Multi-Modality Cross Attention Network for Image and Sentence Matching

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, Feng Wu
University of Science and Technology of China; Kuaishou Technology

2020

Abstract

What challenges do authors had during the process?

Their model should match images with specific sentences. This was done by determining how similar are visual content and semantic content.

While most models are either inter-modality or intra-modality, they propose a model both inter-modality and intra-modality (MultiModality Cross Attention (MMCA)) with accurate results on Flickr30K and MS-COCO benchmarks. But what is inter-modality or intra-modality?

Intra-modality versus Inter-modality

In Intra-modality models there is a connection and interaction within single type of data. In this content, it refers to the type image which it means spatial and contextual relationship between different regions or objects within the same image or it refers to text which means grammatical and semantic connection between words in the same sentence.

On the other hand, in Inter-modality models the relationship is between image regions and sentence words. This model links visual elements in an image to semantic elements in a sentence.

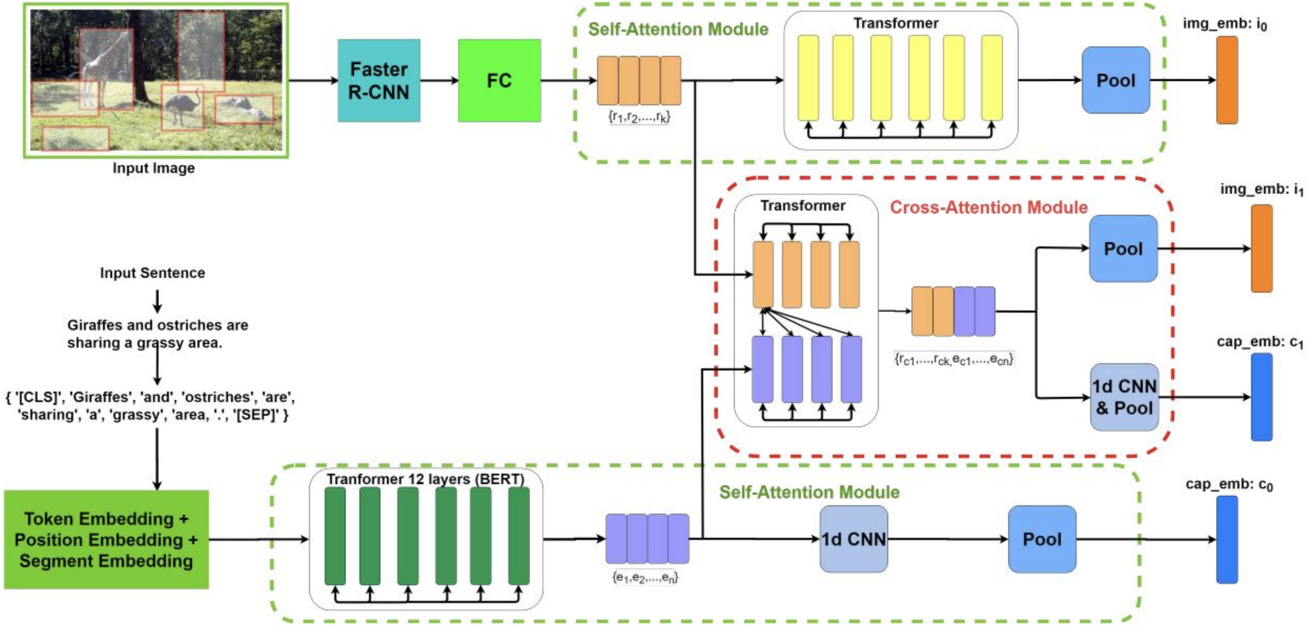
Introduction

Image sentence matching is used in many tasks in vision including image sentence retrieval (the most relevant senteree for a given image), visual grounding (Inking Phrases or words in a sentence to the corresponding regions or objects in an image), Visual Question Answering (answering questions about the content of an image like objects- actions) and many other tasks. Existing methods is these tasks are either one-to-one matching or many-to-many matching task.

One-to-one matching tasks extract global representation for images and sentences and match them by embedding them both into a shared space. However, this will miss the local similarities. Many-to-many matching tasks consider relationship between multiple regions of an image and multiple words in a sentence. This tasks are divided into inter-modality and intra-modality themself.

Their proposed model is both inter-modality and intra-modality. Because has two modules, one self-attention module which extract features of image's regions & uses word token embeddings fr sentences, applying a transformer, a BER model, to discover intra-modality relationships, and another one is a cross-attention module which combines representation of image’s regions and sentence words, processes them with a transformer and a 1D-CNN to fuse intra-modality and inter-modality to predict similarity scores.

The model


Input Image
Faster R-CNN
FC
Self-Attention Module
$\{r_1, r_2, \ldots, r_k\}$
Transformer
Pool
img_emb: i_0
Transformer
Cross-Attention Module
Pool
img_emb: i_1
$\{r_{c1}, \ldots, r_{ck}, e_{c1}, \ldots, e_{cn}\}$
1d CNN & Pool
cap_emb: c_1
Input Sentence
Giraffes and ostriches are sharing a grassy area.
{ '[CLS]', 'Giraffes', 'and', 'ostriches', 'are', 'sharing', 'a', 'grassy', 'area, '.', '[SEP]' }
Token Embedding + Position Embedding + Segment Embedding
Tranformer 12 layers (BERT)
$\{e_1, e_2, \ldots, e_n\}$
Self-Attention Module
1d CNN
Pool
cap_emb: c_0

Proposed method

In the first step, given an image and sentence pair, the model focuses on sentences and images separately. It feeds image using a bottom up attention model which has been pretrained on Genome dataset. This model extract features from different regions in images. The output is set of region features O={o(1), o(2), o(3), ..., o(n)} on which o(i) is mean pooled convolutional feature for the ith region then the result is given to fully-connected layer and the results would be R={r(1), r(2), r(3), ..., r(n}).

On the other side, the words in the sentences are broken down into wordPiece tokens(a subword tokenization method). These are the tokens considered correspondent to the mage regions (inspired from a work in machine translation). Each word in the sentence is a combination of 3 embeddings:

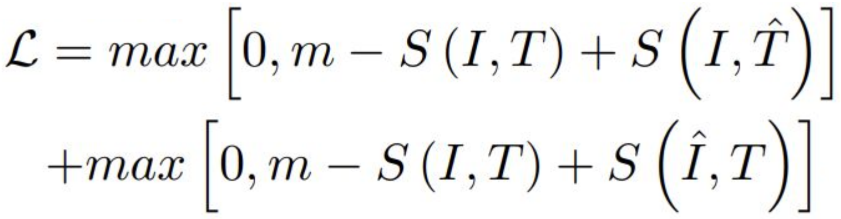
1. Token embedding: actual word or token
2. Position embeddings: position of the word in the sentence.
3. Segment embedding: distinguishes between different parts of the input text

In the alignment part there are i0, the results from self-attention module for images, c0, the results from self-attention module for text, i1, the results from cross-attention module for image and c1, the result from cross-attention module for text are combined together. So the similarity score for image I and sentence T would be:

$$S\left(I,T\right) = i_0 \cdot c_0 + \alpha\left(i_1 \cdot c_1\right)$$

Loss function

They used bi-directional triplet ranking loss. This encourages similarity score of matched images & sentences to be larger than mismatched ones. They also used hard negative mining to improve the performance of the model.



$$\begin{aligned}\mathcal{L} = max\left[0, m - S\left(I, T\right) + S\left(I, \hat{T}\right)\right] \\ +max\left[0, m - S\left(I, T\right) + S\left(\hat{I}, T\right)\right]\end{aligned}$$

where m denotes the margin, (I, T) denotes the true matched image-sentence pair, and I(hat), T(hat) stand for the hard negatives in a mini-batch.