



Under review at Pattern Recognition Letters

Order embeddings and character-level convolutions for multimodal alignment

Jônatas Wehrmann^a, Anderson Mattjie^a, Rodrigo C. Barros^{a,**}

^aPontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga, 6681, Porto Alegre, RS, 90619-900, Brazil

ABSTRACT

With the novel and fast advances in the area of deep neural networks, several challenging image-based tasks have been recently approached by researchers in pattern recognition and computer vision. In this paper, we address one of these tasks, which is to match image content with natural language descriptions, sometimes referred as multimodal content retrieval. Such a task is particularly challenging considering that we must find a semantic correspondence between captions and the respective image, a challenge for both computer vision and natural language processing areas. For such, we propose a novel multimodal approach based solely on convolutional neural networks for aligning images with their captions by directly convolving raw characters. Our proposed character-based textual embeddings allow the replacement of both word-embeddings and recurrent neural networks for text understanding, saving processing time and requiring fewer learnable parameters. Our method is based on the idea of projecting both visual and textual information into a common embedding space. For training such embeddings we optimize a contrastive loss function that is computed to minimize order-violations between images and their respective descriptions. We achieve state-of-the-art performance in the largest and most well-known image-text alignment dataset, namely Microsoft COCO, with a method that is conceptually much simpler and that possesses considerably fewer parameters than current approaches.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

To learn proper feature representation of input data is an essential part of any machine learning problem, as it directly impacts in the precision of the generated data-based models. Thanks to the fast pace in which computer hardware has been evolving in the last decade, along with the rapid development of computer vision and natural language processing technologies, great advances have been made in tasks that require a huge amount of computational power, in particular those that benefit from optimization-based approaches such as deep neural networks. One of these tasks, image-text alignment, has become an important problem in the latest years as it has many applications such as image and video retrieval, captioning, navigation for the blind, and many others. By successfully mapping image-to-text and text-to-image, we can significantly impact the general and broad task of information retrieval.

For this multimodal content retrieval task, state-of-the-art results (Mao et al., 2014; Vinyals et al., 2015; Venugopalan

et al., 2016) rely on either Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Units (GRUs) (Chung et al., 2015) with word-embeddings (Mikolov et al., 2013). Although such an approach has shown promising results, it has some drawbacks that are important to consider. First of all, it is costly due to the need of pre-training a word-embedding with a latent space informative enough to capture semantic relationships among words. Second, it takes a considerable amount of storage and memory for dealing with these word embeddings, depending on the size of the dictionary, in which often larger is better in terms of predictive accuracy.

By taking into consideration the aforementioned drawbacks, we propose a model that, instead of relying in similar recurrent LSTM/GRU-based architectures that depend on pre-trained word-embeddings, learns from scratch, character per character, how to retrieve captions from images and images from captions by making use of convolutional layers alone. Additionally, our model makes no assumptions regarding specific templates, guidelines, or classifications since it learns everything from scratch using the training data. Since image captioning can be seen as a special case of a single visual-semantic hierarchy over words,

^{**}Corresponding author: Tel.: +55-51-3353-8641;
e-mail: rodrigo.barros@pucrs.br (Rodrigo C. Barros)

sentences, and images, we design a loss function based on the so-called order embeddings (Vendrov et al., 2016). This particular type of embeddings are designed for explicitly modeling the partial order structure of the visual-semantic hierarchy existent in image captions. While typical approaches for image captioning rely on mapping words to vectors in a distance-preserving fashion (Socher et al., 2014), we believe order embeddings to be more suitable since the resulting mapping is not distance-preserving but order-preserving between the hierarchy and a partial order over the embedding space, making it easier to relate the naturally-hierarchical concepts within image captions.

In order to evaluate the performance of our model, we execute a series of experiments where we make particular architectural changes to the convolutional neural network by increasing the number of convolutional layers up to 5 and by changing the number of filters. We compare our proposed approach with the current state-of-the-art, and we show that our method achieves state-of-the-art results while often presenting a much lighter, simpler, and easier-to-train architecture.

This paper is organized as follows. Section 2 presents in detail our novel approach for multimodal content retrieval. Section 3 describes the methodology that we adopt for performing the experimental analysis, and the results are discussed in Section 4. Section 5 reviews related work, and we end this paper with our conclusions and future work directions in Section 6.

2. Character-based Language Embeddings

The use of word-embeddings (Mikolov et al., 2013) for text understanding has become a standard approach, being largely used across several tasks such as text classification, machine translation, image and video captioning, and information retrieval. Such an approach consists of representing a given word into a multidimensional latent space, $w \in \mathbb{R}^d$. These embeddings are often projected into a distance-preserving Euclidean space, in which semantic regularities can be easily identified and even manipulated. An example of manipulation of this distance-preserving word-embedding space is the following vector operation over the corresponding words:

$$\text{king} - \text{man} + \text{woman} \sim \text{queen} \quad (1)$$

Current state-of-the-art sentence embedding approaches (Kiros et al., 2015; Vendrov et al., 2016; Mesnil et al., 2013; Kiros et al., 2014b; Karpathy and Fei-Fei, 2015) have demonstrated similar results when using RNNs and word-embeddings for encoding entire sentences into a d -dimensional embedding space. Nevertheless, it is important to emphasize that even though word-embeddings are employed in the current state-of-the-art approaches, they tend to present three major drawbacks: i) they require pre-training word-embeddings or RNNs in very large corpora (with millions or billions of words), which demands both time and computational power; ii) in order to encode a single word or sentence, it is necessary to have “at hand” the whole word-dictionary containing all known words, which largely increases the memory requirements to store all data; iii) for multilingual or informal domains (e.g., twitter and internet searches) the number of words in the dictionary increases

according to the number of languages and, in addition, preprocessing is often employed for correcting typos and standardizing the words.

In this work, we propose a novel architecture for learning textual embeddings based on convolving raw characters. Our approach is designed to be simple, efficient, and fast, though still capable of generating state-of-the-art results. Our hypothesis is that a convolutional layer is capable of learning a proper latent embedding space for encoding text semantics. Hence, we replace the word dictionary by applying f convolutional filters over the input text in a temporal window of size t . Moreover, to keep t unchanged through the layer computation, we perform padded convolutions. The resulting feature map $\in \mathbb{R}^{t \times f}$ can be seen as a text-embedding where each character is projected onto a f -dimensional space by considering its relation to the adjacent characters (defined by the filters’ length). Such an approach does not require fixed-sized dictionaries nor preprocessing of any kind, and its complexity does not raise with the complexity or availability of text.

2.1. Character-based Convolutions

The core idea behind our approach is to replace the use of both RNNs and word dictionaries by applying convolutions to learn textual embeddings. A given text is represented by $\mathcal{T} \in \{0, 1\}^{n \times a}$, where n is the number of characters in the text and a is the alphabet size. Note that n is variable and changes according to the available text, whereas a must be unchanged given that the alphabet contains all known characters. We convolve characters of the input text \mathcal{T} by applying f convolutional filters of length l , where the j^{th} filter in the i^{th} convolutional layer generates feature map \mathcal{F}_{ij} whose x^{th} position is given by :

$$\mathcal{F}_{ij}^x = \phi \left(b_{ij} + \sum_{m=0}^{f_i-1} \sum_{p=0}^{l-1} w_{ijm}^p \mathcal{F}_{(i-1)m}^{(x+p)} \right) \quad (2)$$

where ϕ is an activation function, b_{ij} is the bias for the respective convolutional filter, m iterates over the feature maps (channels), p indexes the position of the kernel, w_{ijm}^p is the filter weight and $\mathcal{F}_{(i-1)m}^{(x+p)}$ is the value of the previous feature map (or input). Note that m iterates over the alphabet size for the case of the first convolution, and over the f_i feature maps of the previous layer for the subsequent convolutions.

A known restriction of applying a single convolutional layer for embedding texts is the size of the receptive field. A convolutional filter of length $l = 7$ is capable of learning information of 7 neighboring characters. This size is probably enough for learning word-based information. Standard strategies for allowing the global learning of the whole text include: i) increasing l , which leads to an exponential growth of parameters, eventually making the learning unfeasible for large-size texts; ii) adding more convolutional layers, hence requiring more processing resources and parameters depending on the number of filters of the subsequent convolutions; and iii) using local or global pooling layers. In order to keep our architecture compact, we explore up to five convolutional layers and a maximum filter size of 7.

The final text-embedding vectors are generated by applying a max-pooling-over-time layer, which selects the most important features across the temporal dimension of feature map \mathcal{F} .

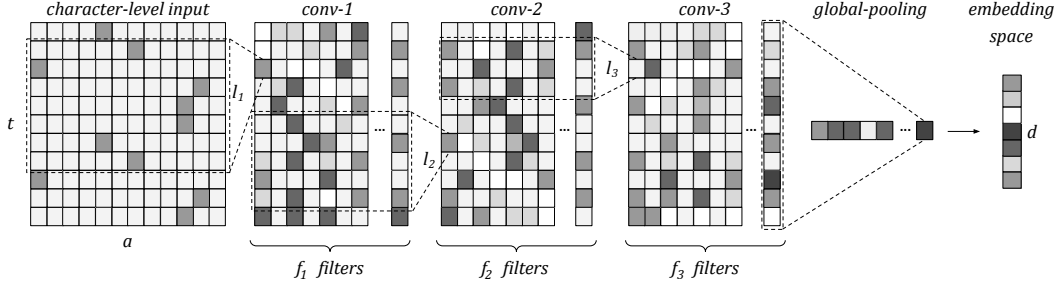


Fig. 1. Convolutions-through-time (CTT) module with 3 convolutional layers for processing a text of t characters based on an alphabet of a symbols.

Note that the number of filters in the last convolutional layer defines the length of the embedding vector. Figure 1 presents a schematic of text-embedding via character-based convolutions, which we have named “convolutions-through-time” (CTT).

2.2. Architecture

Our architecture is designed to approximate two encoding functions, $f_t(\mathcal{T})$ and $f_i(\mathcal{I})$, whose goal is to project both text \mathcal{T} and image \mathcal{I} into the same embedding space. In such a space, correlated image-text pairs should be close to each other, and the distance of non-correlated pairs should necessarily be larger than the correlated ones. For the text encoding function $f_t(\mathcal{T})$, we make use of the CTT module described in the previous section. For the image encoding function $f_i(\mathcal{I})$, we extract image features from the second fully-connected layer of a VGG-19 network (Simonyan and Zisserman, 2015) pre-trained in the ImageNet dataset (Russakovsky et al., 2015). For better feature representation, we use the 10-crop strategy: we scale to 256-pixels the smallest image size, and sample 224×224 crops from the corners, center, and horizontal mirroring. Finally, features from all crops are averaged element-wise.

Let $C(\mathcal{I})$ be features extracted from image \mathcal{I} by the convolutional neural network. Images are projected onto the \mathcal{R}_+^d embedding-space based on a linear mapping:

$$f_i(\mathcal{I}) = |W_i \cdot C(\mathcal{I})| \quad (3)$$

where $W_i \in \mathcal{R}^{d \times 4096}$ is a learned weight matrix and d is the number of dimensions of the embedding space.

For embedding text, we use the proposed character-based approach $f_t(\cdot)$ with four main variations shown in Table 1. We use maxout-based convolutions (Eq. (4)), given that it performs universal approximation of proper activation functions, addressing the rectifier linear unit (ReLU) issue of saturating in the negative region of activations:

$$\mathcal{F}_m = \max\{\psi(W_1, x), \psi(W_2, x)\} \quad (4)$$

where \mathcal{F}_m is a feature map generated by convolving $\psi(\cdot)$ – a given input x with two distinct weight tensors, namely W_1 and W_2 . Each activation value in the output map is selected as the largest of the two values generated by the distinct convolutions.

Since it uses a larger number of parameters, maxout convolutions generate more compact and informative feature maps while saving some memory when compared to ReLU-based networks with the same amount of parameters. In order to keep a reduced

Table 1. Four versions of our approach.

| Version | Architecture | Output | #Filters (f) | Filter Length (l) | #Params |
|---------|--------------|----------------|------------------|-----------------------|-----------|
| A | Input | $t \times 72$ | | | |
| | MaxConv_1 | $t \times 512$ | 512 | 7 | 517,120 |
| | Pooling | 512 | - | t | - |
| | | | | | 0.52M |
| B | Input | $t \times 72$ | | | |
| | MaxConv_1 | $t \times 256$ | 256 | 7 | 258,560 |
| | MaxConv_2 | $t \times 512$ | 512 | 5 | 1,311,744 |
| | Pooling | 512 | - | t | - |
| | | | | | 2.10M |
| C | Input | $t \times 72$ | | | |
| | MaxConv_1 | $t \times 128$ | 128 | 7 | 129,280 |
| | MaxConv_2 | $t \times 256$ | 256 | 5 | 328,192 |
| | MaxConv_3 | $t \times 512$ | 512 | 3 | 787,456 |
| | Pooling | 512 | - | t | - |
| | | | | | 1.24M |
| D | Input | $t \times 72$ | | | |
| | MaxConv_1 | $t \times 512$ | 512 | 7 | 517,120 |
| | MaxConv_2 | $t \times 512$ | 512 | 5 | 2,622,464 |
| | MaxConv_3 | $t \times 512$ | 512 | 3 | 1,573,888 |
| | Pooling | 512 | - | t | - |
| | | | | | 4.713M |

amount of parameters, we decreased the number of filters in the shallower convolutional layers, except for Architecture D, which contains the same number of filters in all layers.

All models are standardized to provide a 512-long vector representation that carries the textual semantic information. Similarly to $f_i(\cdot)$, we linearly project such representation onto \mathcal{R}_+^d by using a learned $W_t \in \mathcal{R}^{d \times 512}$ weight matrix.

Note that the current state-of-the-art approach for image-text alignment (Vendrov et al., 2016) makes use of GRU networks fed with word embeddings, requiring $\approx 8M$ parameters for the word-embeddings and $\approx 4M$ parameters for the GRU itself. Our largest architecture contains $\approx 4.7M$ parameters (almost $3\times$ lighter) and our smallest architecture contains only half a million parameters, $24\times$ lighter than the state-of-the-art.

2.3. Loss function

Let $f_t(\mathcal{T}) = \mathbf{m}$ be the sentence embedding vector and $f_i(\mathcal{I}) = \mathbf{v}$ be the image embedding. We first scale \mathbf{m} and \mathbf{v} to have unit norm, so that the inner product of both results in the cosine distance. Instead of directly optimizing the cosine distance, we decided for optimizing the alignment preserving the

Table 2. Bidirectional results in MS COCO test set. Values of the first section are reported in the 1-fold test set. Values of both second (state-of-the-art) and third (our models) sections are averages of the 5-fold test set. Values in bold depict the current state-of-the-art results. Values underlined outperform the best published results.

| Method | Image to text | | | | Text to image | | | |
|---|---------------|-------------|------------|------------|---------------|-------------|---------|------------|
| | R@1 | R@10 | Med r | Mean r | R@1 | R@10 | Med r | Mean r |
| MNLM (Kiros et al., 2014b) | 43.4 | 85.8 | 2.0 | * | 31.0 | 79.9 | 3.0 | * |
| DVSA (Karpathy and Fei-Fei, 2015) | 38.4 | 80.5 | 1.0 | * | 27.4 | 74.8 | 3.0 | * |
| FV (Klein et al., 2015) | 39.4 | 80.9 | 2.0 | 10.4 | 25.1 | 76.6 | 4.0 | 11.1 |
| m-CNN (Ma et al., 2015) | 38.3 | 81.0 | 2.0 | * | 27.4 | 79.5 | 3.0 | * |
| m-CNN _{ENS} (Ma et al., 2015) | 42.8 | 84.1 | 2.0 | * | 32.6 | 82.8 | 3.0 | * |
| combine-skip-thought (Kiros et al., 2015) | 33.8 | 82.1 | 3.0 | * | 25.9 | 74.6 | 4.0 | * |
| Order-embeddings (symm.) (Vendrov et al., 2016) | 45.4 | 88.7 | 2.0 | 5.8 | 36.3 | 85.8 | 2.0 | 9.0 |
| Order-embeddings (Vendrov et al., 2016) | 46.7 | 88.9 | 2.0 | 5.7 | 37.9 | 85.9 | 2.0 | 8.1 |
| Arch. A [ours] | 42.1 | 87.2 | 2.0 | 6.7 | 34.5 | 85.0 | 2.6 | 8.2 |
| Arch. B [ours] | 46.2 | 88.0 | 2.0 | 5.9 | 36.9 | <u>86.4</u> | 2.0 | <u>7.4</u> |
| Arch. C [ours] | 46.2 | 88.6 | 2.0 | 5.8 | 37.0 | <u>86.6</u> | 2.0 | <u>7.6</u> |
| Arch. D [ours] | 47.2 | 88.9 | 2.0 | <u>5.6</u> | 37.5 | 87.0 | 2.0 | <u>7.3</u> |

order relationships among the visual-semantic hierarchy, given that asymmetric distances are naturally more well-suited for image-sentence alignment. Hence, we apply an order-violation constraint by penalizing an ordered pair (x, y) of points in \mathcal{R}_+^N :

$$s(x, y) = -|\max\{0, y - x\}|^2 \quad (5)$$

The order violation penalties are used as a similarity distance, and optimized by the following contrastive pairwise ranking loss:

$$\mathcal{L} = \sum_{\mathbf{m}} \sum_k \max\{0, \alpha - s(\mathbf{m}, \mathbf{v}) + s(\mathbf{m}, \mathbf{v}_k)\} + \sum_{\mathbf{v}} \sum_k \max\{0, \alpha - s(\mathbf{v}, \mathbf{m}) + s(\mathbf{v}, \mathbf{m}_k)\} \quad (6)$$

where \mathbf{m}_k and \mathbf{v}_k are respectively the sentence and image contrastive examples (i.e., uncorrelated). This loss function encourages the similarity $s(x, y)$ for actual image-text pairs to be larger than the contrastive ones by a margin of at least α .

3. Experimental Setup

3.1. Dataset

For analyzing the performance of our proposed approach, we make use of the Microsoft COCO dataset (Lin et al., 2014). It contains over 100,000 images with at least 5 captions per image, where more than 65% of the captions are written in English. We have used the same data splits from (Karpathy and Fei-Fei, 2015): 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. This dataset has been extensively employed in the recent years for image-text retrieval challenges.

3.2. Hyper-Parameters

We use the performance on validation data as a proxy for choosing the best hyper-parameters, and we perform a non-exhaustive hyper-parameter grid search. We employ Adam (Kingma and Ba, 2014) for optimization, given its capacity in adjusting per-weight learning rates during training. It performs a learning rate annealing in each iteration based on estimates of

first and second-order moments, leading to a deeper exploration of the search space. We use Adam’s default initial learning rate as suggested in (Kingma and Ba, 2014). In addition, we found it was beneficial to reduce the learning rate by 10× whenever the validation error *plateaus*. Inspired in (Vendrov et al., 2016), we use a batch size of 100 (hence, 99 contrastive examples) and margin $\alpha = 0.05$. Note that neither weight decay nor dropout were used, since we believe the loss function itself is enough to regularize the model by including several contrastive examples instead of using only hard-contrastive ones.

3.3. Evaluation Measures

For evaluating the results, we use the same measures as those in (Vendrov et al., 2016): $R@K$ (reads “Recall at K ”) is the percentage of queries in which the ground-truth term is one of the first K retrieved results. The higher its value, the better. We also show the results of *Med r* and *Mean r* , which represent respectively the median and mean of the ground-truth ranking. Since they are ranking-based measures, the smaller their values the better.

4. Results

All recent approaches developed for image-text alignment so far make use of word-embeddings, whilst ours is the first to generate and use character-level embeddings for this kind of task. Our results are compared to the state-of-the-art approaches that rely on word-embeddings, namely: MNLM (Kiros et al., 2014b), DVSA (Karpathy and Fei-Fei, 2015), FV (Klein et al., 2015), *m*-CNN (Ma et al., 2015), *m*-CNN_{ENS} (Ma et al., 2015), combine-skip-thought (Kiros et al., 2015), order-embeddings and order-embeddings (symm.) (Vendrov et al., 2016).

4.1. Quantitative Analysis

Table 2 depicts the results obtained for our 4 architectures along with the published results from the baselines. We first analyze the performance of all methods with respect to the image-to-text task. Note that our approach (Arch. D) outperforms the previous state-of-the-art by 0.5 regarding $R@1$, whereas for $R@10$ we obtain the same performance (88.9) as the method in (Vendrov et al., 2016). Considering the ranking measures, the

best approach according to the median ranking is DVSA (Karpathy and Fei-Fei, 2015), but note that such a result is not directly comparable to ours since they use a single fold as test set whereas we present the average results for all 5 folds like (Vendrov et al., 2016). With respect to the mean ranking, we once again achieve state-of-the-art results with the lowest ranking of all (5.6). If we take into consideration that the architecture in (Vendrov et al., 2016) is more complex than ours, besides relying on word-embeddings and hence requiring much more computational resources, we can see that our approach is clearly preferred over the method in (Vendrov et al., 2016), arguably establishing itself as the novel state-of-the-art for image-to-text retrieval.

Next, we analyze the performance of all methods with respect to the task of text-to-image retrieval. In this task, we only fail at outperforming the state-of-the-art regarding $R@1$, in which our best architecture is outperformed by a margin of 0.4 to (Vendrov et al., 2016). However, three of our architectures (Arch. B, C, and D) outperform the state-of-the-art considering $R@10$, with our best architecture reaching 87% of recall against 85.9% of the best baseline. The same behavior is verified in the mean ranking, where our best approach (Arch. D) reaches an average ranking of 7.3 versus 8.1 of the current state-of-the-art. Hence, we come to the same conclusion as before: our approach, besides simpler and lighter, is also often better than the previously-published approaches, hence establishing itself as the novel state-of-the-art for text-to-image retrieval.

As a secondary note, let us consider Arch. A, our smallest architecture. Even though it did not achieve the best results regarding the 4 architectures we propose, it still reaches quite competitive results, which is *per se* quite surprising considering that it comprises a single convolutional layer and at least fifteen times fewer parameters than the the architecture in (Vendrov et al., 2016). In addition, this very simple model is actually capable of outperforming the following more complex methods: i) combine-skip-thoughts, which had to be pre-trained for an entire month in a huge corpora and makes use of three GRUs with 1024 units each for extracting features from texts; ii) DVSA, which employs R-CNN based features from 20 regions in the images (i.e., features extracted from networks that were trained to perform object detection and localization), a much more complex and computationally demanding approach when compared to simply extracting traditional features from image-classification based networks; iii) FV, that makes use of Fisher Vectors handcrafted features for text encoding; and iv) *m*-CNN and its ensemble version, which are CNN-based architectures that learn similarity metrics.

Figure 2 shows the trade-off analysis between model complexity and predictive performance of our models and the work in (Vendrov et al., 2016). The ideal position is in the upper-left position (largest recall and smallest number of parameters). For generating this visualization, we computed the number of trainable parameters for each method. Note that the word-embeddings are considered trainable parameters as well. For image-to-text retrieval both Arch. C and D are virtually in the same horizontal line than the architectures in (Vendrov et al., 2016), while requiring only a fraction of the processing needed by the baseline.

Regarding the text-to-image results, note that Arch.s-[B,C,D] present superior performance despite being up to ten-fold lighter (Arch. C with 1.2M versus OE with 12M).

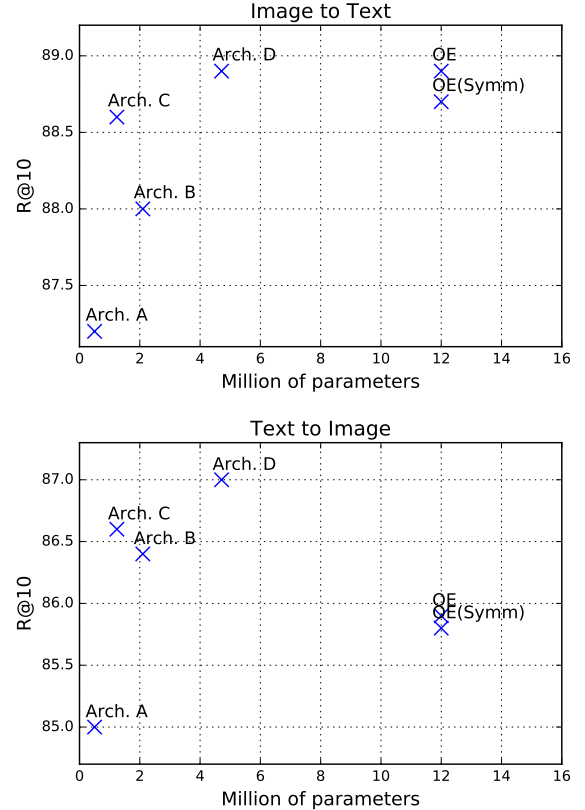


Fig. 2. Trade-off between model complexity (#parameters) and predictive performance ($R@10$) in the test set. Results are for both image-to-text (above) and text-to-image retrieval (below).

4.2. Qualitative Analysis

In this section we discuss qualitative results by analyzing images and text that were retrieved by using our best model, namely Arch. D. Figure 3 shows cases of image retrieval from textual queries. This is equivalent to image search taking into consideration the semantics of the text. Note that by using our strategy, it is possible to recover images from highly-detailed queries. For instance, our method is capable of learning spatial relations, objects, differentiate human genre (“*a man is standing...*”, “*a girl is riding...*”), age-related information (“*baby is on top of...*”), counts (“*two giraffes...*”), and also complex relationships (“*a fire hydrant with two eyes...*”, or “*a bird with red eyes on top of a tree...*”).

Figure 4 depicts qualitative results for image-to-text retrieval. In this experiment, we performed queries based on images to retrieve the most similar captions in the multimodal embedding space. We show the three most-correlated captions indicated by our model for four randomly-chosen images. Note that, in all images, at least two ground-truth captions were retrieved. We did not highlight the wrongly-retrieved texts on purpose so we give the reader a chance to find them.



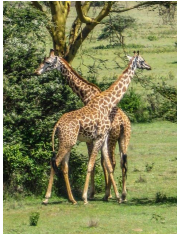


| | | | | | |
|-----------------------|--|---|---|--|---|
| Textual Query | A man standing in front of a metallic refrigerator freezer with a baby on top of it. | A girl is riding her bike down the street. | Two giraffe standing next to each other on a grassy field. | A fire hydrant in which someone put two eyes on it. | A bird with red eyes perched on top of a tree branch. |
| First Retrieved Image |  |  |  |  |  |

Fig. 3. Text-to-image retrieval results.





| | | | | |
|--------------------------|--|---|--|--|
| Query Image |  |  |  |  |
| Top-3 Generated Captions | 1) An old car with surfboards on the top. 2) A colorful vintage truck splattered with paint sits in a field. 3) A car with some surfboards in a field. | 1) A tour bus with advertisement on the side of it 2) A white blue and green passenger bus with an advertisement on the side 3) A large bus with a political ad on the side of it | 1) A black kitten laying down next to two remote controls 2) A black cat is lying next to a remote control 3) A couple of phones are on a blanket next to a black kitten | 1) A row of motorcycles parked next to each other 2) A row of parked motorcycles sitting on the side of a road 3) There are many motorcycles that are parked by each other |

Fig. 4. Image-to-text retrieval results.

Finally, in Figure 5 we show failure cases of image-to-text retrieval. We display the ground-truth captions in green (when recovered). Captions retrieved in the first image are reasonable, given that the model only confused the airplane colors. Descriptions for the second image detail the presence of oranges and a nonexistent glass bowl. Finally, the third and fourth image descriptions are interesting cases where the retrieved texts are plain wrong, probably due to the large variety of images/descriptions in the dataset. Particularly, the fourth image (insect inside a cage) depicts a very unusual scene.

5. Related Work

Recent studies have proposed methods for bidirectional image and sentence retrieval (Kiros et al., 2014b,a; Karpathy and Fei-Fei, 2015; Klein et al., 2015; Kiros et al., 2015; Socher et al., 2014; Vendrov et al., 2016) and automatic image captioning (Donahue et al., 2015; Xu et al., 2015; Vinyals et al., 2015; Chen and Lawrence Zitnick, 2015).

Kiros et al. (2014b,a) propose a novel approach to deal with the problem of image caption generation by using an encoder-decoder architecture. For the encoding part of the architecture, a jointly-trained image-sentence embedding is generated, where sentences are encoded using GRU networks. Their approach is designed to map images and sentences onto the same embedding space so that correlated image-caption vectors present a high cosine similarity score. In addition, they also propose a language model for generating captions based on structure and content.

Karpathy and Fei-Fei (2015) propose an architecture that makes use of features from detection/localization-based systems.

The idea is aligning image regions with a proper sentence fragment. This architecture learns inter-modal correspondences between image and sentence fragments, despite discarding global relations.

Klein et al. (2015) present new Fisher Vectors that were derived from Hybrid Gaussian-Laplacian Mixture Models in order to provide sentence representations. They use Fisher Vectors as a pooling strategy of word-embeddings for representing text. Such representations are used in Laplacian Mixture Models trained with an expectation-maximization algorithm.

Kiros et al. (2015) introduce a sentence embedding strategy, namely *skip-thought vectors*, which is trained with no supervision over a very large book corpus to learn generic sentence representations. They build an encoder-decoder architecture that aims to reconstruct the surrounding sentences of a given encoded passage in a book. The authors encourage the use of a combined architecture of three skip-thought based models trained with different strategies (combine-skip-thought). This ensemble approach often leads to better results, though largely increasing the computational requirements.

Ma et al. (2015) propose a multimodal convolutional neural network for learning a matching score between image representations and text represented by sequences of word-embeddings, by jointly convolving the word-embeddings and image features. The learned similarity score predicts whether a pair is correlated or not. Architectural variations are explored and the best results are achieved with an ensemble version.

Finally, Vendrov et al. (2016) propose the sentence order-embeddings. Those embeddings aim to preserve the partial order structure of a visual-semantic hierarchy. It allows learning

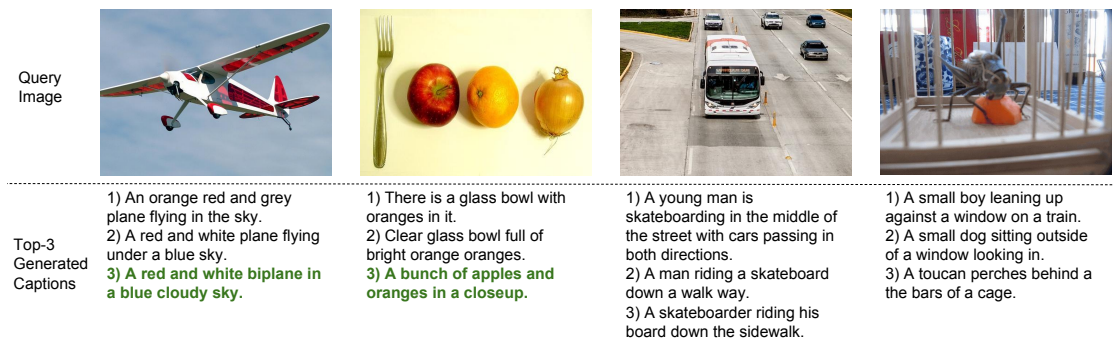


Fig. 5. Failure cases of image-to-text retrieval.

ordered representation by applying order-penalties, and they show that asymmetric measures are better suited for image-sentence retrieval tasks. Their architecture is virtually the same of the one introduced in (Kiros et al., 2014b), with only the loss function being changed to consider the order violations.

6. Conclusions

We presented a simpler and faster architecture capable of learning textual embedding based on raw characters and order-embeddings for image-text alignment. Even though it is conceptually a much simpler architecture than those found in related work, our approach is capable of achieving state-of-the-art results in both text-to-image and image-to-text tasks. A promising future work direction is to analyze the performance of our approach in problems such as multilingual data analysis and caption generation for videos.

Acknowledgments

We would like to thank Motorola Mobility, Google, and Brazilian research agencies CNPq, CAPES, and FAPERGS for funding this research.

References

- Chen, X., Lawrence Zitnick, C., 2015. Mind’s eye: A recurrent visual representation for image caption generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), pp. 2422–2431.
- Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2015. Gated feedback recurrent neural networks, in: International Conference on Machine Learning (ICML 2015), pp. 2067–2075.
- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9, 1735–1780.
- Karpathy, A., Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), pp. 3128–3137.
- Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization, in: International Conference on Learning Representations.
- Kiros, R., Salakhutdinov, R., Zemel, R., 2014a. Multimodal neural language models, in: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 595–603.
- Kiros, R., Salakhutdinov, R., Zemel, R.S., 2014b. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR abs/1411.2539*. URL: <http://arxiv.org/abs/1411.2539>.
- Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Skip-thought vectors, in: Advances in Neural Information Processing Systems (NIPS 2015), pp. 3294–3302.
- Klein, B., Lev, G., Sadeh, G., Wolf, L., 2015. Associating neural word embeddings with deep image representations using fisher vectors, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4437–4446. doi:10.1109/CVPR.2015.7299073.
- Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context, in: European Conference on Computer Vision (ECCV 2014).
- Ma, L., Lu, Z., Shang, L., Li, H., 2015. Multimodal convolutional neural networks for matching image and sentence, in: International Conference on Computer Vision (2015).
- Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L., 2014. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *CoRR abs/1412.6632*. URL: <http://arxiv.org/abs/1412.6632>.
- Mesnil, G., He, X., Deng, L., Bengio, Y., 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding, in: Interspeech, pp. 3771–3775.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Advances on Neural Information Processing Systems (NIPS 2013), USA. pp. 3111–3119.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 211–252.
- Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations (ICLR 2015).
- Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y., 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics (TACL)* 2, 207–218.
- Vendrov, I., Kiros, R., Fidler, S., Urtasun, R., 2016. Order-embeddings of images and language, in: International Conference on Learning Representations (ICLR 2016).
- Venugopalan, S., Hendricks, L.A., Rohrbach, M., Mooney, R.J., Darrell, T., Saenko, K., 2016. Captioning images with diverse objects. *CoRR abs/1606.07770*. URL: <http://arxiv.org/abs/1606.07770>.
- Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2015), pp. 3156–3164.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning (ICML 2015), pp. 2048–2057.