INSTITUTO FEDERAL DE BRASÍLIA

PRÓ-REITORIA DE PESQUISA E INOVAÇÃO

ANEXO I - MODELO DE PROJETO – PIBITI

Desenvolvendo um sistema de legenda para imagens usando técnicas de aprendizado profundo

Campus: Taguatinga
1.03.00.00-7 Ciência da Computação
1.03.03.04-9 Sistemas de Informação

Brasília
2023

Caracterização do produto existente no ambiente produtivo ou social com desenvolvimento tecnológico e/ou inovação proposto(a) no projeto

A geração de uma descrição de uma imagem é chamada de legendagem de imagem. A legendagem de imagens requer o reconhecimento de objetos importantes, seus atributos e seus relacionamentos em uma imagem. Também faz-se necessário gerar frases semanticamente corretas.

Recentemente, a legenda de imagens está se tornando um assunto ativo de pesquisa, atraindo grande interesse dos pesquisadores. Isso decorre do avanço de dois campos amplamente pesquisados em Inteligência Artificial: Visão Computacional e Processamento de Linguagem Natural.

Esta tarefa pode ajudar em uma série de aplicações práticas, incluindo motores de busca, tarefas de otimização ou recuperação de imagens, entre outras. Apesar do grande número de casos de uso prático, a tarefa é complexa.

Técnicas baseadas em aprendizado profundo são capazes de lidar com as complexidades e desafios da legendagem de imagens. Neste projeto de pesquisa, pretendemos apresentar uma revisão inicial de técnicas existentes de legendagem de imagens baseadas em aprendizado profundo. Apresentaremos os fundamentos das técnicas para analisar seus desempenhos, pontos fortes e limitações. Também abordaremos os conjuntos de dados e as métricas de avaliação popularmente usados na geração automática de legendas para imagens baseadas em aprendizagem profunda.

Compreender uma imagem (basicamente componentes e seus atributos) depende em grande parte da obtenção de características da imagem. Plataformas de redes sociais (Twitter, Facebook) permitem que seus usuários postem imagens e, em boa parte dessas publicações, os usuários podem marcá-las com algum descritor: local onde estão (cafeteria, praia, escola, por exemplo), o que estão fazendo, como estão vestidos, entre outros. Tais informações podem ser úteis para a compreensão de uma imagem mas nem sempre estão disponíveis.

As características de uma imagem podem ser aprendidas automaticamente a partir de dados de treinamento e podem lidar com um conjunto grande e diversificado de imagens e vídeos. Essa é a proposta das técnicas baseadas em aprendizado de máquina profunda. Por exemplo, redes neurais convolucionais (CNNs) [1] são amplamente usadas para aprendizado de recursos, e um classificador como Softmax é usado para classificação. A CNN é geralmente seguida por redes recorrentes (RNNs) para gerar legendas.

Nos últimos 5 anos, um grande número de artigos foram publicados sobre legendagem de imagens, com o uso do aprendizado de máquina profundo, que podem lidar com complexidades e desafios de legendagem de imagem. Várias classificações das estratégias e técnicas foram propostas. Usaremos a proposta em [2] para descrever melhor essas categorias:

- **Baseado em espaço visual** - Nos métodos baseados no espaço visual, os recursos da imagem e as legendas correspondentes são passadas independentemente para o decodificador de linguagem.
- **Baseado no espaço multimodal** - A arquitetura de um método baseado em espaço multimodal típico contém uma parte de codificador de linguagem, uma parte de visão, uma parte de espaço multimodal e uma parte de decodificador de linguagem. A parte de visão usa uma rede neural convolucional profunda como um extrator para obter características da imagem. A parte do codificador de linguagem extrai e aprende características de cada palavra. Em seguida, encaminha o contexto temporal semântico para as camadas recorrentes.
O mapa resultante é então passado para o decodificador de linguagem, que gera legendas decodificando o mapa. Os métodos nesta categoria seguem as seguintes etapas:

    (1) Redes neurais profundas e o modelo de linguagem neural multimodal são usados para aprender ambas imagem e texto em um espaço multimodal.

    (2) A parte de geração de idioma gera legendas usando as informações da Etapa 1.

- **Aprendizado não supervisionado** - Técnica que lida com dados não rotulados. Redes adversárias generativas (GANs) [3] são um tipo de técnicas de aprendizado não supervisionadas. O aprendizado por reforço é outro tipo de abordagem de aprendizado de máquina em que os objetivos de um agente são descobrir dados e/ou rótulos por meio da exploração e um sinal de recompensa. Vários métodos de legendagem de imagens usam aprendizado por reforço e abordagens baseadas em GAN.

- **Legendagem densa** - Johnson e outros [4] propuseram um método de legendagem de imagens chamado DenseCap. Este método localiza todas as regiões importantes de uma imagem e então gera descrições para essas regiões. Um método típico desta categoria tem as seguintes etapas:

    (1) A imagem é dividida em diferentes regiões.

    (2) CNN é usado para obter características da imagem com base na região.

    (3) As saídas da Etapa 2 são usadas por um modelo de linguagem para gerar legendas para cada região.

- **Baseado em arquitetura composicional** - Método composto por vários blocos de construção funcionais independentes: primeiro, uma CNN é usada para extrair os conceitos semânticos da imagem; em seguida, um modelo de linguagem é usado para gerar um conjunto de legendas de

candidatos; ao gerar a legenda final, essas legendas candidatas são reclassificadas usando um modelo de similaridade multimodal profunda. Um método típico desta categoria contém os seguintes passos:

(1) As características da imagem são obtidas usando uma CNN.

(2) Conceitos visuais (por exemplo, atributos) são obtidos a partir das características visuais.

(3) Múltiplas legendas são geradas por um modelo de linguagem usando as informações da Etapa 1 e 2.

(4) As legendas geradas são reclassificadas usando um modelo de similaridade multimodal profunda para selecionar legendas de imagem de alta qualidade.

- **Baseado em atenção** - Um método típico desta categoria adota os seguintes passos:

(1) As informações da imagem são obtidas com base em toda a cena por uma CNN.

(2) A fase de geração de linguagem gera palavras ou frases com base na saída da Etapa 1.

(3) Regiões salientes da imagem dada são focadas em cada passo de tempo do modelo de geração da linguagem baseada em palavras ou frases geradas.

(4) As legendas são atualizadas dinamicamente até o estado final do modelo de geração de linguagem.

- **Baseado em conceito semântico** - seletivamente buscam atender a um conjunto de propostas de conceitos semânticos extraídos da imagem. Esses conceitos são então combinados em estados ocultos e as saídas de redes neurais recorrentes.

Os métodos nesta categoria seguem as seguintes etapas:

(1) O codificador baseado em CNN é usado para codificar os recursos da imagem e os conceitos semânticos.

(2) Características de imagem são inseridos na entrada do modelo de geração de linguagem.

(3) Conceitos semânticos são adicionados aos diferentes estados ocultos do modelo de linguagem.

(4) A parte de geração de linguagem produz legendas com conceitos semânticos.

De uma maneira geral, ao se observar uma cena, seres humanos costumam focar, de forma inconsciente, em pessoas, animais e, por último, em objetos presentes na imagem. A etapa de processamento de imagens deve detectar múltiplos objetos/cenas. A etapa seguinte envolve processamento de linguagem natural para construção de legendas e deve estar centrado na atenção. A natureza multimodal da tarefa coloca pressão sobre o processo de treinamento de redes neurais. As redes neurais precisam, não apenas aprender o processo de extração de

características da imagem, mas também codificar heurísticas de linguagem no processo de geração que produz a legenda. Bases de imagens e métricas são fundamentais para validação das abordagens.

Vários conjuntos estão disponíveis para treinar, testar e avaliar os métodos de legendagem de imagens. Os conjuntos de dados diferem em várias perspectivas, como quantidade de imagens, quantidade de legendas por imagem, formato das legendas e tamanho da imagem. Três conjuntos de dados, Flickr8K [5], Flickr30K [6] e MS COCO [7], são usados popularmente

Diversas métricas de avaliação podem ser usadas para medir a qualidade das legendas geradas. Cada métrica aplica sua própria técnica de computação e tem vantagens distintas. Os métodos existentes de legendagem de imagens calculam pontuações de log-verossimilhança para avaliar sua geração de legendas. As métricas mais usadas são: BLEU [8], METEOR [9], ROUGE [10], SPICE [11] e CIDEr [12]. No entanto, BLEU, METEOR e ROUGE não estão bem correlacionados com avaliações humanas de qualidade. SPICE e CIDEr têm melhor correlação, mas são difíceis de otimizar. Liu et ai. [13] introduziu uma nova métrica de avaliação de legenda que é uma boa escolha para avaliadores humanos. É desenvolvida através de uma combinação de SPICE e CIDEr, e foi denominada como SPIDEr [13]. Ela usa um método de gradiente de política para otimizar as métricas.

Justificativa em Desenvolvimento Tecnológico e Inovação

Este projeto pretende implementar uma solução de atribuição de legendas para vídeos a partir do reconhecimento da cena, dos objetos importantes presentes na cena, seus atributos e relacionamentos entre si. A construção desse protótipo envolverá as seguintes etapas:

- estudo das tecnologias envolvidas - visão computacional, processamento de linguagem natural e aprendizagem de máquina;

- proposição e implementação de um modelo; e

- validação e análise dos resultados.

Objetivo geral relacionado ao desenvolvimento tecnológico e/ou inovação projetado(a)

O objetivo deste projeto é desenvolver um sistema de legendagem de imagens usando técnicas de aprendizado profundo, usando técnicas de visão computacional, reconhecimento de imagem e processamento de linguagem natural. Entre as possíveis aplicações, podemos citar: auxiliar pessoas com deficiência, melhorar a compreensão de imagens em sistemas de inteligência

artificial e melhorar a experiência do usuário em mídias sociais e plataformas de comércio eletrônico.

Podemos apresentar as seguintes questões de pesquisa:

- "Como as técnicas de aprendizado profundo podem ser usadas de forma eficaz para gerar legendas de imagem precisas e contextualmente relevantes?" e
- "Quais são as abordagens mais eficazes para integrar o processamento de imagem e as técnicas de processamento de linguagem natural em um sistema de legendagem de imagem?"

Objetivos específicos relacionados ao desenvolvimento tecnológico e/ou inovação projetado(a)

Os objetivos específicos do projeto são:

- Estudo das técnicas de aprendizagem profundo;
- Estudo das técnicas de visão computacional e reconhecimento de imagens;
- Estudo das técnicas de processamento de linguagem natural;
- Proposta de uma estratégia para legendagem de imagens combinando visão computacional, reconhecimento de imagem e processamento de linguagem natural;
- Implementação de um protótipo;
- Avaliação do uso do protótipo.

- Material e Métodos

Este projeto será desenvolvido por um aluno do curso superior ABI em Computação, sob orientação do prof. Raimundo Claudio da Silva Vasconcelos.

O projeto será realizado no campus de Taguatinga do Instituto Federal de Brasília.

Não serão necessários recursos materiais para o desenvolvimento do projeto, além dos equipamentos já disponíveis nos laboratórios de informática da instituição. No laboratório será utilizado um computador desktop com sistema operacional Windows e/ou Linux.

O aluno irá fazer fichamento da bibliografia inicial sobre o assunto a fim de conhecer a as técnicas envolvidas para a realização da tarefa: aprendizado profundo, visão computacional, linguagem natural, estudo de propostas similares em identificar objetos em uma imagem e sua descrição em sentenças semântica e sintaticamente corretas, proposta de uma estratégia, implementação e avaliação de uma solução.

- Principais contribuições para o desenvolvimento tecnológico e/ou inovação proposta no projeto:

As principais contribuições deste trabalho são:

- Identificação de técnicas de visão computacional e reconhecimento de objetos em imagens;
- Análise das técnicas de aprendizagem profunda mais adequadas para esta identificação;

- Seleção de técnicas de processamento de linguagem natural para geração de legendas;
- Descrição das técnicas de aprendizagem profunda mais adequadas para a geração de legendas;
- Seleção do banco de imagens;
- Definição e escolha de métricas para validação;
- Desenvolvimento de uma solução para geração de legendas para imagens usando aprendizado profundo;
- Construção do protótipo para geração de legendas de imagens;
- Validação e análise dos resultados.

- Cronograma de execução do projeto:

A seguir estão descritos no tempo as atividades a serem realizadas durante o desenvolvimento deste projeto.

Atividades	**Meses**											
	1	2	3	4	5	6	7	8	9	10	11	12
Levantamento da bibliografia inicial	X	X	X									
Estudo das técnicas existentes para identificação de objetos em imagens		X	X	X								
Estudo das técnicas existentes para geração de legendas			X	X	X							
Estudo das técnicas de aprendizagem profunda	X	X	X	X	X	X						
Estudo das bases de imagens e métricas para validação					X	X	X					
Construção do modelo						X	X	X				
Implementação do protótipo							X	X	X	X		
Teste e análise dos resultados									X	X	X	X
Escrita e submissão de artigo											X	X

- Plano de trabalho do bolsista

O bolsista irá desenvolver as seguintes atividades:
- leitura e fichamento da bibliografia inicial;
- estudo das técnicas existentes para identificação de objetos em imagens;
- estudo das técnicas existentes para geração de legendas;

- estudo das técnicas de aprendizagem profunda mais adequadas para identificação de objetos em imagens e geração de legendas;
- descoberta e estudo de bases de imagens disponíveis;
- estudo das métricas usadas;
- construção do modelo;
- implementação;
- escrita e submissão de um artigo.

Referências

[1] Yann LeCun, LÃl';on Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 11 (1998), 2278–2324.

[2] Hossain, MD. Zakir and Sohel, Ferdous and Shiratuddin, Mohd Fairuz and Laga, Hamid. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 51, 6, Article 118 (November 2019), 36 pages. https://doi.org/10.1145/3295748.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Advances in Neural Information Processing Systems. 2672–2680.

[4] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4565–4574.

[5] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research 47 (2013), 853–899.

[6] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision. 2641–2649.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft Coco: Common objects in context. In European Conference on Computer Vision. Springer, 740–755.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 311–318.

[9] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Vol. 29. 65–72.

[10] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Vol. 8.

[11] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision. Springer, 382–398.

[12] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4566–4575.

[13] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In Proceedings of the IEEE International Conference on Computer Vision (ICCV'17), Vol. 3. 873–881.