Resumo: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models

De Matheus Loiola Pinto Curado Silva

5 de outubro de 2023

Introdução

Existem poucos *datasets* com a descrição de frases em regiões da imagem. A maioria dos métodos de legendagem tentam aprender mapeamentos diretamente de imagens inteiras a frases inteiras. Não por acaso, tais modelos tendem a reproduzir legendas genéricas a partir dos dados de treinamento, e tem uma perfomance ruim em imagens com composições inéditas, nas quais objetos podem ter sidos vistos no treinamento isoladamente, mas não em combinação com outros.

A principal contribuição do artigo é providenciar um largo *dataset* de descrição de imagens que contenha a descrição região-para-frase das imagens.

Flickr30k contém 31.783 imagens que focam principalmente em pessoas e animais, e 158.915 legendas em inglês (5 por imagem). O *dataset* Flickr30k Entities melhora o dataset Flickr30k ao identificar quais menções ao longo das legendas se referem a imagens no mesmo conjunto, resultando em 244.035 correntes de coreferência.

Trabalhos relacionados

Descrições em nível de região

O conjunto de dados mais recente e em grande escala desse tipo é Microsoft Common Objects in Context (MSCOCO). No entanto, assim como nas Sentenças UIUC, as anotações em nível de região MSCOCO não estão vinculadas às legendas de forma alguma.

Em vez de emparelhar imagens com uma legenda que resume a imagem inteira, alguns conjuntos de dados emparelham objetos específicos em uma imagem com descrições curtas, como o conjunto de dados ReferIt.

Grounded Language Understanding

Grounded Language Understanding é a tarefa de aprender o significado das unidades de linguagem natural, e o problema mais comum da literatura é o legendamento automático das imagens.

Para o artigo, os métodos vistos para esse problema são os que focam em associar uma região local da imagem em palavras ou frases de uma legenda.

Estatísticas

O processo de anotação do *dataset* identificou 513.644 entidades ou cenas, que foram linkadas com 244.035 correntes de coreferência. O processo de desenhar as bounding boxes geraram 275.775 bounding boxes nas 31.783 images.

Objetos mencionados foram:

- Pessoas com 94.2% de frequência.
- Animais com 12.0% de frequência.
- Roupas e partes do corpo com 69.9% e 28.0% de frequência.
- Veículos e instrumentos com 13.8% e 4.3% de frequência, enquanto outros objetos são mencionados em 91.8% das imagens.

Resultado do experimento

A principal motivação em coletar Flickr30k Entities é promover o desenvolvimento de métodos que podem raciocinar sobre correspondências detalhadas entre frases no texto e regiões em uma imagem. Para isso, eles propoem um benchmark para o problema da localização de frases.