1st Place Solution for the Kaggle Image Caption/Matching Competition

Yiming Chen
New Oriental AI Research
chenyiming71@xdf.cn

Penglong Luan
New Oriental AI Research
luanpenglong@xdf.cn

He Zhao
New Oriental AI Research
zhaohe16@xdf.cn

Shaou-Gang Miaou
Chung-Yuan Christian Univ.
miaou@cycu.edu.tw

Abstract

This paper presents our solution to the image caption/matching competition hosted by Kaggle to automatically retrieve the text for online images in Wikipedia. We used VinVL (Vision (V) in Vision Language (VL)) with region feature extraction and a Vision-Language fusion model to perform the image-to-text retrieval task and achieve the 1st place with 0.73413 in terms of NDCG (Normalized Discounted Cumulative Gain Score) on the leaderboard.

1 Introduction

We often rely on online images for knowledge sharing, learning and understanding. In some Wikipedia articles, some images either have no corresponding context or contain too many irrelevant descriptions. To solve this problem, the early practice of the Internet is manual correction, which can be done in a limited time when the amount of information is small. However, with the explosive growth of information on the Internet, human based text-image correction is time-consuming and tedious. Thus, logic based algorithms are utilized to automatically perform these tasks. For instance, current solutions of Wikipedia rely on translations and page interlinks based methods.

In recent years, with the development of multimodal algorithm and the invention of Bert model, the accuracy of image caption task has been greatly improved.Therefore, the image caption based method is gradually used to replace the logical method above. However, the feature distribution of massive image and text information has become increasingly complex. Even the most advanced computer vision based image captioning is not suitable for the images with complex semantics[1].

To improve the current situation above, the image caption/matching competition was hosted by Kaggle. In the competition, a feasible model is required to automatically retrieve the text closest to an image.

In this competition, we build and fine-tune the model based on VinVL (Zhang et al., 2021) and the

[1] https://www.kaggle.com/c/wikipedia-image-caption/overview/description

original data provided by the organizer.

2 Algorithm

The main structure of our proposed method is shown in Figure 1, where it consists of two modules, including region feature extraction and a Vision-Language fusion model. The region feature extraction is the key contribution of Zhang et al. (2021), where their proposed fusion model is an expansion of OSCAR (Li et al., 2020) called OSCAR+ (Zhang et al., 2021). The details of the proposed method will be given next in the following sections.

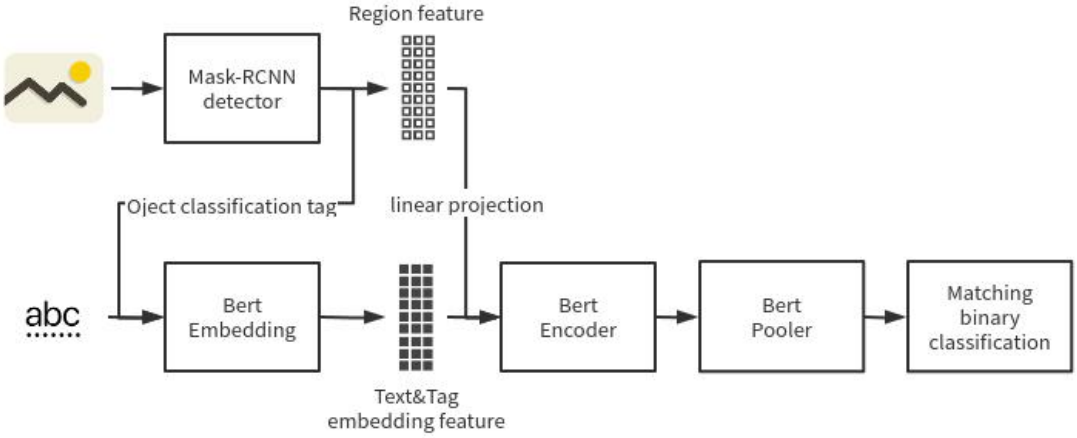

Figure 1 Overview of main framework

2.1 Region feature extraction

Usually, vision-and-language tasks utilize the features from the whole picture (Shuster et al., 2019) or separated picture patches (Zhang et al., 2021).However, on the MS COCO (Lin et al., 2014) dataset, the percentages that an image and its paired text share at least 1, 2, or 3 objects are 49.7%, 22.2%, and 12.9%, respectively (Zhang et al., 2021). In other words, image description has a relatively strong relationship with the salient objects in the image. And the feature vector of salient objects inside the image instead of, traditionally, overall image's feature is considered as input into the image caption network design. To implement the idea above, the most intuitive way is to utilize a pre-trained object detector to collect features in each individual region.

In the proposed algorithm, Mask-RCNN (He et al., 2017) with additional attribute branch is utilized as an object detector, which is pre-trained through COCO (Lin et al., 2014), Objects365 (Shao et al., 2019), OpenImagesV5 (Kuznetsova et al., 2020) and Visual Genome (Krishna et al., 2016) datasets that have a total of 1848 object classes and 524 attribute classes. And for feature extraction purpose, the Mask-RCNN detector eliminates the mask branch and attribute branch which existed originally in Mask-RCNN

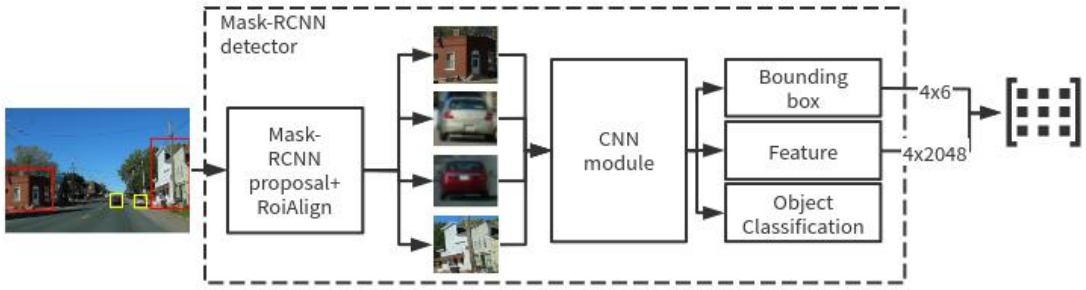

Figure 2 Structure of region feature extraction

In Figure 2, the Mask-RCNN detector selects proposal regions in terms of proposal and ROI align module (He et al., 2017). Then proposal regions are sent directly to the CNN module that can generate 1×2048 feature vector for each region input. In addition, normalized coordinate features with the shape of 1×6 from each proposal region is considered, which are top-left point, bottom-right point, width and height of each region bounding box. As shown in Figure 2, we have 4 proposal regions and 4×2054 embedding features. Besides, object classification results (tags) would be fed into Bert Embedding module for tag embedding feature generation, as presented in Figure 1.

2.2 Vision-Language fusion model

Vision-and-Language Pre-training (VLP) which learns joint image-text representations on large-scale image-text pairs has been shown to improve the performance on various vision-language downstream tasks, such as visual question answering, image-text retrieval, and image captioning. Most existing VLP models are based on multi-layer transformers and simply concatenate image region features and text features as input. In this case, the semantic alignments between image regions and text are roughly learned by the self-attention mechanism.

Inspired by this, OSCAR (Li et al., 2020) innovatively uses object tags detected from images as anchor points to significantly assist the learning of semantic alignments. Further, OSCAR+ (Zhang et al., 2021) improves the performance of VLP with a larger pre-training corpus and more complex pre-training objectives compared with OSCAR and outperforms existing approaches by a significant margin.

The input of our fusion model is an image-text pair represented as a triple (sentence & object tags & region features), of which the sentence is a meaningful text sequence, the object tags and region features are extracted from the image via the region feature extraction module described in the previous section.

The sequence and the object tags are tokenized and fed into Bert Embedding to get the embedding features. To concatenate the region features with word embedding, the region feature vector is transformed using a linear projection to ensure that it has the same vector dimension as that of word

embedding.

The Bert Encoder consisted of Multi-Layer Transforms is applied to the embedding of input. The representation of [CLS] in encoder outputs is sent to the Bert Pooler to get the fused vision-language representation.

In the pre-training phase, a novel pre-training objective is designed in which Masked Token Loss and Contrastive Loss are used to learn the cross-modal representations effectively (Li et al., 2020). The Masked Token Loss is similar to masked language model used by Bert (Devlin et al., 2019). The Contrastive Loss (Zhang et al., 2021) takes into account two types of training samples *x*: the {caption, image-tags, image-features} triplets of the image captioning data denoted by $[c, t, f]$, and the {question, answer, image-features} triplets of the VQA data marked as $[q, a, f]$(as shown in Eq. (1) of which *y* is the label of *x*). A 3-way classifier is used to predict whether the triplet is matched ($y = 0$), contains a polluted caption c' ($y = 1$), or contains a polluted answer a' ($y = 2$). Polluted caption c' is uniform probability sampled from all c and q. Similarly, polluted answer a' is obtained from t and a.

$$y = \begin{cases} 0, & x = [c, t, f] \text{ or } x = [q, a, f] \\ 1, & x = [c', t, f] \\ 2, & x = [q, a', f] \end{cases} \quad (1)$$

OSCAR+ build a larger corpus with 5.65 million unique images and 8.85 million text-tag-image triples on three types of existing datasets, including image captioning datasets, visual QA datasets and image tagging datasets with machine-generated. Two pre-trained models denoted as $\text{OSCAR}+_B$ and $\text{OSCAR}+_L$ with structure of Bert base and large trained on this corpus are released by Zhang et al. (2021).

In this image-to-text retrieval task, the representation of [CLS] in model outputs is used as the input to a classifier to predict a score indicating the similarity of the given image and sentence. In testing, the predicted score is used to rank a given image-text pairs of a query. We fine-tune the classifier model on the pre-trained $\text{OSCAR}+_B$ (Zhang et al., 2021) with the competition dataset, and report the top-5 retrieval results on the test data. The AdamW Optimizer (Devlin et al., 2019) is used with learning rate 0.00002. The model is fine-tuned for 30 epochs with batch size 16.

REFERENCES

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Kaiming He, Georgia Gkioxari, Piotr Doll´ar, and Ross B. Girshick. Mask r-cnn. 2017 IEEE *International Conference on Computer Vision* (ICCV), pp. 2980–2988, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-

Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, andVittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128:1956–1981, 2020.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, LijuanWang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pretraining for vision-language tasks. In *ECCV*, 2020.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8429–8438, 2019.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12508–12518, 2019.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 5579–5588, June 2021.