International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

Image Captioning using Multimodal Embedding

Rachit Jain[1], Samarth Joshi[2]

[1,2]B.Tech students, Department of Computer Science
Bharati Vidyapeeth's College of Engineering, New Delhi, India

---

Abstract

Image captioning still remains a conundrum as it not only focuses on extraction of the visual semantics of a given image but also on combination techniques from the domain of natural language processing. Various models capable of captioning an image using the semantic features and the style of the text corpus are unable to combine the visual semantics of two different images being fed simultaneously. We propose a novel methodology wherein multiple images sharing similar context can be used to generate a single story/caption. Our alignment model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through multimodal embedding. We, then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. The paper encompasses on extracting the visual semantics using existing deep learning architecture followed by a pipeline of NLP model of skip thought vectors. This can be further used along with a matrix of TF-IDF values based on the text corpus extracted from various books. After training our model, we extract and evaluate our vectors on semantic readiness with linear models. The results compare two different models- one based on TF-IDF matrix values and other being skip thought vector representation of bag of words, each considering 2 grams at a time.

Key Words: **visual semantics, natural language processing, convolution neural networks, image regions, recurrent neural network, multimodal embedding, deep learning architecture, skip thought vectors, TF-IDF values**

1. INTRODUCTION

Describing an image is probably the easiest task for a human being. This remarkable ability of humans to describe an image just by looking at it can serve as a motivation for visual recognition models. However, achieving remarkably accurate results has proven to be an elusive task for a machine learning model. The vocabularies of visual concept are more convoluted as compared to the impeccable descriptions by humans. The field of visual recognition has shown various models that achieve feature extraction.

Ever since the starting of ImageNet challenge, there has been an exponential increase in the Convolution architecture that has beckoned the task of image recognition as well as object detection. Plenty of work has been done in visual recognition which focuses on labeling of images with a fixed set of visual categories. The main focus of these works has been to describe a compound multiplex visual scenario in a single line sentence/caption. These models can therefore be of immense significance in describing the visual semantics of an image in form of short sentences. Some pioneering approaches that address the challenge of generating image descriptions have been developed [1, 2]. However, these models often rely on hard-coded visual concepts and sentence templates, which imposes limits on their variety. In this paper, we aim at taking this task to the next level by combining the visual descriptions into a single story (which shares the context similar to the images it has seen at the input). We combined the two well-known architectures namely Neural Talk 2 by Andrej Karpathy for extracting image captions and Skip thoughts, which is an unsupervised learning algorithm to encode these captions. Neural talk 2 is trained on Flicker8K, Flicker30K and MSCOCO datasets while the skip thoughts has a dataset of 16 different genres like romance, fantasy, science fiction, teen, etc. The rest of the paper includes a description of these architectures followed by the approach used by us to combine the captions. We use two approaches- TF-IDF matrix representation and Skip thought vector representation and then compare the results achieved.

2. INDIVIDUAL MODELS

We have used the hybrid model of two existing architectures to generate and combine the annotations to generate meaningful sentences. In the following section, we describe these two architectures followed by the approach used by us to combine these annotations in form of dense vectors.

2.1 Neural Talk 2 [7]

There are two main contributions that this architecture has provided. Firstly, development of a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe. Secondly, introduction of a multimodal Recurrent Neural Network architecture that takes an input image and generates its description in text format.

This model takes a set of images as input and their corresponding sentence descriptions (Figure 2). Firstly, it presents an approach that aligns the sentence snippets to the visual regions through a multimodal embedding. It then treats these correspondences as training data for a second multimodal Recurrent Neural Network model that learns to generate the snippets.

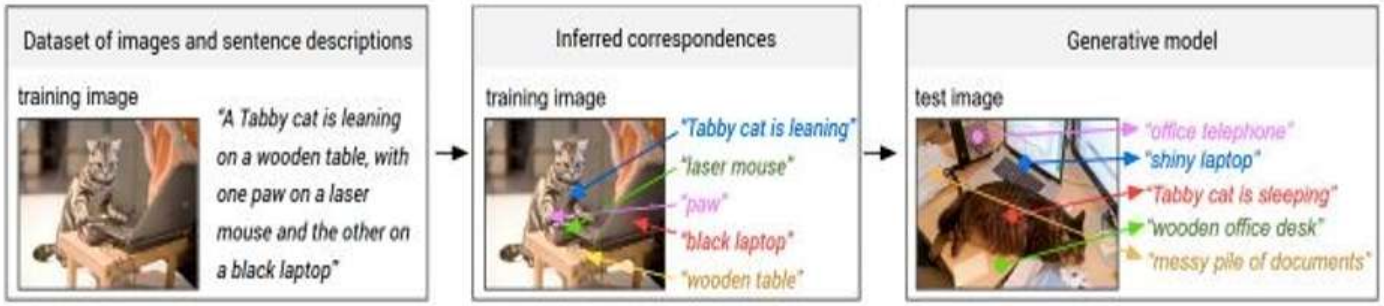

Fig 1. A dataset of images and their sentence descriptions is given as input and the model infers correspondences and learns to generate novel descriptions.

As it is known that sentence descriptions make frequent references to objects and their attributes. Thus, it follows the method of Girshick et al. [5] to detect objects in every image with a Region Convolutional Neural Network (RCNN). The CNN is pre-trained on ImageNet [6] and fine-tuned on the 200 classes of the ImageNet Detection Challenge [4].

Following Karpathy et al. [3], we use the top 19 detected locations in addition to the whole image and compute the representations based on the pixels I_b inside each bounding box as follows:

$$v = W_m[CNN_{\theta c}(I_b)] + b_m$$

$$CNN_{\theta c}(I_b) = \text{Image transformation operation}$$

The above approach is simply a multilayer perceptron with CNN layer consisting of nearly 60 million parameters. The matrix W_m has dimensions h × 4096, where h is the size of the multimodal embedding space (h ranges from 1000-1600 in our experiments). Every image is thus represented as a set of h-dimensional vectors $\{v_i \mid i = 1 \ldots 20\}$.

To address the part of intermodal relationship, it proposes a Bidirectional Recurrent Neural Network (BRNN).

Using a sequence of N words (encoded in a 1-of-k representation) it transforms each one into an h-dimensional vector.

However, the representation of each word is enriched by a variably-sized context around that word. The mathematical representation of BRNN is as follows:

$$d_t = W_w * I_t$$
$$e_t = f(W_e * d_t + b_e)$$
$$h_t^f = f(e_t + W_f h_{t-1}^f + b_f)$$
$$h_t^b = f(e_t + W_b h_{t-1}^b + b_b)$$
$$s_t = f(W_d(h_t^f + h_t^b) + b_d)$$

The BRNN consists of two independent streams of processing, one moving left to right (h_t^f) and the other right to left (h_t^b) (see Figure 3 for diagram). The final h-dimensional representation s_t for the t^{th} word is a function of both the word at that location and its surrounding context in the sentence. Now the objective is to focus at the level of entire images and sentences to formulate an image-sentence score as a function of the individual region-word scores. Intuitively, a sentence-image pair should have a high matching score if its words have a confident support in the image. The model of Karpathy et a. [3] interprets the dot product $v_i^T * s_t$ between the i^{th} region and the t^{th} word as a measure of similarity and uses it to define the score between image k and sentence l as follows:

It then computes a sequence of hidden states ($h_1 \ldots h_t$) and a sequence of outputs ($y_1 \ldots y_t$) by iterating the following recurrence relation for t = 1 to T:

$$b_v = W_{hi}[CNN_{\theta c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \gamma(t = 1) \odot b_v)$$
$$y_t = softmax(W_{oh}h_t + b_o)$$

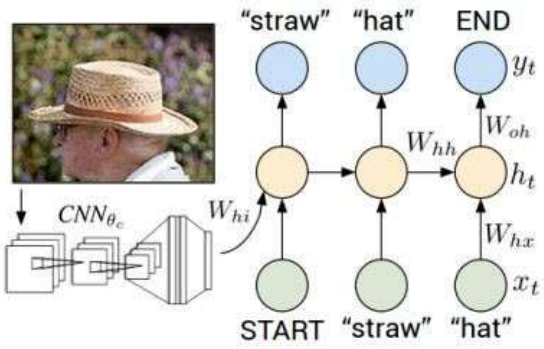

Fig 2. Old Neural Talk2 Model

$$s_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} max(0, v_i^T * s_t)$$

g_k = *set of image segments*
g_l = *set of sentence fragments*
The final representation after hyperparameter could be represented in the form of:

$$s_{kl} = \sum_{t \in g_l} max_{i \in g_k}(v_i^T * s_t)$$

Where, s_t = *single best image region*
For k=l the loss comes out to be:

$$C(\theta) = \sum_T \left[\sum_i max(0, s_{il} - s_{ik} + 1) + \sum_i max(0, s_{it} - s_{kk} + 1) \right]$$

We can interpret the quantity * as the $v_i^T * s_t$ un-normalized log probability of the t^{th} word describing any of the bounding boxes in the image. As the purpose is to annotate each bounding box with a sequence of words it actually represents, so it uses true alignment of these words as a latent variable in Markov Random Field (MRF). The MRF considers the binary interaction between two neighboring words to be aligned in the same region. Thus, it takes a sentence with N words and an image with M bounding boxes and defines latent alignment variables $a_j \in \{1 \ldots M\}$ for $j = 1 \ldots N$ and formulate an MRF in a chain structure along with the sentence.

Here, we use β as a hyper-parameter that controls the affinity towards longer word phrases.

For captioning during training, the Multimodal RNN takes the image pixels I and a sequence of input vectors ($x_1 \ldots x_T$). This can be described as follows:

$$E_a = \sum_{j=1 \ldots N} \omega_j^u(a_j) + \sum_{j=1 \ldots N-1} \omega_j^b(a_j, a_{j+1})$$

Here, $\omega_j^u(a_j = t) = v_i^T * s_t$

$$\omega_j^b(a_j, a_{j+1}) = \beta[a_j = a_{j+1}]$$

2.2 Skip-Thoughts

Skip-thoughts is basically an encoder-decoder framework whose aim is to represent every sentence as a skip-thought vector in which encoder accepts a middle sentence and then one decoder generates the previous sentence while the other one generates the future (next) sentence for the given middle sentence. Skip-thought vectors are used to generate vectors for every sentence to know which sentences are semantically similar. Once the model has been trained, the vector representation of a sentence can be extracted from the learned encoder by inputting the sequence of tokens that makes up the sentence. The encoder-decoder model is composed of gated recurrent units (GRUs) [9]. In order to get vector representation of sentences, we have employed the already trained model provided by kiros et al (2015) [8].

This pre-trained model creates a 4800 dimensional vector for each sentence by concatenating the vector representations from the uni-skip model and the bi-skip model. Uni-skip model encodes the input tokens of a sentence in their original order, and outputs a 2400 dimensional vector. This uni-skip model is unidirectional encoder. The bi-skip model is a bidirectional model that encodes the input tokens of a sentence in their original order and in their reversed order, outputting a 1200 dimensional vector for each direction.

The resemblance between two sentences is then computed with the help of cosine similarity. Cosine similarity of both the sentences is taken in order to get their vector representations. This whole process is described as skip thoughts.

2.3 TF-IDF Matrix

In this approach each sentence in a pair of sentences is depicted as vector, where each dimension corresponds to a word type. In TF-IDF, each dimension holds the TF-IDF weight for the corresponding type in the sentence. IDF values are calculated over a 2015 dump of English Wikipedia from 1 September 2015, which was pre-processed using wp2txt1 to remove markup. Then, the similarity between the two sentences is calculated as the cosine between vectors depicting them. The documents are tokenized using an approach provided by Speriosu et al. (2011) [11] — the text is first split based on whitespace; for each token, if it contains at least one alphanumeric character, then all leading and trailing non-alphanumeric characters are stripped. Stop words are removed based on a stop word list and case folding is applied [10].

3. ALGORITHM AND FLOWCHART

We combined Neuraltalk2 architecture with the two approaches mentioned above:

(i) Skip Thought Vector Matrix

(ii) TF-IDF Matrix

As each of the caption generated by the first model captures the dense representation of the images, we can use the skip thought vector of the corresponding sentences to generate the context being used in them. Each of the vector representing one sentence is converted to skip thought vector and arranged along the rows of the matrices and henceforth keeping the word values filled whereas keeping the other values as zeroes (Sparse matrix). The generated matrix is then combined with the matrix generated using the training phase of the language model. The dot product gives the cosine similarity between the two and thus activating the words that are similar in context of the combined sentences.

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

$$\begin{aligned}\mathrm{idf} &= -\log P(t|d) \\ &= \log \frac{1}{P(t|d)} \\ &= \log \frac{N}{|\{d \in D : t \in d\}|}\end{aligned}$$

$$\text{tf-idf}_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t.$$

$$\mathrm{Score}(q,d) = \sum_{t \in q} \text{tf-idf}_{t,d}.$$

Similarly we evaluate TF-IDF matrix with the language model to get the resultant matrix. The final sentence is thus accumulated using the log likelihood probability of each words from the bag of words considering n-words (n=3) at a time.

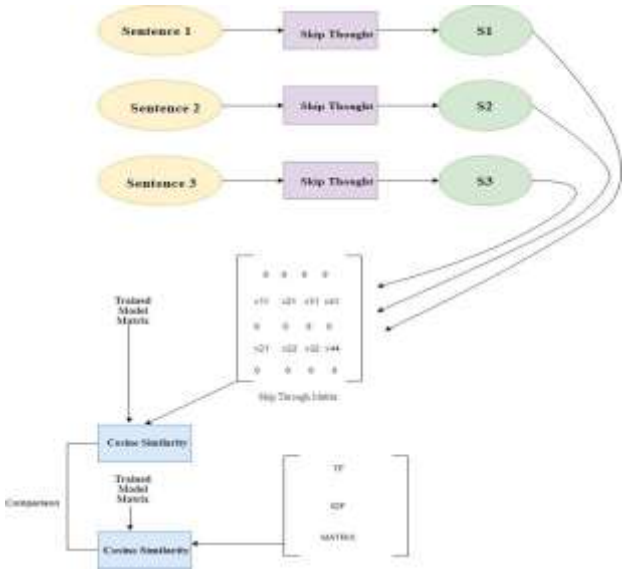

Fig 3. Flowchart of our model

4. RESULTS

We ran our model for both the techniques discussed above. For TF-IDF model, the RMSE and MAE values were 44.3 and 56.4 respectively. For Skip-thought model, the values of RMSE and MAE were 44.1 and 59.6 respectively.

Flickr30K- Dataset				
Model	TF-IDF		Skip thought	
	RMSE	MAE	RMSE	MAE
SDT-RNN	68.2	85.8	78.4	63.5
Our model	44.3	56.4	44.1	59.6
BRNN	66.92	75.67	56.42	79.93
DeFrag	42.16	58.2	45.77	61.46
MSCOCO-Dataset				
SDT-RNN	63.53	80.11	72.71	61.5
Our model	50.31	58.4	42.23	51.6
BRNN	63.23	78.6	56.4	49.9
DeFrag	82.3	61.2	49.7	64.86

Table 1. RMSE and MAE errors of evaluation over Flicker30k and MSCOCO

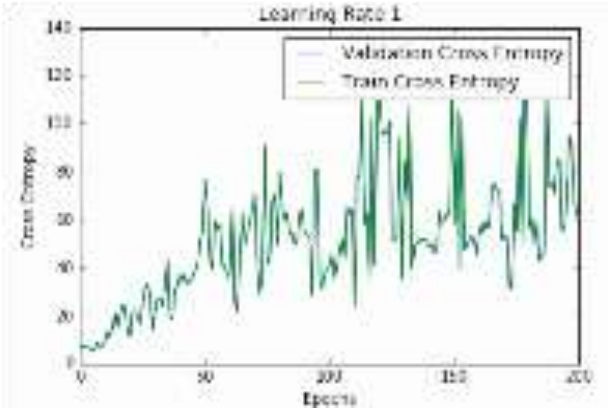

Fig 4. Cross entropy vs learning rate

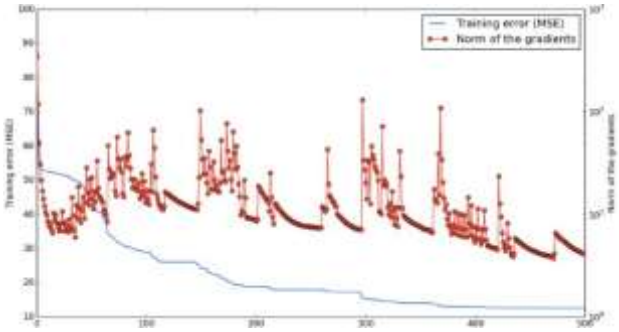

Fig 5. MSE and gradient norm vs epoch

Fig 6. Generated stories

5. CONCLUSION

The best results were obtained using the skip thought vector approach (to represent two sentences and further combining them using semantic relatedness- Cosine similarity). We further aim at improving our model by using the fluid segmentation technique which is the current state-of-the-art algorithm for image recognition. The applications of this model are manifold. It can help in generating reports of crime investigations, automating notes generation from video lectures, helping the patients of autism in medical diagnosis, medical imaging and many more.

6. REFERENCES

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV. 2010.

[2] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, 2011.

[3] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. arXiv preprint arXiv:1406.5679, 2014.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[7] Karpathy, A., & Johnson, J. (2015). Neuraltalk2.

[8] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, Curran Associates, Inc., pages 3276–3284.

[9] Kyunghyun Cho, Bart van Merrienboer, Dzmitry ¨ Bahdanau, and Yoshua Bengio. 2014. on the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar, pages 103–111.

[10] King, M., Gharbieh, W., Park, S., & Cook, P. (2016). UNBNLP at SemEval-2016 Task 1: Semantic Textual Similarity: A Unified Framework for Semantic Processing and Evaluation. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 732-735).

[11] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the First workshop on Unsupervised Learning in NLP. Edinburgh, Scotland, pages 53–63.