NEURAL CAPTION-IMAGE RETRIEVAL

Junyang Qian (junyangq), Giacomo Lamberti (giacomol)

1 INTRODUCTION

In the modern era, an enormous amount of digital pictures, from personal photos to medical images, is produced and stored every day. It is more and more common to have thousands of photos sitting in our smart phones; however, what comes with the convenience of recording unforgettable moments is the pain of searching for a specific picture or frame. How nice it would be to be able to find the desired image just by typing one or few words to describe it? In this context, automated caption-image retrieval is becoming an increasingly attracting feature, comparable to text search.

In this project, we consider the task of content-based image retrieval and propose effective neural network-based solutions for that. Specifically, the input to our algorithm is a collection of raw images in which the user would like to search, and a query sentence meant to describe the desired image. The output of the algorithm would be a list of top images that we think are relevant to the query sentence. In particular, we train a recurrent neural network to obtain a representation of the sentence that will be properly aligned with the corresponding image features in a shared high-dimensional space. The images are found based on nearest neighborhood search in that shared space.

The paper is organized as follows: first, we briefly summarize the most relevant work related to our task; then, we describe the dataset employed for training and the features of our problem. Subsequently, we introduce our models, namely a multi-response linear regression model and a deep learning method inspired by Vendrov et al. (2015). In the results section, we evaluate the accuracy of the different models by computing the Recall@K measure, i.e. the percent of queries for which the desired image is among the top K retrieved ones. We also perform some error analysis to study the influence of the length of the caption to the accuracy of the results. Finally, we conclude with some remarks and ideas for future research.

2 RELATED WORK

Under the umbrella of multimodal machine learning, caption-image retrieval has received much attention in recent years. One main class of strategies is to learn separate representations for each of the modalities and then coordinate them via some constraint. A natural choice of constraint is similarity, either in the sense of cosine distance (Weston et al., 2011; Frome et al., 2013) or the Euclidean distance. Recent advancement of neural networks enables one to build more sophisticated language and images models based on more informative embeddings. For example, Socher et al. (2014) exploits dependency tree RNN to capture compositional semantics.

A different class of constraint considered in Vendrov et al. (2015) is order embedding. There the features are constrained to have non-negative values, and the smaller the feature values are, the more abstract that corresponding concept is. For example, the universe is assumed to be at the origin. In the context of caption-image retrieval, a caption is assumed to be an abstraction of the image and should be enforced to have smaller feature values. That is particularly useful for hypernym prediction. For caption-image retrieval, however, the performance isn't much different from the normal Euclidean distance and it doesn't seem very robust to different architectures and specifications in our experiments.

More recently, there is another line of work that tries to improve retrieval performance with the use of generative models. In Gu et al. (2018), they propose an "imagine" step where the target item in the other modality is predicted based on the query and then a more concrete grounded representation is obtained. However the training would be much slower compared with previous methods.

Under the hood of most state-of-the-art models, the choice of pretrained features/embeddings plays an important role. We use VGG-19 (Simonyan & Zisserman, 2014) as used by Vendrov et al. (2015). Gu et al. (2018) claims ResNet-152 (He et al., 2016) can further improve the retrieval performance.

3 Dataset and Features

We train our models using the Microsoft COCO dataset (Lin et al., 2014), which contains $123,287$ images in total. Each image is associated with 5 human-annotated captions. We use the same split as in Karpathy & Fei-Fei (2015): $113,287$ for training, $5,000$ for validation and test respectively. An example from the dataset is shown below.

- Three teddy bears laying in bed under the covers.
- A group of stuffed animals sitting next to each other in bed.
- A white beige and brown baby bear under a beige white comforter.
- A trio of teddy bears bundled up on a bed.
- Three stuffed animals lay in a bed cuddled together.

To represent images, a common choice is to use a pretrained image model as a feature extractor and use the last layer of the forward pass as the representation. In the present work, we employ the $fc7$ features of the 19-layer VGG network (Klein et al., 2015). In particular, each image is cropped to generate 10 images, which are then passed through the VGG network; the resulting outcomes are averaged to produce a single high-dimensional feature vector.

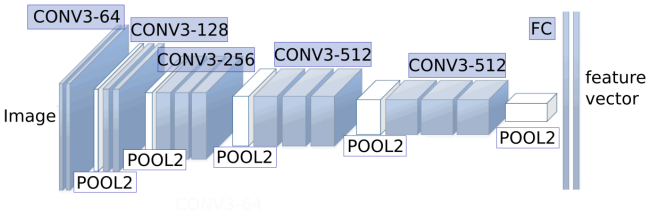

Figure 1: 19-layer VGG network, without last fully-connected layer

For text, or specifically words, a widely used representation is pretrained word vectors, such as the skip-gram model (Mikolov et al., 2013) or the GloVe model (Pennington et al., 2014). In our case, we employ the GloVe word vectors to represent each word of the caption.

4 Methods

In this section, we describe the methods that we use for this task. They include a traditional supervised method based on multiple-response linear regression, and methods based on neural networks.

4.1 Baseline Method

In multimodal machine learning, a common approach is coordinating the representations of different modalities so that certain similarity among their respective spaces are enforced. Our task involves texts and images. To represent the captions, we simply average the GloVe vectors relative to the words of the sentence, though more sophisticated methods exist. Let f_{GloVe} be the sentence features and f_{VGG} be the image features coming from the VGG network. In order to encourage similarity between these two different types of representation, we would like to find a weight matrix such that:

$$\hat{W}_{c,i} = \arg\min_{W} \sum_{k} \|f_{\text{VGG}}(i_k) - W \cdot f_{\text{GloVe}}(c_k)\|_2^2.$$

This is known as multi-response linear regression. As a generalization of linear regression, it has closed-form solution or can be solve by stochastic gradient descent when we have a large dataset. At test time when we are given a caption $c^{(t)}$, we compute the caption feature vector $f_{\text{GloVe}}(c^{(t)})$, and find the image(s) closest to that:

$$i^{(t)} = \arg\min_{i'} \|f_{\text{VGG}}(i') - \hat{W}_{c,i} \cdot f_{\text{GloVe}}(c^{(t)})\|_2^2.$$

4.2 MULTIMODAL NEURAL NETWORK METHODS

Our method is inspired by Vendrov et al. (2015) where they use RNN for language modeling and a pretrained VGG model to generate static image features. We borrow some notations from that paper. Given a set of image-caption pairs, the goal is to learn a similarity score between an image (i) and its caption (c):

$$S(i, c) = - \| f_i(i) - f_c(c) \|_2^2 , \tag{1}$$

where f_i and f_c are embedding functions for images and captions, respectively. There is a negative sign since we would like larger S to indicate more similarity. For the purpose of contrasting correct and incorrect matches, we introduce negative examples that comes handy in the same training batch. The cost can thus be expressed as

$$\sum_{(c,i)} \left(\sum_{c'} \max\{0, \alpha - S(c, i) + S(c', i)\} + \sum_{i'} \max\{0, \alpha - S(c, i) + S(c, i')\} \right), \tag{2}$$

where (c, i) is the true caption-image pair, c' and i' refer to incorrect captions and images for the selected pair. Therefore, the cost function enforces positive (i.e. correct) examples to have zero-penalty and negative (i.e. incorrect) examples to have penalty greater than a margin α.

Feature extraction for both modalities is similar to the baseline. The embedding function f_i is obtained by opportunely weighting the outcome before the output layer of the VGG network and f_c takes the last state of a recurrent neural network (RNN) with gated recurrent unit (GRU) activation functions (Cho et al., 2014), i.e.

$$f_i(i) = W_i \cdot f_{\text{VGG}}(i), \qquad f_c(c) = f_{\text{GRU}}(c). \tag{3}$$

where W_i is a $n \times 4096$ matrix of weights to be trained and n is the number of features of the embedding space. The embedding function f_c is now the outcome of a recurrent neural network (RNN) with gated recurrent unit (GRU) activation functions (Cho et al., 2014):

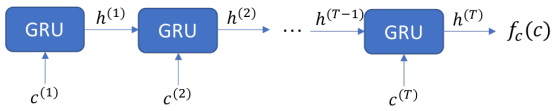

Figure 2: Recurrent neural network to process captions.

We observe that in Vendrov et al. (2015), all word embeddings are treated as parameters and trained from scratch. The training can adapt the embedding parameters to this specific task, but it is also limited by the semantic information contained in the training corpus. We find that the captions in the training set are mostly short phrases or sentences. As a result, the trained embeddings can miss more sophisticated implication within and between the words. Instead, pretrained word vectors on a larger corpus like Wikipedia enables one to exploit richer information encoded in a variety of contexts. Moreover, we can either use them as fixed, non-trainable embeddings or use them as an initialization and fine tune them for our specific task. The latter one is adopted in our method.

In addition, we explore the usage of a different modules in the RNN, namely long short-term memory (LSTM) and different RNN architectures such as stacked bidirectional RNNs.

5 EXPERIMENTS

Metric The metric we use in this project is Recall@K (R@K). Given a list of predicted rankings $r_i (1 \leq i \leq m)$ for m images based on their corresponding input captions, we define

$$\text{R@K} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\{1 \leq r_i \leq K\}.$$

We should notice that this metric also depends on the size of the image database. For example, searching over a one-million-image database is clearly harder than a one-hundred database. In this project, we focus on the size of 1K images. In addition, we will also look at some conditional metrics, such as the length of the caption, to better understand the results.

Hyperparameters We started with a set of hyperparameters suggested in Vendrov et al. (2015), experimented other combinations and chose the optimal based on the performance on the validation set. We train all the models at most 50 epochs and do early stopping when necessary, i.e. the model appears to overfit. Specifically, the training data are divided into random mini-batches of 128 examples and trained using Adam optimizer with a learning rate of 0.05. Moreover, the CNN output has $4,096$ dimensions and the word vectors has 300 dimensions. The shared embedding space for both captions and images has $1,024$ dimensions. In the experiments, a margin of 0.1 in (2) helps us achieve the best performance.

Results In Table 1, we show the performance of different methods evaluated on the test set. The Mean r column computes the average rank of the correct image match. We've listed two baseline results. The pure baseline is based on the method described in the previous section. The Baseline + Weight method computes the average feature vectors of all *five* heldout captions for each test image, while the other methods only use one of them. Although it is a little unfair to the other methods in comparison, it could still be an option in practice. Such averaging in the baseline method is equivalent to assigning different weights to the words and the key words that appear repeatedly in the five captions are automatically highlighted. It is worth considering how to incorporate a user-provided weighting in other nonlinear methods.

Method	**R@1**	**R@10**	**Mean r**
Baseline	10.3	17.1	176.1
Baseline + Weight	19.8	65.5	10.5
GRU (Vendrov et al., 2015)	36.3	85.8	7.6
GRU + GloVe	**37.0**	**86.8**	**7.3**
LSTM + GloVe	35.4	86.2	7.5

Table 1: Recall@K (in %) results on the test set. For GloVe, we use it as initialization of the word vectors, and fine-tune it in the training.

We see that GRU-RNN with GloVe initialization does the best, LSTM-RNN the second, and both better than the results reported in Vendrov et al. (2015). Their architectures are very similar, and we see that using pretrained word vectors indeed help improve the retrieval quality.

Figure 3 on the left shows the evolution of the R@10 measure over the epochs; both GRU and LSTM models have similar behavior: after just a single epoch the accuracy is already higher than the baseline method, and after ~ 30 epochs both curves seem to have reached a plateau.

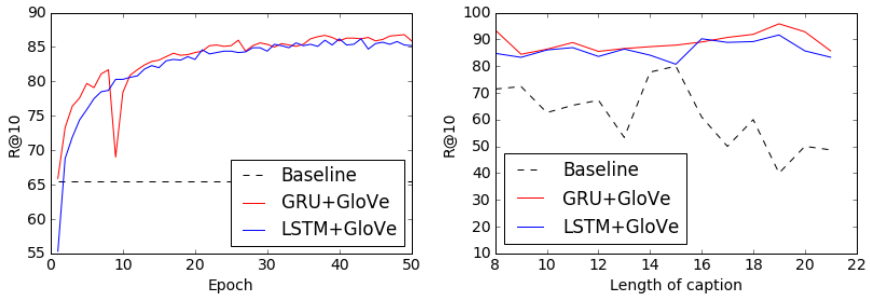

Figure 3: $R@10$ measure variation throughout the epochs (left), and its dependency on the length of the caption (right).

On the right, we compare the sensitivity of the methods on the length of the caption. Initially we thought that long sentences would be a challenge for the RNN and the retrieval quality would signif-

icantly degrade as the input sentence becomes longer. The baseline method clearly fails, as the average of all feature vectors will mask the real important ones. From the plot, we see the GRU/LSTM networks, however, are capable of dealing with long sequences. One possible explanation is that in this dataset, it is rare to have a caption that is unnecessarily long. Long captions there usually carries more information about the image, and in this sense will help the model to identify the correct image. It is likely that the curve on the right shows such two-sided tradeoff by long sequences.

We also compare the baseline method and the neural network solution through some real examples.

Figure 4: Query: **[dog]**

Figure 5: Query: **[a guy riding a bike next to a train]**

We see that the baseline model works well for simple queries like single-word object names. However for longer captions as in Figure 5, it is unable to capture multiple objects and their interactions. The architecture of RNN with GRU/LSTM has the mechanism of adaptively memorizing the words it has seen. That can help identify minor details and complex relationship within the image.

6 Future Work

In this project, our emphasis is more on language models because as a first step we would like to accurately identify the semantics implied by the query. On the image side, we only represent each by its features extracted from a pretrained network. Although we see the image feature is able to capture small details in the image, it can still be the bottleneck as our language model becomes more sophisticated. In the future, we would like to endow a dynamic attention mechanism so that the model will be able to choose adaptively the region(s) to focus on in the image. This might be done either by including some pretrained features in the lower layers or by computing features on sub-regions of the image. There are some initial attemps in this direction such as Chen et al. (2017) and we would like to further develop on that. Another direction we are interested in but don't have enough time to explore in this project the use of generative model to improve retrieval performance. As mentioned in (Gu et al., 2018), that can help us learn more local grounded features than global abstract features.

Link to the code: `https://github.com/giacomolamberti90/CS229_project`

REFERENCES

Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. Amc: Attention guided multi-modal correlation learning for image search. *arXiv preprint arXiv:1704.00763*, 2017.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7181–7189, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4437–4446, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218, 2014.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, pp. 2764–2770, 2011.