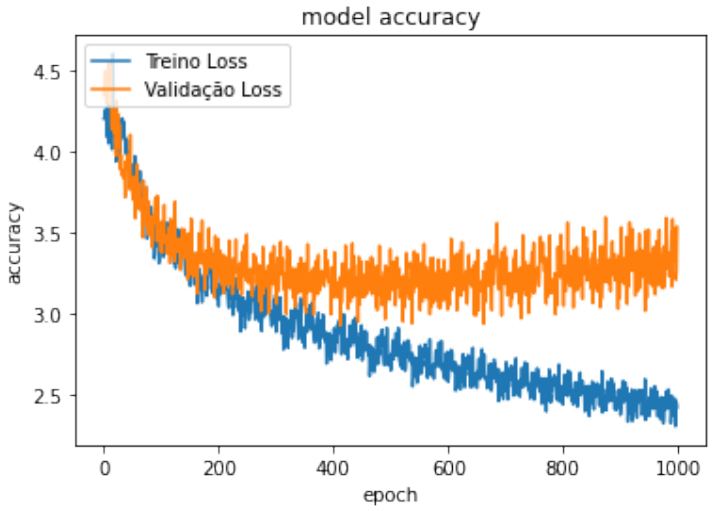

Figura 27 – Perda durante o treinamento do modelo LSTM, usando o otimizador ADAM e taxa de aprendizado de 0.001, com conjunto de dados mesclado COCO+USED. [Fonte: Autoria Própria].

Analisando as métricas depois de 400 épocas é possível observar que todas as métricas ficaram abaixo da referência de anotadores manuais. A METEOR teve um valor proporcionalmente mais próximo daquele das referências, mas para as demais, a diferença foi proporcionalmente maior. Com isso, podemos concluir que na maioria dos casos o modelo gerou legendas coerentes, mas utilizando sinônimos. Porém, em alguns casos a imagem não foi descrita corretamente e nem completa, conforme exemplo da Figura 28. Em outros casos a rede descreveu a imagem corretamente mas de maneira generalista, não capturando detalhes relevantes para o contexto de formaturas, como o exemplo na Figura 29, cuja legenda gerada não descreve que são formandos e estão em uma escada.

Figura 28 – Exemplo de uma fotografia descrita por um anotador manual como: *three women in graduation caps posing for a picture* e pela rede: *the little boy is wearing red helmet and holding the up of the*.

.

4.1.3 Testes com o modelo baseado em atenção

Nesta seção são descritos os processos de treinamento/validação e os resultados para descrição automática de imagens, utilizando exclusivamente o modelo baseado em atenção descrito na Seção 3.2.2.

Assim como no modelo CNN+LSTM, foi realizado um treinamento com o conjunto de dados mesclado COCO+USED, sendo o conjunto dividido em 80% das imagens para treinamento e 20% para testes. Para o treinamento foi utilizado o otimizador ADAM, função de perda entropia cruzada, *batch size* 64, e taxa de aprendizagem de 0.001. Neste primeiro teste, o modelo foi treinado por 1000 épocas. A evolução da perda durante o treinamento é ilustrada na Figura 30,

Figura 29 – Exemplo de uma fotografia descrita por um anotador manual como: *a group of graduates posing for a picture on the steps of a building* e pela rede: *group of people stand together in front of building*.

.

Modelo	METEOR	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Atenção	0.25206	0.03546	0.05974	0.00666	0.05974
Humano	0.37342	0.05990	0.23422	0.03204	0.20554

Tabela 6 – Métricas calculadas sobre as imagens do conjunto de dados de teste, usando o modelo baseado em atenção e sem inicializar a camada de *embbeddings* com pesos pré-treinados.

para os conjuntos de treinamento e teste — usamos aqui o conjunto de teste como validação apenas para fins de visualização, ele não afeta a escolha de pesos durante o treinamento. Após 1000 épocas de treinamento as métricas calculadas sobre o conjunto de dados de testes ficou conforme ilustra a Tabela 6.

Percebemos que após a primeira rodada de treinamento o modelo se tornou, assim como o modelo CNN+LSTM, muito generalista, perdendo alguns detalhes importantes das fotografias para o contexto de formaturas, como, por exemplo: “onde?”, “ fazendo o quê?” como mostra a Figura 31. A rede se tornou muito sensível em relação ao foco da atenção, e sendo assim cometeu alguns erros quando gerou descrições mais detalhadas, isso por conta que a atenção focou no pedaço errado da imagem, descrevendo pontos irrelevantes, conforme é ilustrado nas Figuras 32 e 33, que demonstram a atenção dada a diferentes áreas da imagem para assim formar a legenda.

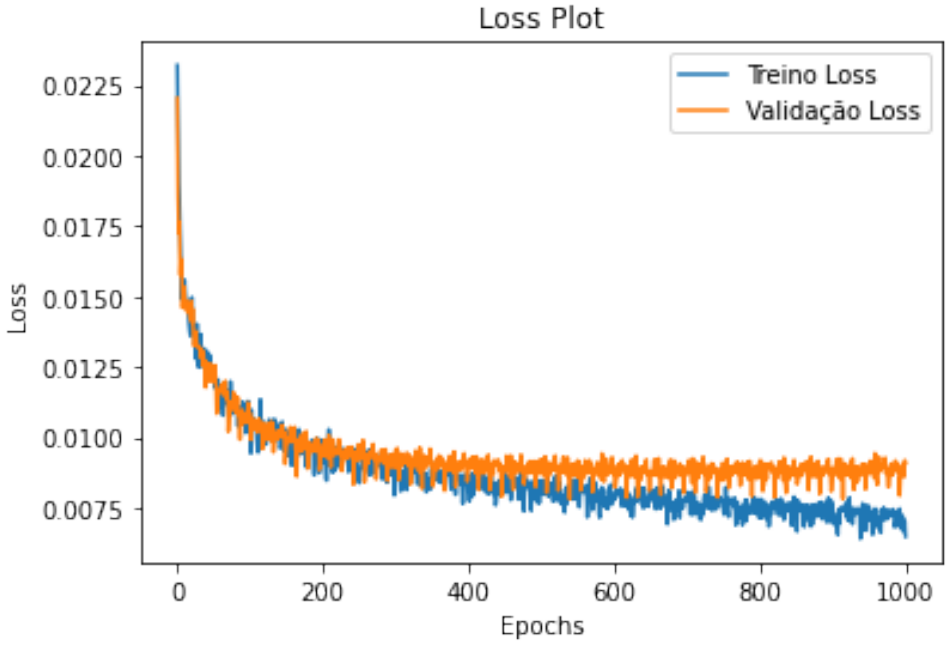

Figura 30 – Perda durante o treinamento, usando o otimizador ADAM e taxa de aprendizado de 0.001, com conjunto de dados mesclado COCO+USED. [Fonte: Autoria Própria].

Figura 31 – Exemplo de uma fotografia descrita por um anotador manual como: *a man and two people posing for a picture with a graduate* e pela rede baseada em atenção treinada por 1000 épocas como: *a man and two people*.

Analisando as métricas é possível observar que a METEOR tem valor mais alto e tanto a ROUGE (baseada em precisão) quanto a BLEU (baseada em revocação) possuem valores baixos. Isto indica que possivelmente a rede descreveu as imagens usando sinônimos das palavras esperadas, os quais apenas a métrica METEOR considera, um exemplo é dado na Figura 34, mas mesmo com este resultado, podemos concluir através deste teste que a rede teve um desempenho

Figura 32 – Exemplo de uma fotografia, descrita por um anotador manual como: *a man holding a billiard cue* e pela rede baseada em atenção treinada por 1000 épocas como: *two man and children*.

Figura 33 – Plot demonstrando a atenção da rede quando gerou a legenda: *two man and children*. É possível observar que a atenção focou na cena lateral, onde existem duas pessoas e uma agachada, o que pode parecer ser uma criança, em nenhum momento a rede focou na pessoa que está em primeiro plano, a qual está segundado um taco de bilhar.

insatisfatório para o objetivo desse trabalho. Embora o gráfico da Figura 30 mostre que após 1000 épocas a perda para o conjunto de testes não estaria melhorando, provavelmente não se trata de um caso de *overfitting*, pois o valor da perda está oscilando em torno de um certo valor, sem aumentar de forma considerável.

Modelo	METEOR	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Atenção	0.25633	0.01542	0.10730	0.00466	0.10475
Humano	0.37342	0.05990	0.23422	0.03204	0.20554

Tabela 7 – Métricas calculadas sobre o conjunto de dados de teste, usando o modelo baseado em atenção e pesos pré-treinados do dataset GloVe na camada de *embeddings*.

Diferentemente do modelo LSTM, o modelo baseado em atenção for inicialmente treinado sem utilizar pesos pré-carregados na camada de *embedding*. Consideramos a hipótese de que isto poderia estar afetando negativamente o desempenho da rede. Por conta disso, realizamos um novo treinamento por 1000 épocas adicionando os pesos pré-treinados do dataset GloVe, com os mesmos parâmetros usados anteriormente. Os resultados após esse ajuste são mostrados na Tabela 7.

Figura 34 – Descrição Feita por Humano: Three people posing for a picture with a graduate. Descrição Feita pelo modelo utilizando sinônimos: Trio of individuals posturing for a photo with a graduated person. [Fonte: Dataset USED].

Analisando as métricas é possível observar que a METEOR teve uma melhora pouco significativa, as métricas ROUGE-1 e ROUGE-L tiverem um incremento pequeno, mas um pouco maior, e o valor para as métricas BLEU e ROUGE-2 diminuiu. Isto indica que possivelmente a rede está produzindo legendas que contêm mais palavras corretas, mas também mais palavras irrelevantes entre elas. Uma possibilidade é que a rede começou a gerar mais legendas sem sentido linguístico, isso por conta que o GloVe na camada de *embedding* força a conexão entre palavras que aparecem geralmente juntas.

Para observar se mais épocas com uma taxa de aprendizagem menor melhoraria os resultados, foi realizado mais um treinamento por 4000 épocas, com os mesmos parâmetros, mas taxa de aprendizagem de 0.0001. O gráfico da função de perda é ilustrado na Figura 35. Percebemos que depois de 1000 épocas o modelo começou a apresentar *overfitting*, por conta disso, concluímos que o problema do mal desempenho da rede não se deve à falta de treinamento e sim à metodologia adotada ou até mesmo à arquitetura do modelo.

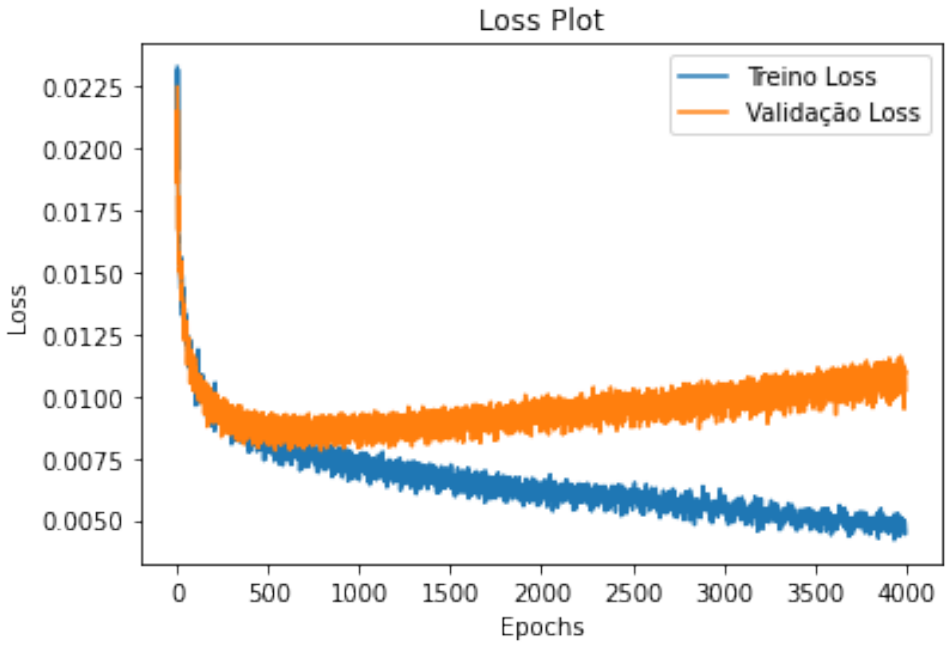

Figura 35 – Acurácia de treinamento, usando o otimizador Adam, uma taxa de aprendizado de 0.0001, com conjunto de dados mesclado COCO+USED. [Fonte: Autoria Própria].

4.1.4 Testes com o modelo OFA

Nesta seção são descritos os processos de treinamento/validação e os resultados para descrição automática de imagens, utilizando exclusivamente o modelo *One For All* (OFA), já descrito na Seção 3.2.3. O modelo utilizado, OFA-Large, já é pré-treinado com uma ampla quantidade de datasets, incluindo o COCO. Por conta disso, antes de realizar um *fine-tuning*, foram calculadas as métricas sobre o modelo pré-treinado. Os resultados são ilustrados na Tabela 8.

Como podemos observar, todas as métricas ficaram com o valor mais alto que as apresentadas pelos outros modelos testados. É possível observar que nesse caso as métricas METEOR e BLEU (baseada em revocação) possuem um valor mais alto do que a referência do dataset COCO, apresentada na seção 4.1.1 a qual apresenta os valores 0.373 e 0.059 respectivamente.

Modelo	METEOR	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
OFA	0.50050	0.08676	0.19697	0.01575	0.18328
Humano	0.37342	0.05990	0.23422	0.03204	0.20554

Tabela 8 – Métricas calculadas sobre imagens do conjunto de dados de teste do dataset COCO+USED, usando o modelo OFA.

Já as métricas ROUGE (baseadas em precisão) apesentaram valores um pouco mais baixos do que a referência. Com isso podemos concluir que a rede gerou as legendas corretamente, e em alguns casos de forma mais eficiente que anotadores humanos, mas não utilizando a mesma palavra ou conjunto delas (n-gramas) e sim sinônimos. Isso já é esperado, dado que anotadores humanos diferentes têm uma tendência a terem vícios em cada anotação, já o modelo tende a gerar legendas mais padronizadas.

Percebemos que a maioria das imagens foi descrita de maneira coerente e com riqueza de detalhes, mesmo que a descrição tenha sido um pouco diferente da realizada por anotadores manuais. Alguns exemplos são dados na Figura 36 e Figura 37, mas em alguns casos específicos o modelo errou em pequenos detalhes, como mostrado na Figura 38 a qual é a mesma que o modelo baseado em atenção descreveu incorretamente, por conta da atenção focada no pedaço errado da imagem.

Figura 36 – Exemplo de uma fotografia, descrita por um anotador manual como: *a graduate with his family* e pela rede OFA como: *a man and a woman posing for a picture with a graduate*.

Após essa avaliação do modelo pré-treinado, prosseguimos com o fine-tuning do OFA utilizando 80% do conjunto de dados COCO+USED para treinamento e 20% para teste. Com

Figura 37 – Exemplo de uma fotografia, descrita por um anotador manual como: ***group of graduates at graduation entrance ceremony*** **e pela rede OFA como:** ***a group of graduates walking on a field holding flags*****.**

Figura 38 – Exemplo de uma fotografia, descrita por um anotador manual como: ***a man holding a billiard cue*** **e pela rede OFA como:** ***a man holding a baseball bat*****.**

esse processo, obtivemos um novo modelo que apresentou um desempenho ligeiramente melhor em algumas métricas em comparação com o modelo pré-treinado, os resultados são apresentados na Tabela 9. Como é possivel observar, O METEOR, o BLEU e o ROUGE-2 aumentaram,

Modelo	METEOR	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
OFA (Pré-treinado)	0.50050	0.08676	0.19697	0.01575	0.18328
OFA (Fine-tuning)	0.51079	0.09676	0.19697	0.02575	0.19328

Tabela 9 – Métricas calculadas sobre imagens do conjunto de dados de teste do dataset COCO+USED, usando o modelo OFA.

Modelo	METEOR	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
CNN+LSTM	0.30408	0.03306	0.11029	0.00674	0.10599
Atenção	0.25633	0.01542	0.10730	0.00466	0.10475
OFA	0.50050	0.08676	0.19697	0.01575	0.18328
Humano	0.37342	0.05990	0.23422	0.03204	0.20554

Tabela 10 – Métricas calculadas sobre imagens do conjunto de dados de teste do dataset COCO+USED, usando o modelo OFA.

indicando melhorias na qualidade e na precisão das descrições geradas pelo modelo. No entanto, o ROUGE-L diminuiu ligeiramente, sugerindo uma variação no estilo das descrições.

Com base nessas análises, podemos concluir que o processo de fine-tuning do modelo OFA com o conjunto de dados COCO+USED resultou em melhorias no desempenho, com métricas geralmente mais altas.

4.1.5 Discussão Sobre as Técnicas de Legendagem Automática

Nesta seção, discutimos e compararamos os resultados dos três tipos de rede utilizados para legendagem automática: CNN+LSTM, baseado em atenção e OFA.

Ao comparar os resultados das três redes com os anotadores humanos para a tarefa de geração de legendas, conforme ilustrado na Tabela 10, observamos que, na maior parte dos casos, o desempenho das redes foi inferior ao obtido pelos anotadores humanos. As duas únicas exceções são as métricas BLEU e METEOR obtidas pelo modelo OFA. Tanto o modelo CNN+LSTM quanto o baseado em atenção foram capazes de gerar legendas coerentes em alguns casos, mas frequentemente as legendas geradas foram genéricas ou imprecisas, com descrições vagas ou confundindo importantes elementos da cena. Por outro lado, as legendas geradas pelo modelo OFA se mostraram mais completas, mesmo sendo o modelo *baseline*, treinado em um conjunto genérico de imagens (ou seja, sem especialização para fotografias de formaturas). Apesar de ainda existirem imprecisões, o bom desempenho obtido para a métrica METEOR mostra que, de forma geral, a rede utilizou muitos sinônimos, mas ainda assim foi capaz de descrever boa parte da cena com sucesso.

Uma diferença inesperada nos resultados é que o modelo baseado em atenção obteve

menos sucesso do que o modelo CNN+LSTM, mais simples. Esta diferença foi observada independente de termos ou não pesos pré-treinados na camada de *embeddings*. Observando alguns casos individuais, notamos que o maior problema é que o mecanismo de atenção não foca nos elementos principais da cena, escolhendo de forma pouco "inteligente". Isto pode ocorrer devido à forma incremental que este modelo gera suas legendas — uma vez que o foco da atenção esteja sobre um elemento da cena, todas as palavras seguintes da legenda podem ser afetadas por ele. Já no modelo OFA, que utiliza *self-attention*, a legenda é observada como um todo, de forma paralela, e o foco da atenção pode ser melhor direcionado aos elementos importantes da cena.

Portanto, o modelo OFA apresentou um desempenho superior na tarefa de geração de legendas, em comparação com os outros modelos avaliados (CNN+LSTM e baseado em atenção). Esses resultados indicam que o OFA é uma escolha promissora para aplicações que exigem geração de descrições de imagens mais precisas e contextuais, bem como identificação de elementos-chave nas imagens.

4.2 IDENTIFICAÇÃO DO SUJEITO PRINCIPAL

Para a identificação do sujeito principal, partimos da premissa de que um modelo treinado para gerar legendas já possui certo entendimento sobre quais elementos estão em destaque em uma imagem. Concentramos nossa análise no modelo OFA, que obteve os resultados mais promissores na tarefa de legendagem. Inicialmente, buscamos abordar o OFA com perguntas diretas sobre o *main subject* da imagem, utilizando o conhecimento adquirido durante o treinamento prévio. Contudo, os resultados não atenderam às expectativas, como evidenciado na Figura 39. A análise de diversos casos revelou que as respostas obtidas careciam de precisão, muitas vezes não capturando adequadamente o elemento central da imagem.

Para aprimorar o desempenho da identificação do sujeito principal, optamos por realizar um processo de *fine-tuning* na tarefa de *visual grounding*, utilizando um conjunto de 600 imagens previamente anotadas. O conjunto de dados foi dividido em 20% para teste e 80% para treinamento. A evolução do *fine-tuning* é ilustrada no gráfico da Figura 40, onde também apresentamos os resultados da validação utilizando o conjunto de teste em cada época — o conjunto de testes não foi usado para validação de fato, apenas para observar a evolução do desempenho após várias épocas. Ao final do procedimento de *fine-tuning*, para o cálculo das métricas finais, selecionamos o modelo gerado na última época de treinamento.

A Tabela 11 apresenta os valores médios de IoU calculados sobre imagens do conjunto

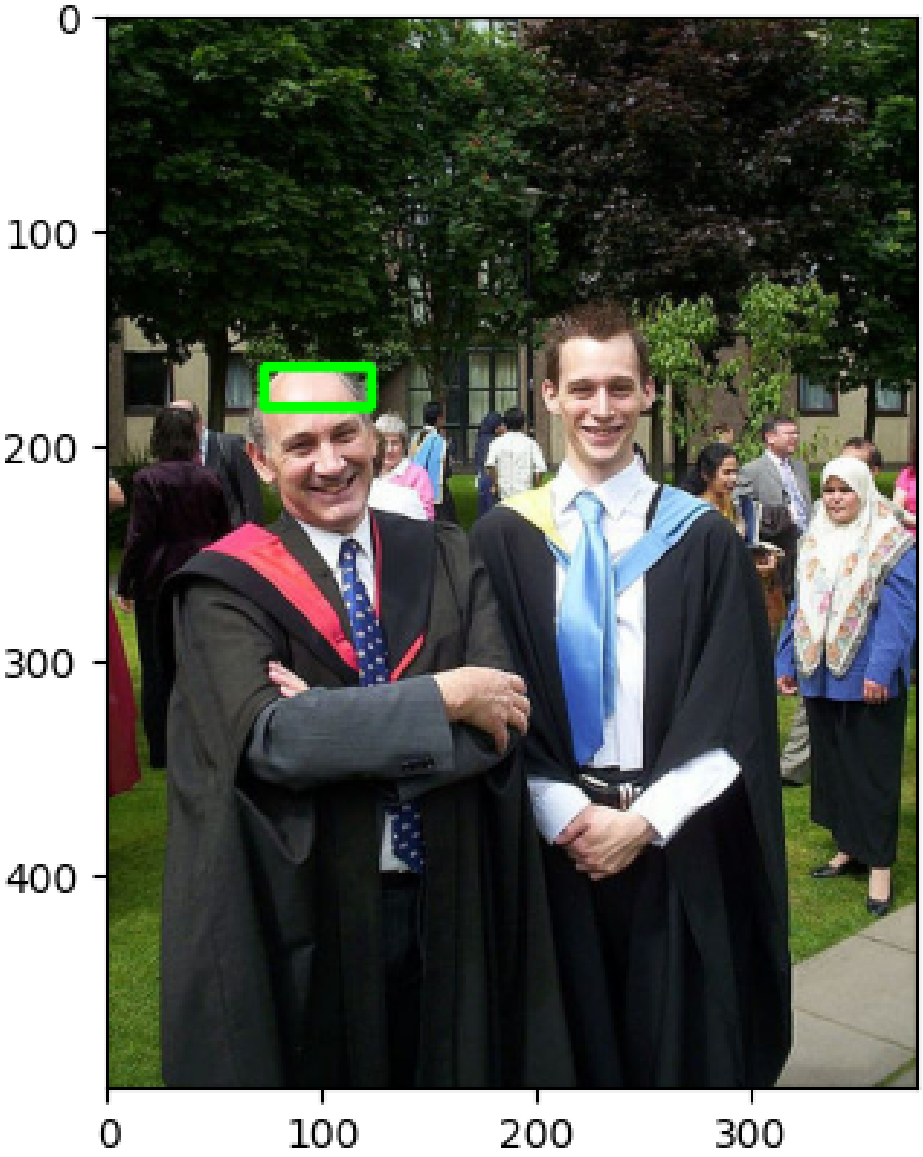

Figura 39 – Exemplo de uma fotografia, e seu respectivo *main subject* anotado pelo OFA, sem *fine-tuning*. O IoU obtido para esta imagem foi de IoU 0.07

de dados de teste, utilizando o modelo OFA Large, antes e depois do processo de *fine-tuning*. Foi possível constatar um aumento significativo do IoU médio. Os histogramas das Figuras 41 e 42 mostram os valores de IoU obtidos para todas as imagens, agrupados por faixas. Nota-se que a maior parte das imagens ficou com o IoU próximo de 0.5, valor considerado como "verdadeiro positivo" para detecções em desafios como Pascal VoC[1] e COCO. As Figuras 43 e 44 mostram exemplos de casos nos quais o modelo apresentou o comportamento esperado. Apesar do bom desempenho, ainda existiram casos nos quais o modelo apresentou resultados insatisfatórios, como mostram as Figuras 45 e 46. Mesmo assim, observa-se que o *fine-tuning*

[1] http://host.robots.ox.ac.uk/pascal/VOC

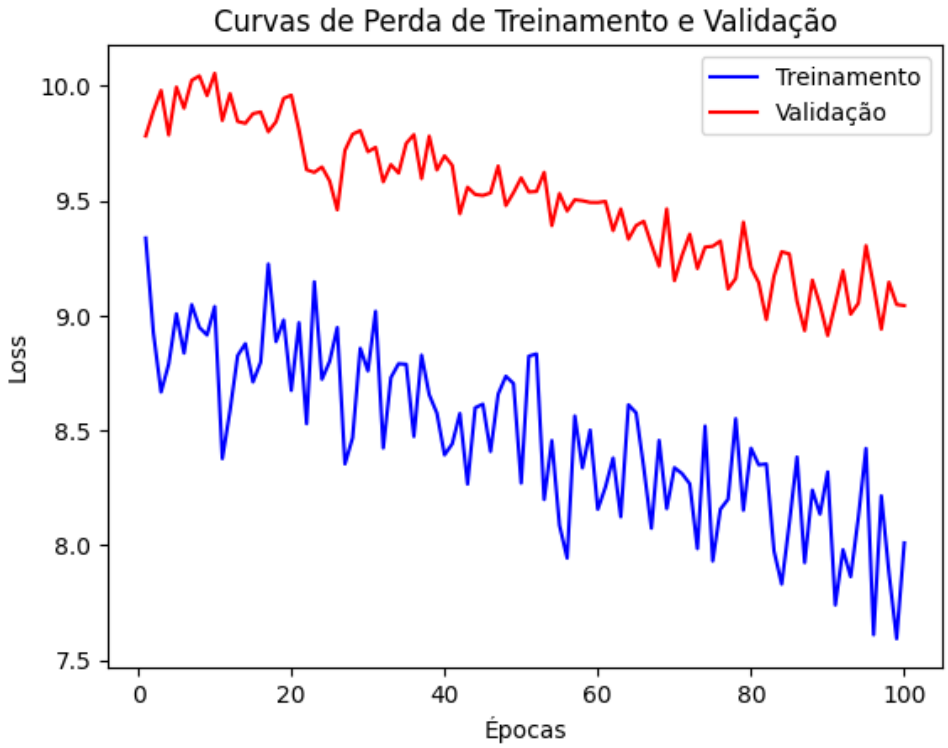

Figura 40 – Perda durante o treinamento, usando o otimizador Adam e taxa de aprendizado de 0.001, com conjunto de dados próprio. [Fonte: Autoria Própria].

melhorou o desempenho para estas imagens — os resultados com o modelo pré-treinado original são mostrados nas Figuras 47 e 48.

Modelo	IoU
Sem *fine-tuning*	0.172947
Com *fine-tuning*	0.473188

Tabela 11 – IoU médio calculado sobre imagens do conjunto de dados de teste, usando o modelo OFA Large.

Consideramos que os resultados do *fine-tuning* do modelo OFA para a identificação do sujeito principal foram satisfatórios. Foi possível observar um aumento significativo do índice de sobreposição (IoU), o que indica que o modelo melhorou sua capacidade de identificar com mais precisão o sujeito principal nas imagens após o treinamento especializado. A identificação do sujeito principal é um passo importante para o desenvolvimento de aplicações que envolvem a classificação e a organização de fotografias. No caso específico de fotografias de formatura, a identificação do sujeito principal pode ser utilizada para agrupar fotos de estudantes que estão juntos na mesma foto, por exemplo. Isso pode ser útil para a criação de álbuns de fotos, para a localização de fotos específicas e para a organização de fotos por turma. A abordagem proposta para a identificação do sujeito principal pode também ser utilizada para desenvolver uma variedade de aplicações que envolvem a classificação e a organização de fotografias.

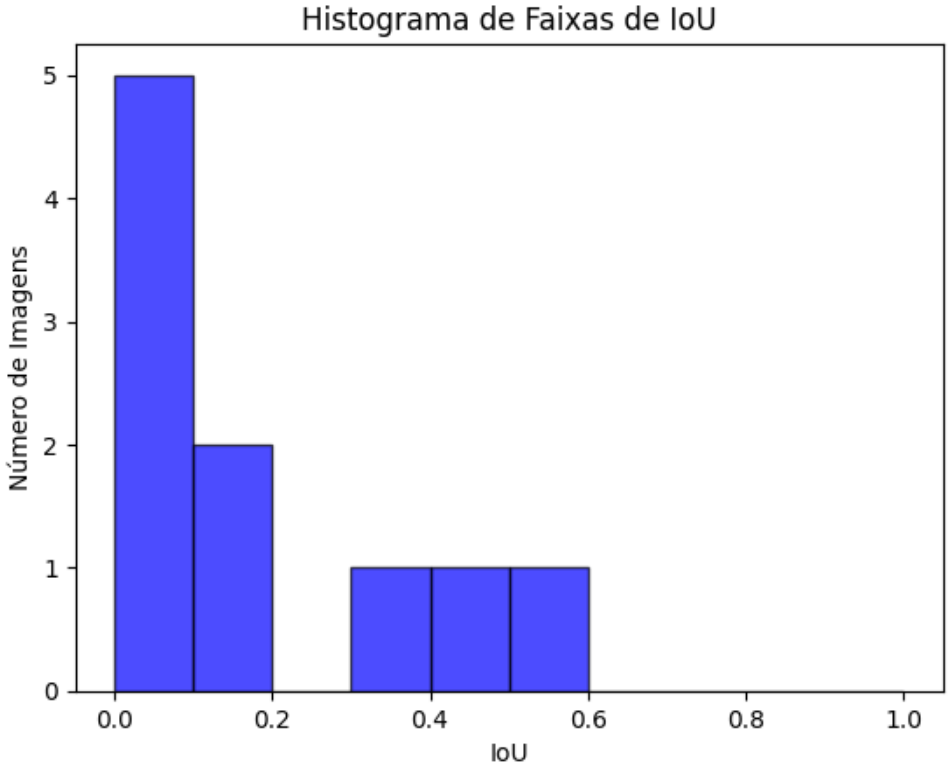

Figura 41 – Histograma mostrando a quantidade de imagens com IoU em diferentes faixas, antes do *fine-tuning*.

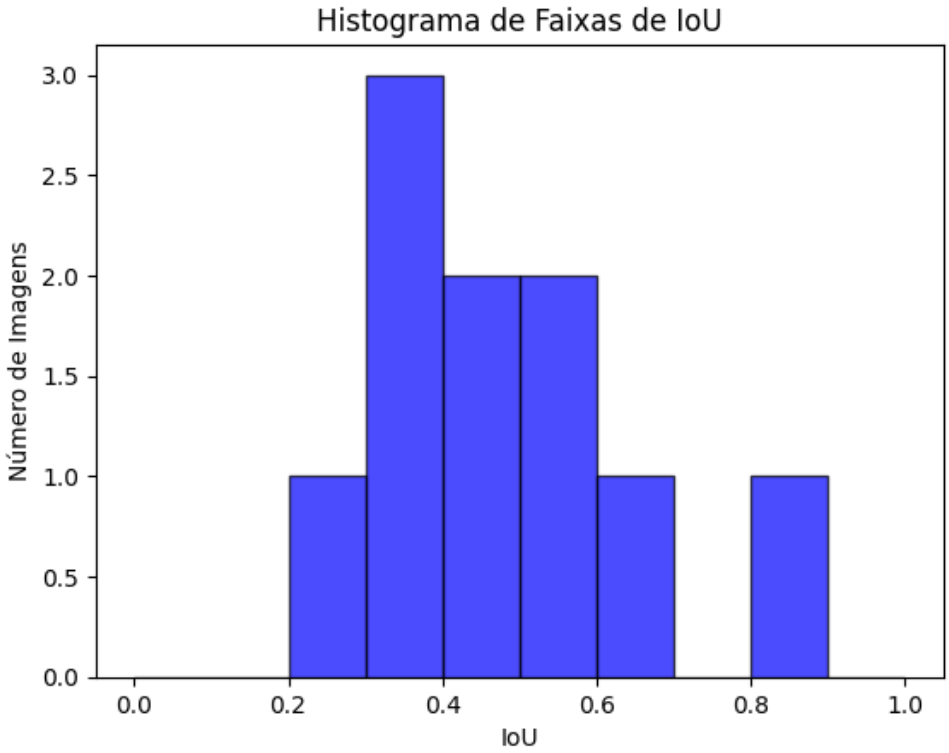

Figura 42 – Histograma mostrando a quantidade de imagens com IoU em diferentes faixas, após o *fine-tuning*.

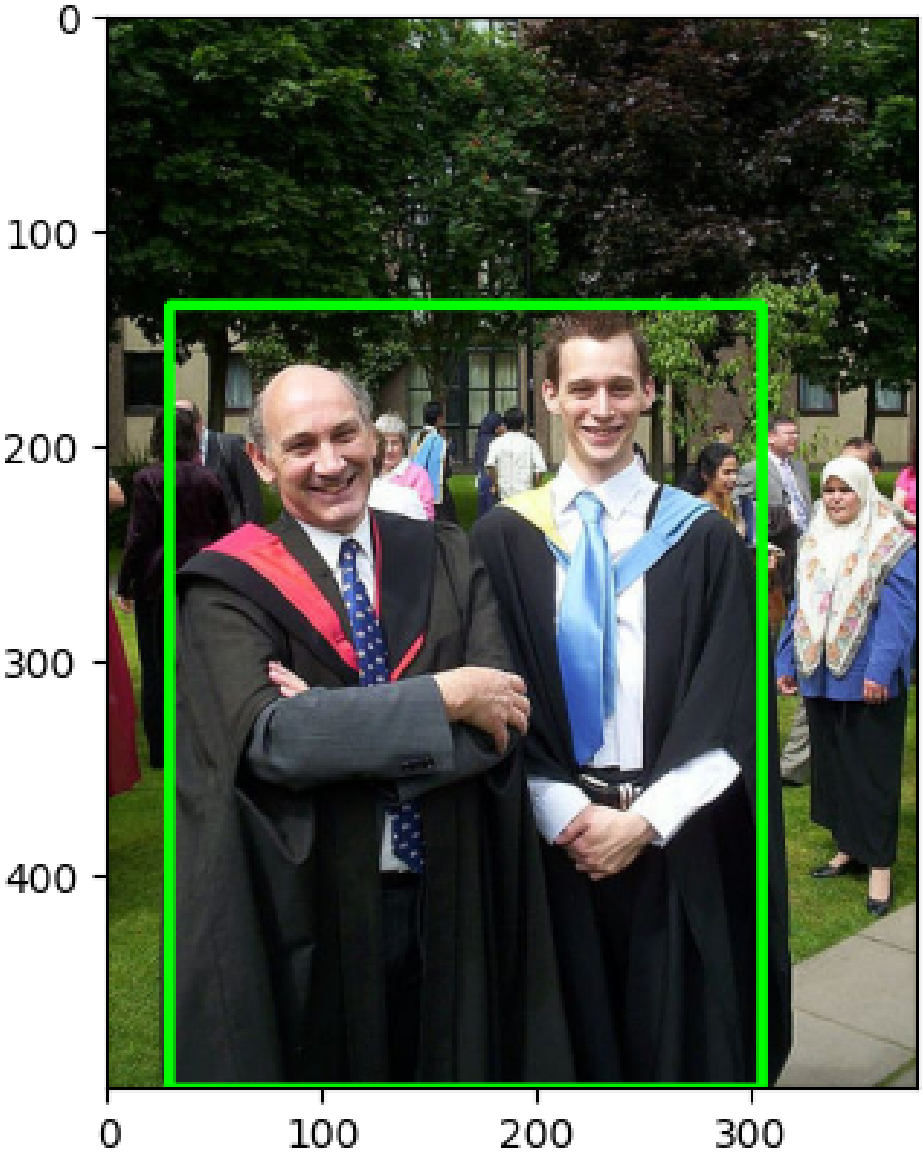

Figura 43 – Exemplo de uma fotografia e seu respectivo *main subject* anotado pelo OFA, com *fine-tuning*. O IoU obtido para esta imagem foi de IoU 0.67

4.3 CRIAÇÃO DE NUVEM DE PALAVRAS

A criação de nuvens de palavras é vista neste trabalho como uma tarefa secundária, que demonstra outra forma de explorar as legendas geradas automaticamente, além da identificação do sujeito principal. Não foi realizada uma avaliação quantitativa desta tarefa, visto que a qualidade dos resultados depende de uma análise subjetiva. Desta forma, foi realizada a criação somente para um conjunto de exemplo. Para criar a nuvem de palavras, empregamos o método explorado na Seção 3.4. Nesse processo, utilizamos um conjunto de 6800 legendas geradas pela etapa de legendagem do OFA. A nuvem de palavras obtida é ilustrada na Figura 49.

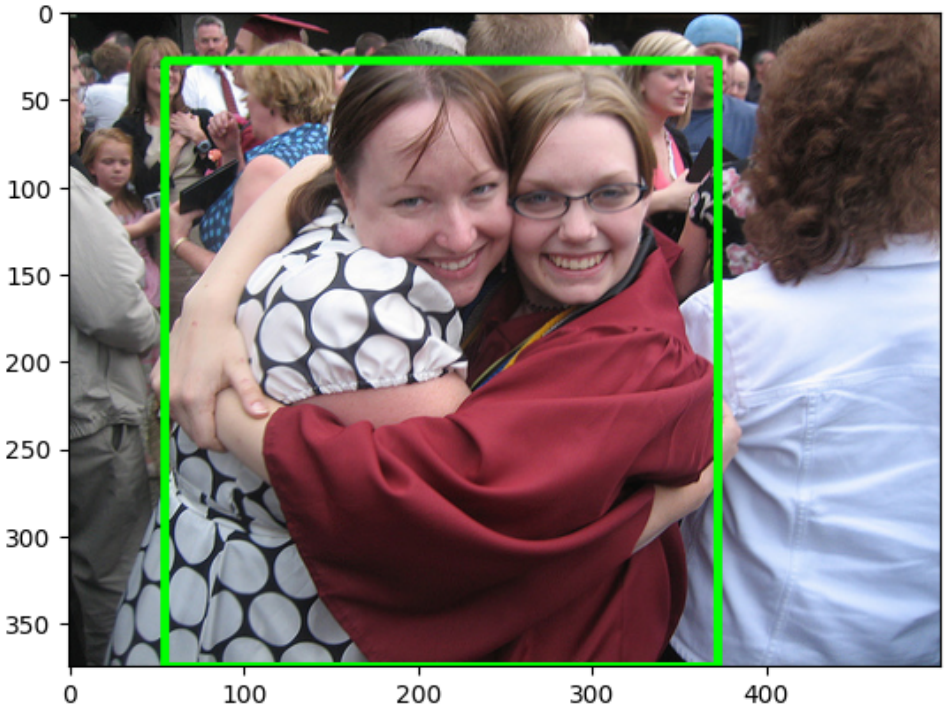

Figura 44 – Exemplo de uma fotografia, e seu respectivo *main subject* anotado pelo OFA, com *fine-tuning*. O IoU obtido para esta imagem foi de 0.7

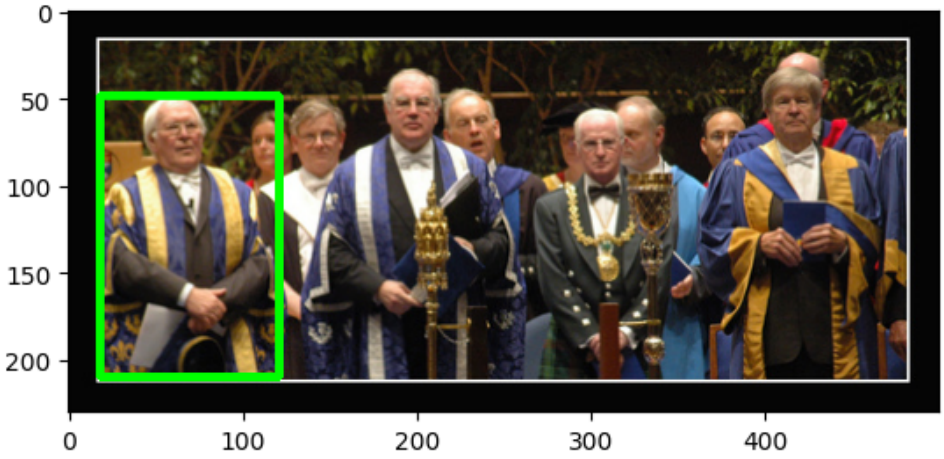

Figura 45 – Exemplo de uma fotografia, e seu respectivo *main subject* anotado pelo OFA, com *fine-tuning*. O IoU obtido para esta imagem foi de 0.23

Este processo permite uma análise visual das palavras mais relevantes presentes nas legendas, oferecendo *insights* iniciais sobre os tópicos e temas abordados nas legendas processadas. Além disso, a nuvem de palavras revelou-se uma ferramenta valiosa na compreensão da distribuição temática das legendas. A visualização proporcionada pela nuvem de palavras oferece uma perspectiva abrangente das principais palavras-chave presentes nas legendas, auxiliando na identificação de áreas de enfoque e ênfase. Essa representação visual poderia ser efetivamente empregada como um recurso de filtragem em uma plataforma de busca de imagens, facilitando

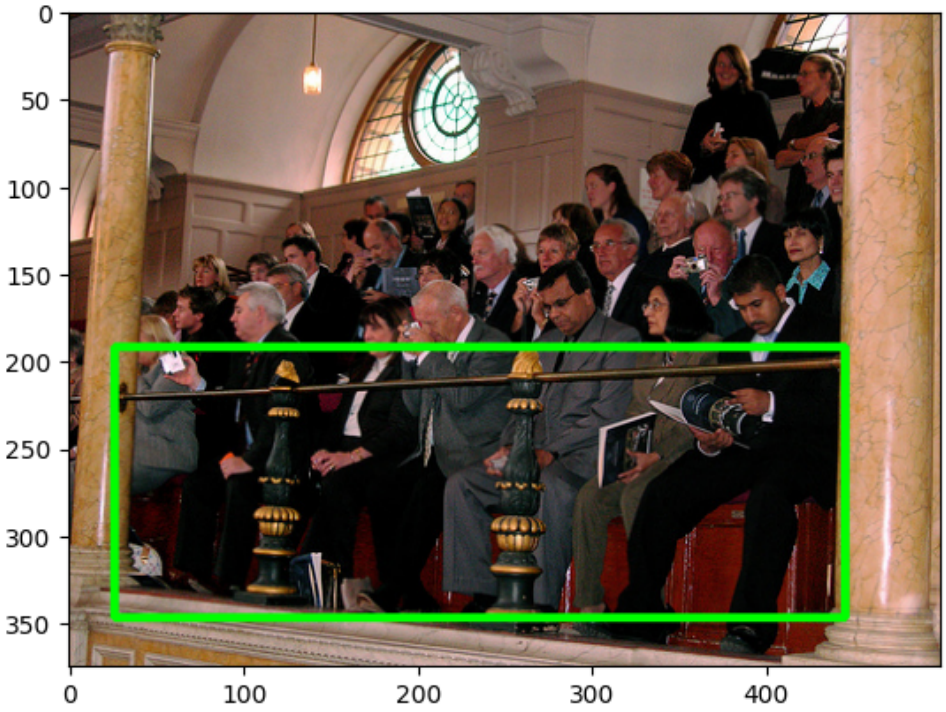

Figura 46 – Exemplo de uma fotografia, e seu respectivo *main subject* anotado pelo OFA, com *fine-tuning*. O IoU obtido para esta imagem foi de 0.39

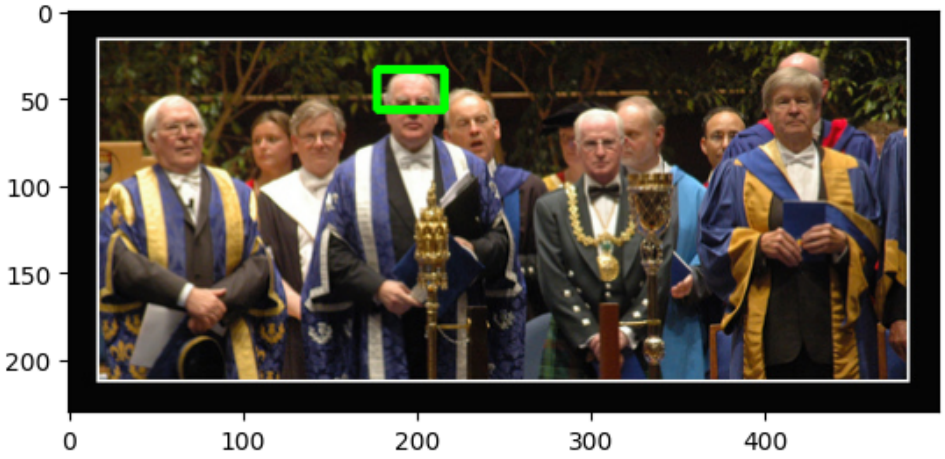

Figura 47 – Exemplo de uma fotografia, e seu respectivo *main subject* anotado pelo OFA, sem *fine-tuning*. O IoU obtido para esta imagem foi de 0.1

a localização rápida e precisa de imagens específicas entre um vasto conjunto de fotografias. Portanto, a utilização da nuvem de palavras demonstrou seu potencial não apenas como uma ferramenta analítica, mas também como um meio prático de otimizar a busca e seleção de imagens em plataformas de gerenciamento e compartilhamento de conteúdo fotográfico.

Figura 48 – Exemplo de uma fotografia, e seu respectivo *main subject* anotado pelo OFA, sem *fine-tuning*. O IoU obtido para esta imagem foi de 0.33

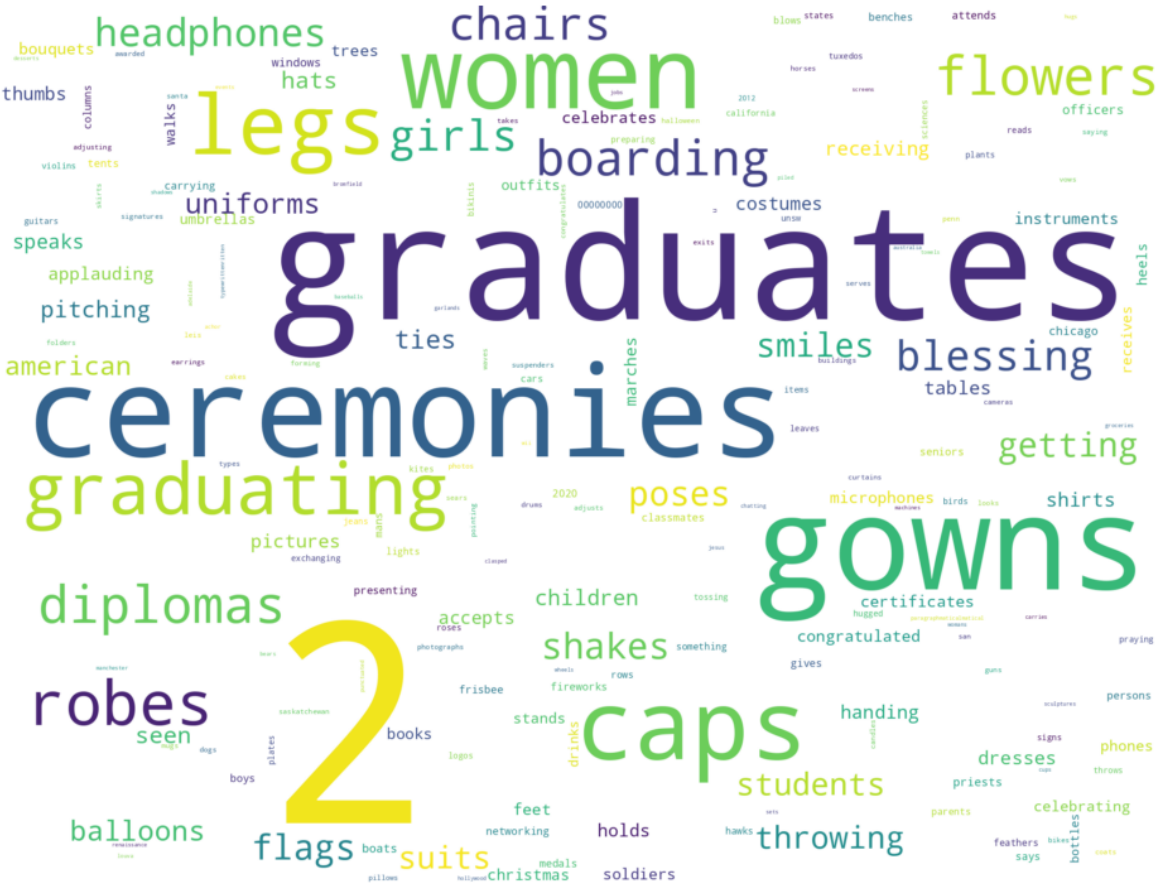

Figura 49 – Nuvem de palavras gerada a partir das legendas processadas.

5 CONCLUSÕES

Neste trabalho, abordamos o desafio da legendagem automática de imagens, com ênfase na descrição de fotografias de eventos de formatura, bem como a identificação do sujeito principal. O trabalho visa auxiliar a criação de álbuns de formatura, uma tarefa de grande importância, porém intensiva em termos de tempo e esforço humanos. A automação de partes desse processo surge como uma possibilidade promissora para acelerar o trabalho e reduzir custos. As tentativas iniciais de atribuir rótulos exclusivos a cada fotografia se revelaram insuficientes e ineficazes na organização das imagens de forma útil para os usuários humanos. Em contrapartida, a abordagem a partir da legendagem automática de imagens, empregando técnicas baseadas em *transformers*, emergiu como uma solução promissora para gerar descrições para imagens, levantando também a hipótese de que um modelo deste tipo já possui implicitamente o conhecimento necessário sobre quais seriam os sujeitos principais de uma fotografia.

Os modelos empregados na legendagem automática geralmente buscam compreender a interação entre objetos e cenários, com o objetivo fundamental de compreender como o mundo funciona e como os elementos interagem dentro de uma imagem. Assim, um modelo pré-treinado em um conjunto de dados como o MS-COCO pode ser adaptado facilmente ao contexto de formaturas por meio da técnica de transferência de aprendizado, na qual alguns dos pesos de um modelo pré-treinado são transferidos e reajustados utilizando diferentes imagens e classes. Além disso, exploramos o conhecimento implícito contido na camada de atenção dos modelos de legendagem automática para identificar o sujeito principal em uma imagem.

Neste trabalho, criamos um conjunto de dados de imagens, combinando dados dos conjuntos COCO e USED. Posteriormente, realizamos testes com diversas abordagens computacionais para descrever uma imagem, incluindo Redes Neurais Convolucionais para extração de características, camadas Long Short-Term Memory para geração de texto em linguagem natural, técnicas baseadas em atenção, e a arquitetura OFA, que se baseia em *transformers*. Em nossos experimentos, empregamos métricas como METEOR, BLEU, ROUGE-1, ROUGE-2 e ROUGE-L para validação. Concluímos que a abordagem OFA se destacou como a mais eficaz entre todas, com o ajuste fino o modelo descreveu com precisão a maioria das fotografias de formatura, frequentemente fornecendo detalhes específicos que são iguais ou superam descrições humanas.

Considerando o desempenho superior do modelo OFA na tarefa de legendagem, especi-

alizamos o mesmo para identificar o sujeito principal por meio da tarefa de *visual grounding*, obtendo medidas de IoU médias de 0,47, em comparação com 0,17 sem a especialização. Isso representa um progresso notável na organização eficiente de álbuns de formatura, economizando tempo e recursos humanos.

Adicionalmente, exploramos a utilidade das legendas geradas pelo modelo OFA na criação de nuvens de palavras, um recurso valioso para filtragem de fotografias, simplificando a organização de álbuns de formatura e facilitando a identificação de imagens relevantes para os usuários.

Em resumo, este estudo destaca o potencial das redes neurais pré-treinadas e das técnicas de aprendizado profundo na automação da legendagem automática de imagens e na identificação do sujeito principal, oferecendo benefícios significativos na organização de álbuns de formatura, uma tarefa de grande valor tanto para empresas do setor quanto para formandos. Também demonstramos que é possível utilizar o conhecimento implícito de uma rede para legendagem automática de imagens para identificar o sujeito principal de uma fotografia. Trabalhos futuros devem se concentrar na criação de ferramentas que explorem de outras formas as descrições extraídas das fotografias — por exemplo, permitindo o agrupamento de fotografias por contextos informados por um usuário, ou mesmo a recuperação de imagens a partir de *queries* em linguagem natural. Para isso, podem ser investigados outros modelos multimodais e multitarefa além do OFA, ou mesmo grandes modelos de linguagem.

REFERÊNCIAS

ANTOL, S.; AGRAWAL, A.; MITTAL, V.; NOROUZI, M.; SUKTHANKAR, R.; PARIKH, D. Vqa: Visual question answering. *In*: **Proceedings of the IEEE conference on computer vision and pattern recognition**. *[S.l.: s.n.]*, 2015. p. 2536–2544.

AUNG, San Pa Pa; PA, Win Pa; NWE, Tin Lay. Automatic Myanmar image captioning using CNN and LSTM-based language model. *In*: . Marseille, France: European Language Resources association, 2020. p. 139–143. ISBN 979-10-95546-35-1. Disponível em: https://aclanthology.org/2020.sltu-1.19.

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. **Neural Machine Translation by Jointly Learning to Align and Translate**. arXiv, 2014. Disponível em: https://arxiv.org/abs/1409.0473.

BESL, Paul J.; JAIN, Ramesh C. Three-dimensional object recognition using range data. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 11, n. 7, p. 752–769, 1988.

CHEN, Xinlei; FANG, Hao; LIN, Tsung-Yi; VEDANTAM, Ramakrishna; GUPTA, Saurabh; DOLLÁR, Piotr; ZITNICK, C. Lawrence. Microsoft COCO captions: Data collection and evaluation server. **ArXiv**, abs/1504.00325, 2015.

CHO, Kyunghyun; MERRIENBOER, Bart van; GULCEHRE, Caglar; BAHDANAU, Dzmitry; BOUGARES, Fethi; SCHWENK, Holger; BENGIO, Yoshua. **Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation**. arXiv, 2014. Disponível em: https://arxiv.org/abs/1406.1078.

CHUNG, Junyoung; GÜLÇEHRE, Çaglar; CHO, KyungHyun; BENGIO, Yoshua. Empirical evaluation of gated recurrent neural networks on sequence modeling. **CoRR**, abs/1412.3555, 2014. Disponível em: http://arxiv.org/abs/1412.3555.

DAI, Zihang; LIU, Hanxiao; LE, Quoc V.; TAN, Mingxing. **CoAtNet: Marrying Convolution and Attention for All Data Sizes**. arXiv, 2021. Disponível em: https://arxiv.org/abs/2106.04803.

DENG, Jia; DONG, Wei; SOCHER, Richard; LI, Li-Jia; LI, Kai; FEI-FEI, Li. Imagenet: A large-scale hierarchical image database. *In*: **2009 IEEE Conference on Computer Vision and Pattern Recognition**. *[S.l.: s.n.]*, 2009. p. 248–255.

DENKOWSKI, Michael; LAVIE, Alon. Meteor universal: Language specific translation evaluation for any target language. *In*: **Proceedings of the Ninth Workshop on Statistical**

Machine Translation. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. p. 376–380. Disponível em: https://aclanthology.org/W14-3348.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *In*: **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. *[S.l.: s.n.]*, 2018. v. 1, p. 4171–4186.

DONAHUE, Jeff; HENDRICKS, Lisa Anne; GUADARRAMA, Sergio; ROHRBACH, Marcus; VENUGOPALAN, Subhashini; SAENKO, Kate; DARRELL, Trevor. Long-term recurrent convolutional networks for visual recognition and description. *In*: IEEE. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. *[S.l.]*, 2014. p. 2625–2634.

DOSOVITSKIY, Alexey; BEYER, Lucas; KOLESNIKOV, Alexander; WEISSENBORN, Dirk; ZHAI, Xiaohua; UNTERTHINER, Thomas; DEHGHANI, Mostafa; MINDERER, Matthias; HEIGOLD, Georg; GELLY, Sylvain; USZKOREIT, Jakob; HOULSBY, Neil. An image is worth 16x16 words: Transformers for image recognition at scale. *In*: **9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021**. *[S.l.]*: OpenReview.net, 2021.

ELLIOTT, Desmond; KELLER, Frank. Image description using visual dependency representations. *In*: **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**. Seattle, Washington, USA: Association for Computational Linguistics, 2013. p. 1292–1302. Disponível em: https://aclanthology.org/D13-1128.

FARHADI, Ali; HEJRATI, Mohsen; SADEGHI, Mohammad Amin; YOUNG, Peter; RASHTCHIAN, Cyrus; HOCKENMAIER, Julia; FORSYTH, David. Every picture tells a story: Generating sentences from images. *In*: DANIILIDIS, Kostas; MARAGOS, Petros; PARAGIOS, Nikos (Ed.). **Computer Vision – ECCV 2010**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 15–29. ISBN 978-3-642-15561-1.

GUPTA, Abhinav; DAVIS, Larry S. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. *In*: FORSYTH, David; TORR, Philip; ZISSERMAN, Andrew (Ed.). **Computer Vision – ECCV 2008**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 16–29. ISBN 978-3-540-88682-2.

HASAN, Rakibul; CRANDALL, David; FRITZ, Mario; KAPADIA, Apu. Automatically detecting bystanders in photos to reduce privacy risks. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 16, n. 11, p. 2824–2838, 2021.

HE, Kaiming; ZHANG, X.; REN, Shaoqing; SUN, Jian. Deep residual learning for image recognition. **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, p. 770–778, 2016.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HSU, Ting-Yao; GILES, C Lee; HUANG, Ting-Hao. SciCap: Generating captions for scientific figures. *In*: **Findings of the Association for Computational Linguistics: EMNLP 2021**. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 3258–3264. Disponível em: https://aclanthology.org/2021.findings-emnlp.277.

JONES, Karen Sparck. A vector space model for information retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11–21, 1972.

KAMEOKA, Hirokazu; TANAKA, Kou; KWASNY, Damian; KANEKO, Takuhiro; HOJO, Nobukatsu. **ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion**. arXiv, 2018. Disponível em: https://arxiv.org/abs/1811.01609.

KAZEMZADEH, A. R.; DANCE, C. R.; COHN, J. F.; DARRELL, T.; SHEIKH, Y. Referit: A dataset for referring expression comprehension. *In*: **Proceedings of the IEEE conference on computer vision and pattern recognition**. *[S.l.: s.n.]*, 2014. p. 3029–3037.

KIROS, Ryan; SALAKHUTDINOV, Ruslan; ZEMEL, Richard. Unifying visual-semantic embeddings with multimodal neural language models. *In*: CURRAN ASSOCIATES INC. **Proceedings of the 27th International Conference on Neural Information Processing Systems**. *[S.l.]*, 2014. p. 3058–3066.

KRISHNA, S.; GUPTA, A.; MALIK, J. Visual genome: A large-scale dataset for visual relationship understanding. **International Journal of Computer Vision**, Springer, v. 123, n. 1, p. 32–60, 2017.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. *In*: PEREIRA, F.; BURGES, C. J. C.; BOTTOU, L.; WEINBERGER, K. Q. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2012. v. 25. Disponível em: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

L., Jane; ZHAI, Kevin Y.; ECHEVARRIA, Jose; FRIED, Ohad; HANRAHAN, Pat; LANDAY, James A. Dynamic guidance for decluttering photographic compositions. *In*: **The 34th Annual ACM Symposium on User Interface Software and Technology**. New York, NY, USA: Association for Computing Machinery, 2021. (UIST '21), p. 359–371. ISBN 9781450386357. Disponível em: https://doi.org/10.1145/3472749.3474755.

LAMB, Alex; GOYAL, Anirudh; ZHANG, Ying; ZHANG, Saizheng; COURVILLE, Aaron; BENGIO, Yoshua. **Professor Forcing: A New Algorithm for Training Recurrent Networks**. arXiv, 2016. Disponível em: https://arxiv.org/abs/1610.09038.

LI, Hong; ANDREAS, Jacob; KIELA, Douwe. Visualbert: A simple and performant baseline for vision and language. **arXiv preprint arXiv:1908.03557**, 2019.

LI, Li-Jia; FEI-FEI, Li. What, where and who? classifying events by scene and object recognition. **2007 IEEE 11th International Conference on Computer Vision**, p. 1–8, 2007.

LI, Li-Jia; SOCHER, Richard; FEI-FEI, Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. *In*: **2009 IEEE Conference on Computer Vision and Pattern Recognition**. *[S.l.: s.n.]*, 2009. p. 2036–2043. ISSN 1063-6919.

LI, Xirong; GAO, Yixing; LYU, Si; REN, Cairong; LIU, Qijie; HAN, Xiaoguang; XU, Jie; FENG, Jianfeng. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. **arXiv preprint arXiv:2007.01301**, 2020.

LI, Xiujun; YIN, Xi; LI, Chunyuan; HU, Xiaowei; ZHANG, Pengchuan; ZHANG, Lei; WANG, Lijuan; HU, Houdong; DONG, Li; WEI, Furu; CHOI, Yejin; GAO, Jianfeng. Oscar: Object-semantics aligned pre-training for vision-language tasks. **ECCV 2020**, 2020.

LIN, Chin-Yew. ROUGE: A package for automatic evaluation of summaries. *In*: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: https://aclanthology.org/W04-1013.

LIN, Tsung-Yi; MAIRE, Michael; BELONGIE, Serge; BOURDEV, Lubomir; GIRSHICK, Ross; HAYS, James; PERONA, Pietro; RAMANAN, Deva; ZITNICK, C. Lawrence; DOLLáR, Piotr. Microsoft COCO: Common objects in context. **ArXiv**, abs/1405.0312, 2015.

LIU, Yinhan; OTT, Myle; GOYAL, Naman; DU, Jingfei; JOSHI, Mandar; CHEN, Danqi; LEVY, Omer; LEWIS, Mike; ZETTLEMOYER, Luke; STOYANOV, Veselin. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

MAO, Junhua; XU, Wei; YANG, Yi; WANG, Jiang; HUANG, Zhiheng. Explain images with multimodal recurrent neural networks. *In*: IEEE. **Proceedings of the International Conference on Computer Vision**. *[S.l.]*, 2014. p. 1149–1157.

MILLER, George A. Wordnet: A lexical database for english. **Communications of the ACM**, ACM Press, v. 38, n. 1, p. 39–41, 1995.

PAPINENI, Kishore; ROUKOS, Salim; WARD, Todd; ZHU, Wei-Jing. Bleu: a method for automatic evaluation of machine translation. *In*: **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318. Disponível em: https://aclanthology.org/P02-1040.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. Glove: Global vectors for word representation. *In*: **EMNLP**. *[S.l.: s.n.]*, 2014. v. 14, p. 1532–1543.

RADFORD, Alec; WU, Jeff; CHILD, Rewon; LUAN, David; AMODEI, Dario; SUTSKEVER, Ilya. Improving language understanding by generative pre-training. *In*: **OpenAI**. *[S.l.: s.n.]*, 2018.

RADFORD, Alec; WU, Jeffrey; CHILD, Rewon; LUAN, David; AMODEI, Dario; SUTSKEVER, Ilya. Language models are unsupervised multitask learners. 2018. Disponível em: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

RAMPAL, Harshit; MOHANTY, Aman. Efficient CNN-LSTM based image captioning using neural network compression. **ArXiv**, abs/2012.09708, 2020.

SHEN, C.T.; LIU, J.C.; SHIH, S.W.; HONG, J.S. Towards intelligent photo composition-automatic detection of unintentional dissection lines in environmental portrait photos. **Expert Systems with Applications**, v. 36, n. 5, p. 9024–9030, 2009. ISSN 0957-4174. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417408009214.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *In*: **International Conference on Learning Representations**. *[S.l.: s.n.]*, 2015.

SU, Hao; GUO, Xiaodong; LI, Jianfeng; LU, Wei; HUANG, Ming Zhou; MITCHELL, Margaret. Vl-bert: Pre-training of generic visual-linguistic representations. *In*: **International Conference on Learning Representations**. *[S.l.: s.n.]*, 2019.

SUTSKEVER, Ilya; VINYALS, Oriol; LE, Quoc V. **Sequence to Sequence Learning with Neural Networks**. arXiv, 2014. Disponível em: https://arxiv.org/abs/1409.3215.

SZEGEDY, Christian; IOFFE, Sergey; VANHOUCKE, Vincent; ALEMI, Alexander A. Inception-v4, inception-resnet and the impact of residual connections on learning. *In*: **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**. *[S.l.]*: AAAI Press, 2017. (AAAI'17), p. 4278–4284.

SZEGEDY, Christian; VANHOUCKE, Vincent; IOFFE, Sergey; SHLENS, Jonathon; WOJNA, Zbigniew. **Rethinking the Inception Architecture for Computer Vision**. 2015.

TAN, Mingxing; BANSAL, Mohit. Lxmert: Learning cross-modality encoder representations from transformers. *In*: **International Conference on Computer Vision**. *[S.l.: s.n.]*, 2019. p. 7393–7403.

VAIOULIS, Fotios; POULOS, Marios; BOKOS, George. Main subject detection of image by cropping specific sharp area. **WSEAS Transactions on Information Science and Applications**, v. 2, 08 2005.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N; KAISER, Ł ukasz; POLOSUKHIN, Illia. Attention is all you need. *In*: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

VEDANTAM, Ramakrishna; ZITNICK, C. Lawrence; PARIKH, Devi. Cider: Consensus-based image description evaluation. **ArXiv**, abs/1411.5726, 2015.

VINYALS, Oriol; TOSHEV, Alexander; BENGIO, Samy; ERHAN, Dumitru. **Show and Tell: A Neural Image Caption Generator**. arXiv, 2014. Disponível em: https://arxiv.org/abs/1411.4555.

VINYALS, Oriol; TOSHEV, Alexander; BENGIO, Samy; ERHAN, Dumitru. Show and tell: A neural image caption generator. *In*: IEEE. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. *[S.l.]*, 2014. p. 3128–3137.

VIOLA, Paul; JONES, Michael. Rapid object detection using a boosted cascade of simple features. *In*: **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)**. *[S.l.: s.n.]*, 2001.

WANG, Cheng; YANG, Haojin; MEINEL, Christoph. Image captioning with deep bidirectional lstms and multi-task learning. **ACM Trans. Multimedia Comput. Commun. Appl.**, Association for Computing Machinery, New York, NY, USA, v. 14, n. 2s, apr 2018. ISSN 1551-6857. Disponível em: https://doi.org/10.1145/3115432.

WANG, Peng; YANG, An; MEN, Rui; LIN, Junyang; BAI, Shuai; LI, Zhikang; MA, Jianxin; ZHOU, Chang; ZHOU, Jingren; YANG, Hongxia. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. **CoRR**, abs/2202.03052, 2022.

XU, Kelvin; BA, Jimmy; KIROS, Ryan; CHO, Kyunghyun; COURVILLE, Aaron C.; SALAKHUTDINOV, Ruslan; ZEMEL, Richard S.; BENGIO, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. **CoRR**, abs/1502.03044, 2015. Disponível em: http://dblp.uni-trier.de/db/journals/corr/corr1502.html#XuBKCCSZB15.

XU, Kelvin; BA, Jimmy; KIROS, Ryan; CHO, Kyunghyun; COURVILLE, Aaron; SALAKHUDINOV, Ruslan; ZEMEL, Rich; BENGIO, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. *In*: BACH, Francis; BLEI, David (Ed.).

Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR, 2015. (Proceedings of Machine Learning Research, v. 37), p. 2048–2057. Disponível em: https://proceedings.mlr.press/v37/xuc15.html.

YANG, Zhilin; DAI, Zihang; YANG, Yiming; CARBONELL, Jaime; SALAKHUTDINOV, Ruslan; LE, Quoc. Xlnet: Generalized autoregressive pretraining for language understanding. **arXiv preprint arXiv:1906.08237**, 2019.

ZHANG, Botong; TIAN, Lihua; LI, Chen; YANG, Yi. Heatmap and edge guidance network for salient object detection. **Computers and Electrical Engineering**, v. 105, p. 108525, 2023. ISSN 0045-7906. Disponível em: https://www.sciencedirect.com/science/article/pii/S0045790622007406.