Review on "Modeling Relationships in Referential Expressions with Compositional Modular Networks

Irdi Balla

22/11/2017

1 Introduction

Detecting object that belong to a predefined set from an image, is something that we know how to do an a lot of work has been put to it. But when it comes to localizing entities from natural expressions the task becomes very challenging. The work on this field focuses on identifying a bounding box that encapsulates the object. But this work either treats referential expressions holistically or relies on fixed sets defined a priori. Compositional Modular Networks (CMNs), focuses on inter-object relationships, and it is an end-to-end model that learns language representation and image recognition jointly.

2 Related Work

2.1 Grounding Referential Expressions

To perform grounding of referential expressions we could first extract a set of candidate regions and then score them with respect to the expression and based on the scores output the result. However this is often insufficient to determine if a region matches an expression. More recent work adds also contextual feature extracted from other regions in an image. These methods represent language holistically using a recurrent neural model, by predicting a distribution or by encoding expressions into vector representations. CMNs learns to parse the expressions into textual components and align them with the image regions end-to-end.

2.2 Handling Inter-Object Relationships

Recent work rains detectors based on RCNN, and uses a linguistic prior to detect relationships in images. CMN does not build on a fixed inventory of classes but rather learns expression parsing and visual entity localization.

2.3 Compositional Structure with Models

Neural Module Networks assemble e network architecture after decomposing the question. THis method relies on an external parser. CMN use a modular structure as well, but they also learn the language representetion end-to-end from the words.

3 CMN

CMNs are compositional in the sense that they ground the components of an expression and use their relationship. CMNs focus on the relationship in the expression which can be represented as a 3-component element (subject, relationship, object). To handle these relationships pairs of regions are considered. CMNs are composed of 2 modules, a localization module and a relationship module. The pairwise score is computed as follows:

$$s_{pair}(b_i, b_j) = f_{loc}(b_i, q_{subj}; \Theta_{loc}) + f_{loc}(b_j, q_{obj}; \Theta_{loc}) + f_{rel}(b_i, b_j, q_{rel}; \Theta_{rel}), \quad (1)$$

The best pairwise score is then grounded to the highest scoring region.

$$s_{subj}(b_i) \triangleq \max_{b_j \in B} s_{pair}(b_i, b_j). \quad (2)$$

$$b^*_{subj} = \arg\max_{b_i \in B}(s_{subj}(b_i)). \quad (3)$$

3.1 Expression Parsing with Attention

Deciding on which words are subject, object and relationship is not an easy task. CMNs follow several steps to do that. First each word is embedded to a vector using GloVe, then, using a 2-layer LSTM, CMNs scan through the word embedding sequence. The first layer outputs the concatenation of 2 hidden states(at every time step). The second layer takes that as input and outputs its concatenation with 2 more hidden states Then the attention weights over each word are obtained as follows:

$$a_{t,subj} = \frac{\exp\left(\beta_{subj}^T h_t\right)}{\sum_{\tau=1}^{T} \exp\left(\beta_{subj}^T h_\tau\right)} \quad (4)$$

$$a_{t,rel} = \frac{\exp\left(\beta_{rel}^T h_t\right)}{\sum_{\tau=1}^{T} \exp\left(\beta_{rel}^T h_\tau\right)} \quad (5)$$

$$a_{t,obj} = \frac{\exp\left(\beta_{obj}^T h_t\right)}{\sum_{\tau=1}^{T} \exp\left(\beta_{obj}^T h_\tau\right)} \quad (6)$$

The language representation of the trio is calculated as

$$q_{subj} = \sum_{t=1}^{T} a_{t,subj} e_t \tag{7}$$

$$q_{rel} = \sum_{t=1}^{T} a_{t,rel} e_t \tag{8}$$

$$q_{obj} = \sum_{t=1}^{T} a_{t,obj} e_t \tag{9}$$

3.2 Localization Module

The localization module outputs a score $s_{loc} = f_{loc}(b, q_{loc}; \Theta_{loc})$ which represents how likely the region bounding box matches subject or the object. The module takes a visual feature(extracted from an image region using CNN) and a 5-dim spatial feature. These two are then concatenated in a vector. To get a score in the end, the following is done

$$\tilde{x}_{v,s} = W_{v,s} x_{v,s} + b_{v,s} \tag{10}$$

$$z_{loc} = \tilde{x}_{v,s} \odot q_{loc} \tag{11}$$

$$\hat{z}_{loc} = z_{loc} / \|z_{loc}\|_2 \tag{12}$$

$$s_{loc} = w_{loc}^T \hat{z}_{loc} + b_{loc}. \tag{13}$$

3.3 Relationship Module

The score of this module represents how likely the pair of regions matches the relationship in the expression. He spatial features of both regions are used in the same way as in the localization module; concatenated and then

$$\tilde{x}_{s1,s2} = W_{s1,s2} x_{s1,s2} + b_{s1,s2} \tag{14}$$

$$z_{rel} = \tilde{x}_{s1,s2} \odot q_{rel} \tag{15}$$

$$\hat{z}_{rel} = z_{rel} / \|z_{rel}\|_2 \tag{16}$$

$$s_{rel} = w_{rel}^T \hat{z}_{rel} + b_{rel}. \tag{17}$$

3.4 End-to-end Learning

Tha pairwise score cannot be always directly optimized. CMNs treat the object region as a latent variable. and optimize the score of the subject. It can be optimized with weak supervision using softmax

$$Loss_{weak} = -\log\left(\frac{\exp\left(s_{subj}(b_{subj\_gt})\right)}{\sum_{b_i \in B} \exp\left(s_{subj}(b_i)\right)}\right) \tag{18}$$

The whole system is then trained end-to-end.