Resumo: Facilitated Deep Learning for Image Captioning

De Matheus Loiola Pinto Curado Silva

21 de setembro de 2023

Para saber mais sobre o modelo e menos sobre o planemento e o processo de pesquisa, pule para a seção de arquitetura do modelo e treinamento.

Introdução

O objetivo da legendagem de imagens é ajudar em inúmeras aplicações práticas, como engines de busca, recuperação de imagens, entre outros...

Trabalhos relacionados

Trabalhos iniciais na área proporam métodos de template para extrair atributos das imagens, como objetos, atributos visuais e relações espaciais, e assim gerar palavras usando um modelo *n-gram* para gerar legendas.

Em métodos de Deep Learning, a arquitetura *Encoder-Decoder* foi muito utilizada, e ela consiste em ler uma sentença e convertê-la para uma representação em vetor de tamanho fixo, que é usada inicialmente em uma *hidden layer* para gerar a sequência desejada. Após isso, o vetor de características é levado ao decodificar LSTM para gerar a legenda palavra por palavra. Entretanto, esse método considera a imagem inteira, e não foca em partes específicas da imagem.

Nos modelos de linguagem, RNN tradicionais sofrem com o problema de *vanishing/exploding gradient* em sentenças longas, e por isso, modelos LSTM são utilizados.

Assim, o modelo proposto pelo autor é um modelo híbrido que considera objetos significantes na imagem pela sua importância, e assim gerar uma legenda mais precisa.

Metodologia

O modelo proposto recebe uma imagem, que é processada e enviado ao detector de objetos no formato apropriado. Dessa forma, o detector de objetos retorna uma lista de objetos detectados, que é utilizada em uma heurística para transformar os objetos em um ponto inicial para o legendador, que é um híbrido de um extrator de características e um modelo de linguagem com enriquecimento por atenção, e esse ponto é adicionado em cada iteração no modelo de linguagem. Assim, em conjunto com as características extraídas, o modelo de linguagem é utilizado para gerar a legenda iterativamente.

Arquitetura do Modelo e Treinamento

Extrator de características

O primeiro estágio do artigo foi utilizar um extrator de características já treinado, e o escolhido foi *ResNet50* pre-treinado da ImageNet. O modelo foi utilizado da forma como veio, apenas retiraram o módulo de classificação, para apenas utilizar o vetor de características.

Detector de objetos

O modelo escolhido para esse estágio foi o *ResNet50* pre-treinado do RetinaNet, e foi treinado do *dataset COCO*, com mAP de 0.45, utilizando 70 classes.

Modelo de legendamento

O modelo precisaria identificar objetos em uma imagem de forma semântica e estrutural, para retornar uma descrição de forma coerente. Assim, dois modelos foram utilizados no estudo do autor: um LSTM de três camadas e um modelo de linguagem com enriquecimento por atenção.

Além disso, os experimentos foram feitos a partir de variações e legendas originais do *COCO dataset*, e variações do *Flickr30k* para treinamento.

Os dois modelos foram treinados com uma versão pré-processada das legendas, e variando os hiper-parâmetros. O modelo de linguagem com atenção convergiu melhor que o LSTM.

Staggering Stages

Algoritmos de atenção ajudam o modelo a focar em regiões de interesse da imagem, e não olhar ela inteiramente, o que a permite focar nas informações importantes. No começo do modelo de geração de legendas, esse algoritmo não é tão útil, pois existe pouco contexto para se basear. Dessa forma, isso pode ocasionar em erros, visto que, nos estágios avançados do modelo, quando o algoritmo de atenção começa a mostrar seu valor, a informação pode ter convergido anteriormente para um erro, prejudicando todo o processo.

Assim, utilizando um ponto inicial pré-definido, esse problema pode ser evitado, pois o modelo de geração de legendas terá um contexto desde o inicio, e ele servirá como referência para moldar a legenda ao longo das iterações do modelo.

Resultados

Os resultados são comparados a partir da *Similaridade entre cosenos*. A partir dele, foi observado que o modelo com esse auxílio da legenda inicial foi superior em 66% das imagens, enquanto o modelo sem auxílio superou ele em apenas 22% das imagens. Dessa forma, o modelo facilitado ofereceu uma melhor perfomance em comparação com os outros modelos.