Tópicos Avançados em Visão Computacional

Jorge Augusto Salgado Salhani
Leonardo Rossi Luiz
Pedro Dias Batista

I. Introdução

Diversos termos vinculados às áreas científicas são utilizados ao longo do tempo por grandes veículos de mídia como forma de direcionar a atenção geral para determinado avanço tecnológico, em particular àqueles apresentam maior potencial de mudança de paradigma na sociedade. Ao fim dos anos 1970, na era pós moderna,[7] eventos como o lançamento de computadores pessoais [3] e smartphones [1] são notoriamente característicos pela cobertura jornalística massiva e especulações sobre os seus efeitos na sociedade. Até mesmo a invenção do mouse, que constitui periférico cotidiano para usuários de computadores, advém de idealizações de interfaces usuário-máquina que pudessem aprimorar o intelecto humano. [10]

Atualmente o conceito de aprendizado de máquina e redes neurais têm ganhado espaço em noticiários, em particular devido ao processamento de linguagem realizado por modelos como o ChatGPT, da OpenAI, [14] e na sua utilização como ferramenta central para o funcionamento de veículos autônomos, por meio de detecção de objetos e de segmentação semântica (quais pixels da imagem capturada pela câmera do carro constituem uma pessoa, por exemplo) para reconhecimento simultâneo e em tempo real de estradas, pessoas e placas de trânsito, por exemplo. [12]

Vale ressaltar que, embora não tão amplamente divulgadas quanto em contextos mencionados acima, estas mesmas tecnologias são fundamentais para diversas outras áreas. Podemos citar o reconhecimento facial como biometria, [8] sistemas de recomendações de conteúdos [15], tradutores de texto [2] e detecção de doenças, como câncer de pulmão. [11]

Com base nesse contexto, neste trabalho[1] apresentamos o desenvolvimento de redes neurais sob a perspectiva do campo da visão computacional [9] na realização de tarefas de geração de legendas, também denominado Image Captioning. [5] As seções serão apresentadas como segue: em **Baseline e dataset** apresentamos os modelos que serão utilizados como base para a construção das primeiras versões da nossa rede neural, assim como as bases de dados que utilizaremos para treinamento e teste; em **Resultados iniciais** explicamos os resultados obtidos na implementação deste primeiro modelo, com destaque às estruturas da rede que serão modificados em análises futuras; já em **Metodologia** apresentamos o modelo proposto para aprimorar os resultados de base que obtivemos; subsequentemente, em **Resultados** destacamos os ganhos e perdas obtidos com as alterações do modelo de base utilizado; por fim, em **Conclusão** retomamos os detaques ao longo do projeto e motivações para novos estudos.

[1] Código disponível em: https://github.com/jorgesalhani/TopicsVisComp

II. Baseline e dataset

Algumas das abordagens propostas para a resolução de problemas vinculados à rotulação automática (captioning) de imagens são construídas a partir de modelos de machine learning (ML) tradicionais, com extração de atributos (features) e subsequente classificação (por exemplo, por meio de support vector machines - SVM), ou utilizando redes neurais convolucionais (convolutional neural networks - CNN). [5, 13] Uma vez que o uso de CNNs para a resolução de problemas desta categoria é frequente, assim como sua ampla utilização para demais campos vinculados à classificação de imagens, optamos por utilizar deste modelo para a nossa proposta.

Em resumo, CNNs são constituídas de cadeias de processamentos realizados sobre uma amostra de imagens. Como as componentes responsáveis por cada etapa de processamento são essencialmente funções, seu reposicionamento ou adição de novas camadas (i.e. novas componentes) é bastante flexível, embora não completamente livre. Algumas das camadas mais utilizadas são convoluções, subamostragem (pooling), funções de ativação e camadas densas. [9]

Após a extração de features de uma imagem por meio de CNNs, é necessário que um texto com coerência semântica possa ser produzido em função da imagem analisada. Nesta parte, geralmente são utilizados modelos de redes neurais recorrentes (Recurrent Neural Networks - RNNs), em particular arquiteturas com memória longa chamadas Long Short Term Memory (LSTM). [5, 6, 13]

Como baseline optamos pelas arquiteturas VGGNet16[2] e ResNetV2[3] como pré-treino utilizando o dataset ImageNet[4]. Já para testes, utilizaremos datasets MS COCO[5] e Flickr30k[6] [4, 6]

[2]Disponível em: https://keras.io/api/applications/vgg/

[3]Disponível em: https://keras.io/api/applications/resnet/

[4]Disponível em: https://www.image-net.org/

[5]Disponível em: https://cocodataset.org/

[6]Disponível em: https://shannon.cs.illinois.edu/DenotationGraph/

Referências

[1] Ben Agger. ≪iTime: Labor and life in a smartphone era≫. Em: *Time & Society* 20.1 (2011), pp. 119–136. DOI: 10.1177/0961463X10380730. eprint: https://doi.org/10.1177/0961463X10380730. URL: https://doi.org/10.1177/0961463X10380730.

[2] Michael Auli. ≪Joint Language and Translation Modeling with Recurrent Neural Networks≫. Em: *Proc. of EMNLP*. Out. de 2013. URL: https://www.microsoft.com/en-us/research/publication/joint-language-and-translation-modeling-with-recurrent-neural-networks/.

[3] Brian Cogan. ≪"Framing usefulness:" An examination of journalistic coverage of the personal computer from 1982–1984≫. Em: *Southern Communication Journal* 70.3 (2005), pp. 248–265. DOI: 10.1080/10417940509373330. eprint: https://doi.org/10.1080/10417940509373330. URL: https://doi.org/10.1080/10417940509373330.

[4] Jiuxiang Gu. ≪An Empirical Study of Language CNN for Image Captioning≫. Em: out. de 2017, pp. 1231–1240. DOI: 10.1109/ICCV.2017.138.

[5] MD. Zakir Hossain. ≪A Comprehensive Survey of Deep Learning for Image Captioning≫. Em: *ACM Comput. Surv.* 51.6 (fev. de 2019). ISSN: 0360-0300. DOI: 10.1145/3295748. URL: https://doi.org/10.1145/3295748.

[6] S. Kalra. ≪Survey of convolutional neural networks for image captioning≫. Em: *Journal of Information and Optimization Sciences* 41.1 (2020), pp. 239–260. DOI: doi:10.1080/02522667.2020.1715602.

[7] Douglas Kellner. ≪Postmodernism as Social Theory: Some Challenges and Problems≫. Em: *Theory, Culture & Society* 5.2-3 (1988), pp. 239–269. DOI:

10.1177/0263276488005002003. eprint: https://doi.org/10.1177/0263276488005002003. URL: https://doi.org/10.1177/0263276488005002003.

[8] Relly Victoria Petrescu. ≪Face Recognition as a Biometric Application≫. Em: *Journal of Mechatronics and Robotics* 3 (2019), pp. 237–257. URL: http://dx.doi.org/10.2139/ssrn.3417325.

[9] Moacir Antonelli Ponti. ≪Everything You Wanted to Know about Deep Learning for Computer Vision but Were Afraid to Ask≫. Em: *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*. 2017, pp. 17–41. DOI: 10.1109/SIBGRAPI-T.2017.12.

[10] Albrecht Schmidt. ≪Augmenting Human Intellect and Amplifying Perception and Cognition≫. Em: *IEEE Pervasive Computing* 16.1 (2017), pp. 6–10. DOI: 10.1109/MPRV.2017.8.

[11] Fatma Taher. ≪Lung cancer detection by using artificial neural network and fuzzy clustering methods≫. Em: *2011 IEEE GCC Conference and Exhibition (GCC)*. 2011, pp. 295–298. DOI: 10.1109/IEEEGCC.2011.5752535.

[12] Yu-Ho Tseng. ≪Combination of computer vision detection and segmentation for autonomous driving≫. Em: *2018 IEEE/ION Position, Location and Navigation Symposium (PLANS)*. 2018, pp. 1047–1052. DOI: 10.1109/PLANS.2018.8373485.

[13] H. Wang. ≪An Overview of Image Caption Generation Methods.≫ Em: *Computational intelligence and neuroscience* (2020), pp. 1–13. DOI: https://doi.org/10.1155/2020/3062706.

[14] Tianyu Wu. ≪A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development≫. Em: *IEEE/CAA Journal of Automatica Sinica* 10.5 (2023), pp. 1122–1136. DOI: 10.1109/JAS.2023.123618.

[15] Fei Yu. ≪Network-based recommendation algorithms: A review≫. Em: *Physica A: Statistical Mechanics and its Applications* 452 (2016), pp. 192–208. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2016.02.021. URL: https://www.sciencedirect.com/science/article/pii/S0378437116001874.