

Geração de Legendas de Imagens em Português Utilizando Deep Learning

Bruno dos Anjos, Ravi Figueiredo,
Richardson Andrade

26 de julho de 2022

Sumário

1 Introdução

- Motivação
- Objetivo

2 Método

- Pipelines
- Base de Dados
- Pré Processamento

3 Resultados

4 Conclusão

5 Referências

Motivação

- Existem muitas fontes de imagens que não contém descrições. Mesmo que seja uma atividade fácil para o humano, é uma tarefa complexa para a máquina.



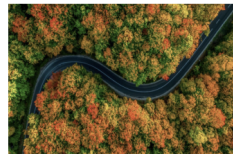
A politician receives a gift from politician.



A collage of different colored ties on a white background.



Silhouette of a woman practicing yoga on the beach at sunset.



Aerial view of a road in autumn.

Motivação II

- Nos campos, de visão computacional e linguagem natural, a descrição automática de imagens é uma área com limitações e desafios abertos.
- Maioria dos conjuntos de dados na literatura são apenas com legendas em inglês.
- E escassos conjuntos de dados com legendas descritas em outros idiomas.

Image Caption

O objetivo de image caption é descrever o conteúdo visual de uma imagem, empregando um sistema de entendimento visual e um modelo de linguagem capaz de gerar sentenças com significado e sintaticamente corretas.

Aplicações

Geração de relatórios médicos [1]



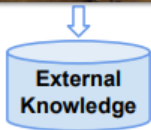
Aplicações

geração de descrição de arte [2]

Art Painting Image



Art
Describer



Vasari
Signorelli
Lamentation
1502
Christ
Cortona
...

(Context) An account of Vasari says that Signorelli wanted to represent in the figure of the naked Christ his own son, who died of plague in 1502. (Content) In the middle of the painting, there is an unreal landscape, clear and ... (Form) It strikes the observer with great power on account of its dimensions, the liveliness of ...

Aplicações

geração de descrição de notícias [3]



President **Obama** and **Mitt Romney** debate in **Hempstead NY** on **Tuesday**.



A bunch of people who are holding red umbrellas.



Virginia Cavaliers fans celebrate on the court after the **Cavaliers** game against the **Duke Blue Devils** at **John Paul Jones Arena**.



A baseball player hitting the ball during the game.

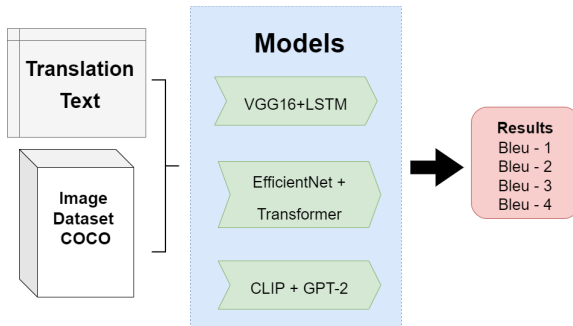
Hipóteses

- Os modelos de aprendizado profundo podem gerar legenda da imagem de maneira compreensiva ao nativo no idioma português?
- Qual modelo de aprendizado profundo gera uma saída similar a legenda da imagem esperada em português?
- As técnicas da aplicação de aprendizado profundo sobre a base de dados em português tem um resultado igual ou superior do que as técnicas de aprendizado profundo usando a base de dados em inglês e com tradução da legenda para o português?

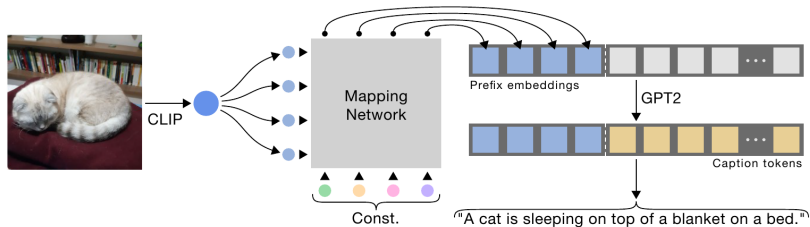
Objetivo

- Encontrar alto nível de similaridade entre a legenda original e a legenda predita da imagem, em português usando modelos de aprendizado profundo.

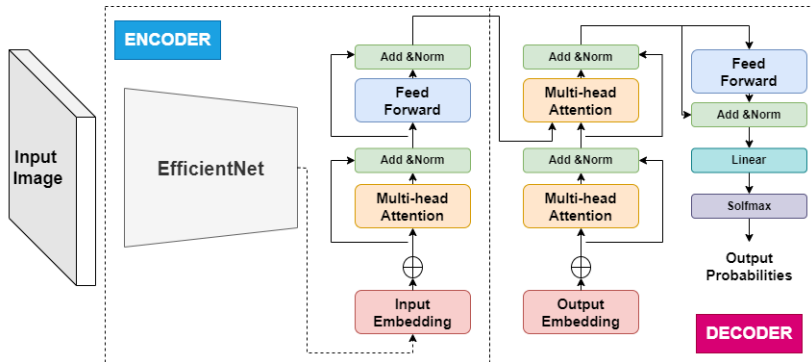
Overview do método



CLIP + GPT-2



InceptionResNetV2+Transformer



Base de Dados

- Criou-se uma versão em português do MS-COCO.
- O dataset MS-COCO contém **530.874** legendas para treinamento, **58.986** de validação e **5000** para teste, em que cada imagem tem 5 legendas diferentes.

Instância






- Um gato laranja está dormindo em um teclado. Um close-up de um gato sentado ao lado do teclado. Um gato dormindo com é ouvir descansando em um teclado. Um gato com a cabeça apoiada em um teclado.

Resultados

Architecture	Blue-1	Bleu-2	Bleu-3	Bleu-4
VGG16 + LSTM	0.118	0.153	0.097	0.047
EfficientNet + Transformer	0.610	0.411	0.244	0.127
CLIP + GPT2	0.664	0.492	0.329	0.209

Resultados

			
Ground Truth	Uma cena de rua com foco nas placas de rua em um viaduto.	uma mulher esta batendo uma bola com uma raquete	Uma banheira branca embaixo de uma janela.
VGG16 + LSTM	Um Caminhão De Bombeiros Está Estacionado Em Um Campo.	Um Homem De Terno E Gravata.	Um Gato Está Sentado Em Uma Cadeira De Madeira.
EfficientNet + Transformer	Um carro esta dirigindo por uma rua movimentada.	Uma mulher de vestido azul e branco jogando tenis	Uma banheira branca ao lado de uma janela uma.
CLIP + GPT2	Uma estrada cheia de tráfego ao lado de um sinal de autoestrada.	Uma mulher batendo uma bola de tênis com uma raquete de tênis.	Um banheiro sujo com pia, vaso sanitário e banheira.

Conclusão

- Neste trabalho geramos legendas de imagens em língua portuguesa sobre o conjunto de dados MSCOCO usando diferentes modelos.
- Criamos um novo conjunto de dados traduzido do conjunto de dados MSCOCO com boa qualidade gramatical usando o Google Translate.
- Os resultados mostram que o CLIP+GPT2 tem os melhores resultados em todos os cenários com uma melhoria de até 64% em comparação com as outras arquiteturas. O VGG16+LSTM apresenta os piores resultados em todos os casos.
- Para trabalhos futuros, pretendemos criar mais conjuntos de dados em português como traduzir o Flickr30k.
- Inserir um modulo de conversão de texto para voz para criação de um aplicativo de descrição.

Referências I

- [1] Xingyi Yang et al. **Writing by Memorizing: Hierarchical Retrieval-based Medical Report Generation**. 2021. DOI: 10 . 48550 / ARXIV . 2106 . 06471. URL: <https://arxiv.org/abs/2106.06471>.
- [2] Zechen Bai, Yuta Nakashima e Noa Garcia. **Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation**. 2021. DOI: 10 . 48550 / ARXIV . 2109 . 05743. URL: <https://arxiv.org/abs/2109.05743>.
- [3] Fuxiao Liu et al. **Visual News: Benchmark and Challenges in News Image Captioning**. 2020. DOI: 10 . 48550 / ARXIV . 2010 . 03743. URL: <https://arxiv.org/abs/2010.03743>.

Obrigado!!!