Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220539

Wiki-match: A multi-model pipeline for image-caption matching task on Wikipedia dataset

Yibing Chen[1,4], Siyu Lei[2] and Zhouhang Sun[3]

[1]School of Computing and Information System, Singapore Management University, 188065, Singapore
[2]School of Computer Science, Sichuan University, 610207, China
[3]School of Data Science, The Chinese University of Hongkong (Shenzhen), 518172, China

[4]yibing.chen.2018@smu.edu.sg

Abstract. We propose a multi-model pipeline for image-caption matching tasks on Wikipedia-based dataset which leverages object-detection technique and attention mechanism to achieve fine-grained matching between textual representation and image representation. Different from the prior research, we not only evaluate our pipeline effectiveness on common benchmark dataset such as MS-COCO and Flickr30k, but also a new dataset that is derived from Wikipedia which is rich in natural entities and abstract concepts. Our findings show: 1) our model pipeline improves R@1 by 113.4%, R@3 by 86.1%, and R@5 by 114.4% compared to the original pipeline provided by the Wikipedia-based dataset. 2) our model pipeline has close to the state-of-the-art performance in common benchmark dataset including Flickr30k and MS-COCO. 3) images that are from Wikipedia creates bigger challenges for models to understand compares to MS-COCO or Flickr30k due to the abstract concepts and broad topics covered by Wikipedia.

Keywords: deep learning, natural language processing, computer vision, attention

1. Introduction

Vision and language are the two most important ways for humans to perceive the world. Human's eyes are constantly capturing visual information, and the human's brain is capable of transforming it into information that can be delivered by oral activity. According to Mei et al. [1], one of the most common routines for the human's brain is the interaction between vision and language information since human beings rely on the ability to talk about what they see. This is crucial for human's daily life and at the same time inspiring the research of artificial intelligence. Researchers start to think, "is there a way for computers to learn the connection between vision and language?". Therefore, in 1966, Marvin Minsky asked an undergraduate student to "... spend the summer linking a camera to a computer and getting the computer to describe what it saw" [2]. Although such ambitious projects failed, this was the first time when human researchers started to ask a computer to understand both vision and language.

Nowadays, the task involves enabling computers to understand and reason both vision and language information which is often described as the vision-language understanding task. According to Li et al. [3], common vision-language understanding tasks such as visual reasoning, question answering, and captioning require a computer to understand the semantic behind the vision object (e.g. an image) and

DOI: 10.54254/2755-2721/2/20220539

associate the concepts embedded in the vision object to the appropriate natural language expression. Besides those common tasks, a new type of vision-language tasks called image-caption matching is becoming more and more popular these days and draw a significant amount of attention due to its helpfulness in the context of web search and image understanding. According to Ordonez et al. [4], captions are textual information that describes the most important content expressed by the image and thus a good caption will convey the same meaning to a human being as its corresponding image. Aligning each image with a descriptive caption will benefit a user who wants to do a web search for images that belong to a typical theme. According to Hodosh et al. [5], current search engines are poor with image search. Specifically, instead of searching for a certain type of image directly, the search engine tries to match the user query with the descriptive text of an image (e.g. alt-text, caption). As a result, image-caption matching is an important field to study since teaching a machine to match an image and its most related caption can benefit the web search and other fields where an image needs to be associated with some descriptive textual representations. Besides, finding the most useful caption for an image also facilitates a user who tries to understand an ambiguous image online. According to Wikipedia - image/caption matching [6], most of the images from Wikipedia do not have any descriptive text associated with them. Since textual representation has less ambiguity compared to an image, finding the most relevant captions for Wikipedia images without any contextual description can greatly improve the user's understanding of the image.

In this paper, a novel model pipeline called Wiki-match is proposed. It leverages the advanced vision understanding and textual understanding technique and efficiently matches the images from Wikipedia with the most relevant captions. We hope that the Wiki-match pipeline can help to retrieve descriptive textual information for Wikipedia images and benefits tasks such as the image understanding and searching that may potentially bring better experiences to editors and users of Wikipedia.

2. Related work

This section will go through the history involved in the design of image-caption matching model, and particularly the techniques that are used to represent visual information (image) and textual information (caption).

2.1. Image-caption matching

The focus of image-caption matching is measuring the semantic relevance between captions and images. During the early period of research, Convolutional Neural Network (CNN) was considered as a common approach for visual features extraction and could also be used in textual feature extraction. Ma et al. [7] proposed a Multimodal CNN, leveraging multi-CNN to do image-caption matching from word, phrase and sentence level, aggregating the similarity score of different level CNNs. Based on the structure of CNN, Wang et al. [8] proposed two network structures (embedding network and similarity network) that produce different output representations for learning the similarity between image and caption, which achieved high accuracies on phrase localization and bi-directional image-sentence retrieval tasks. As the attention mechanism showed great potential in the fine-grained feature extraction of images, it had also been implemented in image-caption matching tasks on a large scale [9-13] to obtain salient regional information from images and texts. Later, Lee et al. [12] established a common embedding space and projected the words and image regions to a common space using stack-across attention mechanism to predict the semantic relevance between the whole image and the caption. With the advancement of attention mechanisms (elaborated in the later section) , more and more research has been focusing on the combining of visual attention and textual attention. Therefore, a dual attention network for multimodal matching between text and images was proposed by Nam et al. [11]. Similar to previous approaches that drew attention to both image and text, a novel bi-directional spatial-semantic attention network was proposed by Huang et al. [13], utilizing both the word to regions relation and visual object to words relation for more effective matching. In order to extract more precise and in-depth features from the images and text, a fusion layer was proposed by Wang et al. [10] to perform feature extraction

DOI: 10.54254/2755-2721/2/20220539

and combination, embedding image and text according to the distance between their individual features and common features.

2.2. Image representation and text representation

From the perspective of image representation, Girshick et al. [14] proposed an object detection method known as Fast Region-based Convolutional Network. Inspired by the work of Girshick et al. [14], Ren et al. [15] introduced a Region Proposal Network (RPN) and further merged RPN and Fast R-CNN to design an object detection algorithm called Faster RCNN, which has achieved state-of-the-art object detection accuracy and made great contributions to the research of object detection.

For text representation, word embedding is commonly used since it captures the abstract semantic concept of words via numerical representation [16]. Recently, the advancement of word embedding is driven by a deep learning approach. ELMo, proposed by Peters et al. [17], trains a bidirectional LSTM model on a large corpus and later utilizes LSTM to obtain the word representation.

3. Dataset

The Wiki-match pipeline is evaluated on two common benchmark dataset MS-COCO and Flickr30k. To further demonstrate the effectiveness of Wiki-match, a dataset named Wiki is constructed based on the Wikipedia-based image text dataset (WIT), which contains more abstract concepts and natural entities given by its abundant images and text from Wikipedia [18].

According to Srinivasan et al. [18], WIT dataset contains 37.6 million text examples across 108 types of languages and 11.5 million images. As it is shown in Table 1, WIT is about 30 times larger than MS-COCO and 300 times larger than Flickr30k in terms of image availability and text availability. Three types of description text including reference, attribution, and alt-text are provided for each image. Their research determined that the reference description, which is also known as caption, is the text property that is most related to the image itself. Therefore, Wiki-match focuses on the matching task between image and its corresponding "reference description" in the dataset.

Since Wiki-match does not address the multilingual problem yet, we filter the dataset based on the language type and only focus on image-caption pairs with English captions. We sample evenly from the original dataset which has English as caption language. The final dataset has 7.5k image and caption pairs in total and is named as Wiki.

Table 1. Comparison of images, text, languages availability across dataset.

Dataset	Images	Text	Languages
Flickr30k	32k	148K	<8
MS-COCO	330K	1.5M	<4
WIT	11.5M	37.6M	108

Note. Text column shown in the table is equivalent to caption. For images availability, the WIT dataset is 34 times larger than MS-COCO and 359 times larger than Flickr30k. For text availability, it is 25 times larger than MS-COCO and 254 times larger than Flickr30k.

DOI: 10.54254/2755-2721/2/20220539

4. Methodology

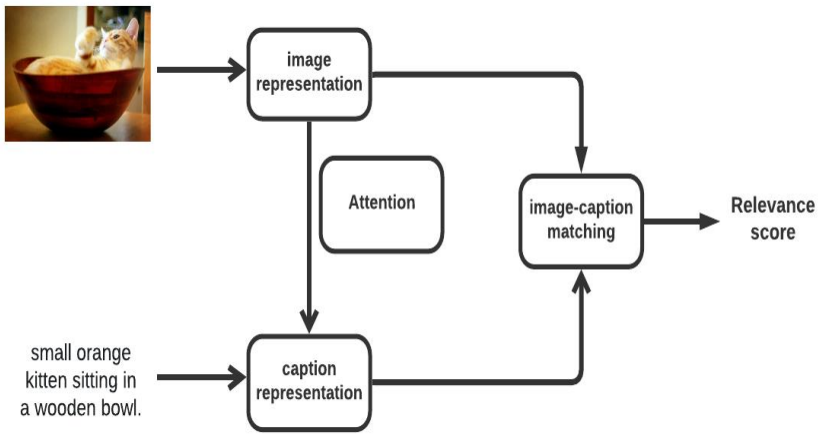

Figure 1. Model pipeline illustration for image-caption matching.

Note. Image representation (upper-left) and word (re-)representation (bottom left) component will project the images and words (captions) to the same feature space for relevance calculation.

The section 4 introduces the design details of Wiki-match, which consists of three components: image representations, caption representations, and image-caption matching as shown in Figure 1 and Figure 2. The image representation approach is introduced in section 4.1 and caption representation approach is in section 4.2. Finally, section 4.3 will present how Wiki-match utilizes the image and caption representation from 4.1 and 4.2 to teach the neural network to match images with the most descriptive caption.

The notations of this paper are presented as below for later discussions. A image is denoted as I, splitted into $|I|$ number of region $\{r_1, r_2 \dots, r_i\}$. A caption is denoted as C, splitted into $|C|$ number of word $\{w_1, w_2 \dots, w_j\}$. Each of the word has a set of hidden states $\{h_1, h_2 \dots, h_k\}$. Cosine similarity between 2 objects a,b is denoted as cos(a, b).

DOI: 10.54254/2755-2721/2/20220539

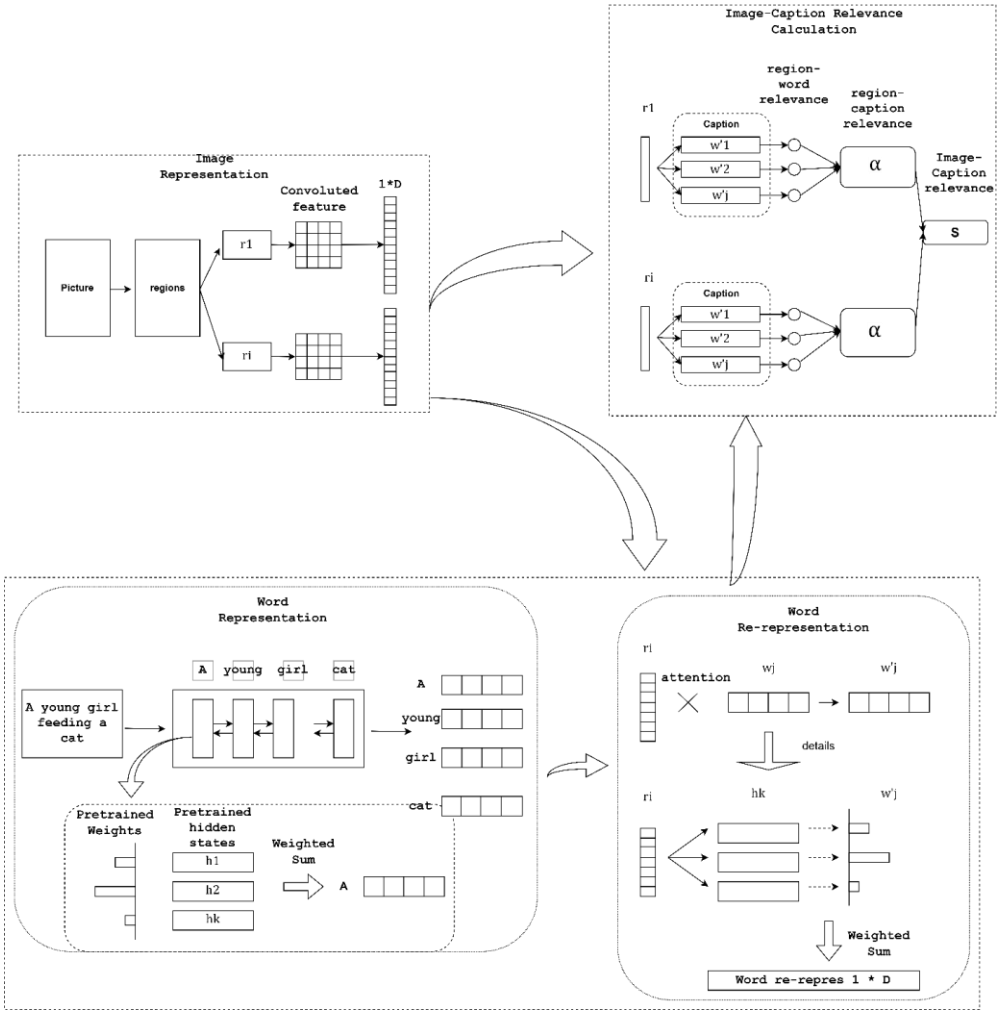

Figure 2. Model pipeline illustration for image-caption matching.

Note. Image representation (upper-right) and word (re-)representation component (upper-left) will project the images and words (captions) to the same feature space for similarity calculation.

4.1. Image representation and region proposal

In this work, an image is split into a collection of regions representing the key features of the image. The image is fed into ResNet101 to obtain the feature representations. The image representation was subsequently used for Faster-RCNN detecting objects as the region proposal. Finally, each image is represented as R regions and each region r is represented as 1*D dimensional vector. In the experiment setup, D is 2048.

4.2. Caption representations

4.2.1. Word representation. A caption consists of a set of words and each word is represented as a vector contributing a certain degree of the words' information. Wiki-match utilizes a vocabulary lookup table which contains all the words appearing in every caption. The vocabulary table contains N rows representing N number of words, each of the rows is one vector as the word representation. Besides the vocabulary table, Wiki-match has a dictionary containing words as keys and one-hot vectors as corresponding values. Each of the one-hot vectors indicates the index of the word representations in the lookup table. The dictionary size is N, the same as the vocabulary lookup table row numbers.

DOI: 10.54254/2755-2721/2/20220539

Each word's representation is pre-trained by ELMo. Different from previous word embedding methods using the last hidden state representing the word's meaning, ELMo made use of the weighted sum of each hidden layer to represent the word. The intuition is that the first one or two hidden layers could capture the low level syntactic, such as part-of-speech tagging, syntactic dependencies and name-entity recognition. The last one or two hidden layers are better for higher-level semantics representations, such as sentiment, semantic role labelling, and question answering.

4.2.2. Word re-representation. Implementing on top of EMLo, Wiki-match utilized the attention mechanism with region representation as a query and each hidden layer of one word as the set of values to re-represent the word. The intuition is that the region should decide how important is low level syntactic and high level semantics of a word contributing to the word representation when mapping to this specific region.

The attention score e_{ijk} for each region r_i and hidden state $h_{jk}$is calculated as a dot product.

$$e_{i\,jk} = r_i \cdot h_{jk} \tag{1}$$

Then normalize e_{ijk} to get attention distribution α_{ijk}

$$\alpha_{ijk} = softmax(e_{itk}) = \frac{exp(e_{ijk})}{\sum_m exp(e_{ijm})} \tag{2}$$

Finally the word represent w'_{ij} is obtained

$$w'_{ij} = \sum_k \alpha_{ijk} \cdot h_{jk} \tag{3}$$

The re-represented words contain information of the word meaning with respect to the caption context as well as how each of the image regions perceives the words, which is important for the following image-caption relevance calculation in the 4.3 section.

4.3. Image caption relevance

In this work, cosine similarity is used to calculate the relevance of region r_i with regard to each word w'_{ij}, denoted as cos(r_i , w'_{ij},). By taking the average sum of each cosine similarity region r_i and word w'_{ij} pairs, the region-caption (r_i- C) relevance is determined, denoted as α_{r_ic} in equation (4).

$$\alpha_{r_ic} = \frac{\sum_j \cos(r_i, w'_{ij})}{|C|} \tag{4}$$

Analogously, image-caption relevance (I- C) could be determined as the average sum across each region r_i and caption C as S.

$$S = \frac{\sum_i \alpha_{r_ic}}{|I|} \tag{5}$$

After obtaining the image caption relevance score, triplet loss is used to assess the quality of the current relevance score since it is commonly used in the prior research for image-caption matching [3, 9, 10, 12]. Specifically, for all the image-caption samples in the mini-batch, each image I and its corresponding caption C is labeled as a positive pair ($I - C$) and each combination of the image I with the caption C' that does not associate with I in the mini-match is labeled as negative pairs ($I - C'$). Similarly, for each caption C, all the unrelated image I' to C will also yield a negative pair ($I' - C$) . The intuition is that the neural network encourages all the positive pairs ($I - C$) to produce a higher image-caption relevance score and all the negative pair ($I - C'$) or ($I' - C$) to produce a small image-caption relevance score. If a positive pair produces a small relevance score or a negative pair produces a high

DOI: 10.54254/2755-2721/2/20220539

relevance score, such pair will be considered as a mistake (loss) and the neural network will refine the calculation by gradient descent accordingly. In the design of Wiki-match, for each image caption pair $(I - C)$ in a mini-match, the model pipeline learns from the hardest negative pair. A negative image - caption pair is defined by $(I - C')$ or $(I' - C)$ and the hardest negative pair is a negative pair which the current neural network falsely produces a high relevance score. Therefore, to calculate the loss for each $(I - C)$ in the mini-batch, the model need to find the hardest image I_h for C by equation (6) and the hardest caption C_h for I by equation (7) to obtain the hardest negative pair for $(I - C)$.

$$I_h = argmax_{I' \neq I} S(I', C) \tag{6}$$

I_h is the the image in the batch that leads to the hardest negative pair for I, denoted as $(I_h - C)$

$$C_h = argmax_{C' \neq C} S(I, C') \tag{7}$$

C_h is the caption in the batch that leads to the hardest negative pair for C, denoted as $(I - C_h)$ Finally, the loss of image-caption pair $(I - C)$ is given by:

$$loss(I, C) = [\alpha - S(I, C) + S(I, C_h)]_+ + [\alpha - S(I, C) + S(I_h, C)]_+ \tag{8}$$

α in equation (8) denotes the margin and $[\cdot]_+$ denotes the hinge loss.

5. Experiment

Wiki-match is tested on the Wiki dataset that is derived from the WIT dataset and also on the MS-COCO and Flickr30K datasets. To demonstrate the outstanding performance of Wiki-match, its performance is also compared to the image-text retrieval pipeline in the WIT [18] and other state-of-the-art pipelines.

The recall at top K (R@K) is used to assess the performance of Wiki-match, which is commonly used in image-captioning and information retrieval. R@K is defined as the percentage of correct items in the top K retrieved results. In the experiment, we compare the R@1, R@3, R@5 and R@10 values with the models proposed by the prior researcher.

6. Result analysis

The effectiveness of Wiki-match is measured in two ways. One is to compare the Wiki-match with previous research results using the same data set, and the other is to compare different data sets under the same model pipeline.

6.1. Comparison with prior image-caption pipelines

The performance of Wiki-match is compared to the baseline method provided by the WIT dataset on the Wiki dataset, which is proposed by Srinivasan et al. [18].

Table 2. Comparison of model pipelines effectiveness on Wiki dataset.

Methods	R @ 1	R@3	R@5
Wiki-match	**0.143**	**0.227**	**0.326**
WIT baseline	0.067	0.122	0.152

Note. Srinivasan et al. [18] propose WIT baseline to address image-caption matching task and Wiki-match outperform WIT baseline by increasing R@1 by 113.4%, R@3 by 86.1%, and R@5 by 114.4%

Table 2 shows that Wiki-match has a higher R@K value than the WIT baseline, which demonstrates the effectiveness of the Wiki-match. The reason could be the better representation of captions and fine-grained matching brought by the attention mechanism in the pipeline design of Wiki-match. Since the design of Wiki-match involves re-representing the captions according to regions in the image via attention mechanism, words in captions and images interact with each other and enable a better representation of captions. The better representation of captions is later used to find region-word relevance, region-caption relevance, and finally the image-caption relevance. Therefore, Wiki-match enables a multi-level granularity for relevance calculation and thus leads to superior performance.

DOI: 10.54254/2755-2721/2/20220539

Besides, Wiki-match also yields close to state-of-the-art performance on common benchmark datasets Flickr30k and MS-COCO.

Table 3. Comparison between Wiki-match and state-of-the-art pipelines on MS-COCO.

Methods	Image-to-Text retrieval		
	R @ 1	R @ 5	R @ 10
VSE++ [19]	0.646	-	0.957
SCAN [12]	0.727	0.948	0.984
PFAN t-i+ i-t [20]	0.765	0.968	0.99
Wiki-match	**0.681**	**0.847**	**0.884**

Note. On benchmark dataset MS-COCO, Wiki-match performance is close to the state-of-the-art pipeline PFAN proposed by Wang et al. [8] in terms of R@1, R@5, and R@10.

As shown in Table 3, Wiki-match's performance on benchmark dataset MS-COCO is close to the state-of-the-art pipeline such as PFAN [20], SCAN [12] and VSE+ [19]in terms of R@1, R@5, and R@10. It shows that the image-caption matching enabled by Wiki-match pipeline is very effective. Similarly, Table 4 suggests that Wiki-match performance on benchmark dataset Flickr30k is also close to the state-of-the-art pipeline proposed by prior researchers, which further demonstrates the effectiveness of the design of Wiki-match pipeline.

Table 4. Comparison between Wiki-match and state-of-the-art pipelines on Flickr30k

Methods	Image-to-Text retrieval		
	R @ 1	R @ 5	R @ 10
VSE++ [19]	0.529	-	0.872
SCAN [12]	0.674	0.903	0.958
PFAN t-i+ i-t [20]	0.700	0.918	0.950
Wiki-match	**0.619**	**0.817**	**0.864**

Note. On benchmark dataset Flickr30k, Wiki-match performance is close to the state-of-the-art pipeline PFAN proposed by Wang et al. [8] in terms of R@1, R@5, and R@10.

6.2. Comparing Wiki-match performance across datasets

Table 5. Wiki-match performance across datasets.

Dataset	Image-to-Text			
	R @ 1	R@3	R @ 5	R @ 10
Flickr30k	0.619	-	0.817	0.864
MS-COCO	0.681	-	0.847	0.884
Wiki	0.143	0.227	0.326	0.482

Note. The performance of Wiki-match in terms of R@1, R@5, and R@10 degrades heavily on the Wiki dataset, which is retrieved from the WIT dataset.

DOI: 10.54254/2755-2721/2/20220539

The experiment shows that the performance of Wiki-match is generally good on benchmark datasets such as MS-COCO and Flickr30k but degrades heavily on the Wiki dataset according to Table 5. It is because the image-caption pairs from the Wiki dataset contain concepts that are more abstract and challenging to learn and interpret.

Figure 3. An abstract image-caption pair in Wiki dataset [18].

Note. The correct caption for the image above is "Erskine River [SEP] Erskine River at Lorne". "[SEP]" is a separator used by pretrained models such as BERT. The image clearly shows a river but geographic information such as Erskine and Lorne is not shown.

As shown in Figure 3, an image could have a caption that contains information such as "Erskine river". It is possible for object detection to find regions associated with the river from the image and compute relevance score according to the "river" keyword in the caption. Thus, all captions that have the keyword "river" could yield high relevance scores. However, to determine the real caption, the object detection needs to consider the association between regions in the image and the "Erskine" keyword. It is very hard because "Erskine" is just an abstract name given by the caption. Similarly, Figure 4 shows another challenging case where specific geographic information such as "Wisconsin" and "Deer park" and the abstract concept like "downtown" are expected to be interpreted from the image. Objects involved in the picture are a road, some cars, and houses, but they are not enough to determine the correct caption since many captions could contain similar keywords. Therefore, when a caption contains abstract and detailed information about an object, retrieving the right caption based on an image is very hard.

Figure 4. An abstract image-caption pair in Wiki dataset [18].

Note. The correct caption for the image above is "Deer Park, Wisconsin [SEP] Downtown Deer Park". "[SEP]" is a separator used by pretrained models such as BERT. The image clearly shows a road, houses, and cars but abstract information such as "Downtown Deer Park" is not shown.

DOI: 10.54254/2755-2721/2/20220539

7. Conclusion

In this work, Wiki-match is introduced to conduct an image-text matching task on the Wiki dataset. The experiment reveals that adding the attention mechanism to word re-representation enables the words to obtain more fine-grained representation from images' regions, and thus making the image-caption matching more effective. Therefore, the Wiki-match pipeline is much better than the original image-text retrieval pipeline in the WIT paper (increase R@1 by 113.4%, R@3 by 86.1%, and R@5 by 114.4%) proposed by Srinivasan et al. [18]. Comparing the performance of Wiki-match on the Wiki dataset to the benchmark datasets, it is obvious that images from the Wikipedia raises great challenges for Wiki-match to retrieve correct captions. Object detection can only yield limited information from the image and is not enough to match with the abstract concepts expressed by the caption. To conclude, the images and captions from Wikipedia that involve abstract concepts raise greater challenges for machine learning models to match the semantic meaning behind images towards the corresponding captions.

In the future, more work needs to be done to address the retrieval of captions containing abstract concepts. We will also adapt our model to multilingual datasets. We hope we can also leverage more in-depth attention mechanisms in the context of vision and text to achieve better image-caption matching results.

References

[1] Mei, T., Zhang, W., & Yao, T. (2020). Vision and language: from visual perception to content creation. APSIPA Transactions on Signal and Information Processing, 9:1-8.

[2] Szeliski, R. (2010). A brief history. In: Gries D. Schneider F.B. (Eds.), Computer vision: algorithms and applications. Springer Science & Business Media. New York; London. pp. 11.

[3] Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

[4] Ordonez, V., Kulkarni, G., & Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems, 24: 1143-1151.

[5] Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47: 853-899.

[6] Wikimedia Foundation. (2021). Wikipedia - image/caption matching. https://www.kaggle.com/c/wikipedia-image-caption.

[7] Ma, L., Lu, Z., Shang, L., & Li, H. (2015). Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE international conference on computer vision. pp. 2623-2631.

[8] Wang, L., Li, Y., Huang, J., & Lazebnik, S. (2018). Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41: 394-407.

[9] Xu, X., Wang, T., Yang, Y., Zuo, L., Shen, F., & Shen, H. T. (2020). Cross-modal attention with semantic consistence for image–text matching. IEEE transactions on neural networks and learning systems, 31: 5412-5425.

[10] Wang, D., Wang, L., Song, S., Huang, G., Guo, Y., Cheng, S., ... & Du, A. (2021). Fusion layer attention for image-text matching. Neurocomputing, 442: 249-259.

[11] Nam, H., Ha, J. W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 299-307.

[12] Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 201-216.

[13] Huang, F., Zhang, X., Zhao, Z., & Li, Z. (2018). Bi-directional spatial-semantic attention networks for image-text matching. IEEE Transactions on Image Processing, 28: 2008-2020.

[14] Girshick, R. (2015). Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 1440-1448.

DOI: 10.54254/2755-2721/2/20220539

[15] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28: 91-99.
[16] Wang, S., Zhou, W., & Jiang, C. (2020). A survey of word embeddings based on deep learning. Computing, 102: 717-740.
[17] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv 2018. arXiv preprint arXiv:1802.05365, 12.
[18] Srinivasan, K., Raman, K., Chen, J., Bendersky, M., & Najork, M. (2021). WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. arXiv preprint arXiv:2103.01913.
[19] Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612.
[20] Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., & Fan, X. (2019). Position focused attention network for image-text matching. arXiv preprint arXiv:1907.09748.