Pattern Recognition Letters 111 (2018) 51–57

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Explicit ensemble attention learning for improving visual question answering

Vasileios Lioutas, Nikolaos Passalis*, Anastasios Tefas

*Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece*

ARTICLE INFO

*Article history:*
Received 21 August 2017
Available online 22 April 2018

*Keywords:*
Visual question answering
Explicit attention
Pictorial superiority effect

ABSTRACT

Visual Question Answering (VQA) is among the most difficult multi-modal problems as it requires a machine to be able to properly understand a question about a reference image and then infer the correct answer. Providing reliable attention information is crucial for correctly answering the questions. However, existing methods usually only use implicitly trained attention models that are frequently unable to attend to the correct image regions. To this end, an explicitly trained attention model for VQA is proposed in this paper. The proposed method utilizes attention-oriented word embeddings that allows efficiently learning the common representation spaces. Furthermore, multiple attention models of varying complexity are employed as a way of realizing a mixture of experts attention model, further improving the VQA accuracy over a single attention model. The effectiveness of the proposed method is demonstrated using extensive experiments on the Visual7W dataset that provides visual attention ground truth information.



## 1. Introduction

Due to the recent developments in the Natural Language Processing and Computer Vision areas, in combination with the rapidly increasing computational power, significant research efforts have been focusing on tackling the problem of building machines that interlink multiple modalities [33–35]. One of the most prominent multi-modal problems is the task of Visual Question Answering (VQA) [4,29,42], which has become one of the most active research directions, not only in academia but also in the industry. VQA requires a machine to be able to properly understand a question about a reference image and then infer the correct answer.

To provide fine-grained information regarding the visual content, *attention mechanisms* have been developed [17,42]. These methods work under the assumption that the image regions, which contain relevant information to the question at hand, will eventually be strongly associated with the corresponding questions, while the irrelevant regions will exhibit diminishing association. However, these attention mechanisms are trained *implicitly*. It was recently demonstrated that utilizing explicitly trained attention models can improve the accuracy of automatic caption generation [19]. Even though a recently released VQA dataset, the Visual7W dataset [42], contains attention ground truth information for some of the available questions, no technique has yet exploited this information to train more accurate attention models and evaluate their performance using the Visual7W dataset.

The paper proposes a novel explicit attention model for VQA tasks. Inspired by the theory of the pictorial superiority effect, we propose employing separate word embeddings for the attention model that is independent from the embeddings, which are used for answering the questions. The theory of pictorial superiority effect refers to the phenomenon of humans remembering images easier compared to words [23,24]. Thus, it is easier to learn common representation spaces, where each word is closer to the visual representation of its semantic content. Finally, recognizing the difficulty of training reliable attention models we use multiple attention models of varying complexity as a way of realizing a mixture of experts attention model [22] that is able to provide more accurate answers than a single attention model. We demonstrate the effectiveness of the proposed method, over both implicit attention models as well as other state-of-the-art VQA techniques, using the Visual7W dataset [42].

The rest of the paper is structured as follows. The prior work is discussed in Section 2 and the proposed method is presented in Section 3. The experimental evaluation is presented in Section 4 and conclusions are drawn in Section 5.

## 2. Related work

Visual Questioning Answering (VQA) methods fall into two categories: (a) The *generative* methods, in which the answer is gen-

* Corresponding author.
*E-mail addresses:* lioutasv@csd.auth.gr (V. Lioutas), passalis@csd.auth.gr (N. Passalis), tefas@aiia.csd.auth.gr (A. Tefas).

https://doi.org/10.1016/j.patrec.2018.04.031
0167-8655/

erated in free-form text, and (b) the *classification*-based methods, in which the correct answer is chosen among a set of predefined answers. The generative methods usually employ recurrent models, such as Long Short-Term Memory Units (LSTMs), to generate the answer to the question [3,21,35]. However, generating the answer in free-form text significantly complicates the evaluation procedure, since there are multiple correct answers for the same question [12]. On the other hand, classification-based methods extract features from the input modalities and then employ a classifier to determine the correct answer [2,12,20,40]. Many of these method, e.g., [3,28,42] also utilize recurrent models to provide better encoding of the input modalities. However, it is worth mentioning, as it was recently established, that employing a simple triplet-based scheme (question-answer-image) [12] can significantly improve the answering accuracy over the rest of the methods proposed in the literature. In this work, a triplet-based classification scheme is also combined with the proposed explicit attention model. The interested reader is referred to [13,36] for an extensive review of VQA methods as well as of the currently available VQA datasets.

A rich literature on using implicit attention models to improve visual analysis tasks also exist. These models work by learning weighting coefficients (or a probability distribution) over the extracted feature maps. Implicit attention model are capable of improving the accuracy of the models for various tasks, e.g., [10,17,30,31,38], including VQA tasks [42]. An attention model that was trained with ground truth human attention information (explicit attention) has been applied for caption generation tasks in [19], and it was shown to improve the accuracy compared to implicit attention models. Also, an extensive discussion regarding the differences between implicit attention models and human attention is provided in [6]. This work also highlights the potential of utilizing explicitly trained attention models for the task of VQA.

To the best of our knowledge, in this paper we propose the first explicitly trained *ensemble* attention model for VQA tasks that is capable of utilizing multiple attention distributions generated by models of varying complexity. Another explicit attention model was proposed by Qiao et al. [27]. This model used the multimodal low-rank bilinear pooling (Kim et al., 2017) to provide several smaller attention maps that were then applied to infer the final attention distribution. In contrast to these approaches, our method is capable of combining several different attention distributions that are provided by *multiple* attention models. This increases the probability of attending to the correct image regions. The ability of our ensemble approach to increase the question answering accuracy is experimentally demonstrated in Section 4. Also, inspired by the pictorial superiority phenomenon, we propose a biologically justified approach that decouples the attention process from the answering process utilizing two separate word embeddings. This further increases the expressive power of the proposed attention model. Finally, instead of utilizing bilinear pooling, we employ a simpler and more lightweight correlation approach through a series of non-linear operators (*tanh* and *relu*).

## 3. Proposed method

The used notation is introduced and the proposed explicit attention model, along with the complete pipeline of the proposed visual question answering system, are described in detail in this Section.

### 3.1. Explicit attention model

The architecture of the proposed explicit attention model is summarized in Fig. 1. The goal of the proposed model is to reduce the semantic gap between textual and image representations. To achieve this, the proposed method *directly* learns to attend the parts of an image that correspond to the given question using the supplied ground truth information. Thus, only the image regions that are actually related to the question at hand are used to infer the correct answer. Instead of utilizing the same word embedding for providing both the correct answer and the attention information, two separate word embedding models are employed as shown in Fig. 2. In that way, it is possible to learn a separate word embedding model that it is only utilized for providing the visual attention information and another one for providing the correct answer. This decoupling allows for increasing the expressive power of the attention model that is equipped with a separate (visual oriented) word embedding model, which does not tie to the word embedding employed for providing the correct answers. We inspired this idea from the theory of pictorial superiority effect [23,24] that states that "human memory is extremely sensitive to the symbolic modality of presentation of event information" [39]. It was also experimentally shown that decoupling the word representations used for providing the attention from the word representations utilized for answering the question at hand improves the overall accuracy of the system.

Consider the proposed explicit attention model (Fig. 1). Let $Q = \{\mathbf{q}_1, \ldots, \mathbf{q}_N\}$ denote a question, where $N$ is the number of words in the question, $\mathbf{q}_i \in \mathbb{R}^{D_w}$ is the embedding vector for the $i$-th word, and $D_w$ denotes the dimensionality of the word embedding. Also, the notation $\mathbf{I}_m \in \mathbb{R}^{D_m \times D_m \times D_d}$ is used to refer to the feature map utilized for providing the attention, where $D_m \times D_m$ is the size of the extracted feature map and $D_d$ is the number of convolutional filters.

First, the words of a question $Q$ are embedded into a textual vector space employing a word embedding model. After that, the representation $\mathbf{Q}_f \in \mathbb{R}^{D_w}$ of the question $Q$ is extracted by averaging the word embedding vectors extracted from the question $Q$, where $D_w$ is the dimensionality of the word embedding. Then, the attention distribution $\mathbf{p}_I$ over the convolutional feature map $\mathbf{I}_m$, for a given the question $Q$, is calculated as:

$$\mathbf{h}_c = [\tanh(\mathbf{I}_m \times \mathbf{W}_I); \mathbf{1}_{D_m \times D_m \times 1} \times \tanh(\mathbf{Q}_f \times \mathbf{W}_Q)] \in \mathbb{R}^{D_m \times D_m \times 2D_c}, \tag{1}$$

$$\mathbf{p}_I = \text{softmax}(\text{relu}(\mathbf{h}_c \times \mathbf{W}_{h1}) \times \mathbf{W}_{h2}) \in \mathbb{R}^{D_m \times D_m}, \tag{2}$$

where $\mathbf{1}_{D_m \times D_m \times 1}$ is a matrix used for stacking $\tanh(\mathbf{Q}_f \times \mathbf{W}_Q)$ $D_m \times D_m$ times in $\mathbf{h}_c$, and $\mathbf{W}_I \in \mathbb{R}^{D_d \times D_c}$ and $\mathbf{W}_Q \in \mathbb{R}^{D_w \times D_c}$ are the weights employed for projecting the question into a common representation space. The dimensionality of the common representation space is controlled by $D_c$. Eq (2) provides the attention distribution over the image regions, as they are expressed through the extracted feature map. Note that a Multilayer Perceptron (MLP) with $D_h$ hidden units is utilized to provide the attention distribution, where $\mathbf{W}_{h1} \in \mathbb{R}^{(2D_c) \times D_h}$ and $\mathbf{W}_{h2} \in \mathbb{R}^{D_h \times 1}$ denote the weight matrices of the MLP. Finally, the extracted attention distribution $\mathbf{p}_I \in \mathbb{R}^{D_m \times D_m}$ is employed to provide the attention-based representation:

$$\mathbf{I}_{m'} = \sum_{i=1}^{D_m} \sum_{j=1}^{D_m} \mathbf{p}_{Iij} \mathbf{I}_{mij} \in \mathbb{R}^{D_d}. \tag{3}$$

The ground truth bounding boxes $B_T$, which associate the correct answer with different regions of the image, are employed to train the proposed explicit attention model. The attention targets are defined as follows:

$$\hat{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}}{||\boldsymbol{\alpha}||_0} \in \mathbb{R}^{D_m \times D_m}, \tag{4}$$