DL-TXST NewsImages: Contextual Feature Enrichment for Image-Text Re-matching

Yuxiao Zhou, Andres Gonzalez, Parisa Tabassum, Jelena Tešić
Computer Science Department, Texas State University, San Marcos, TX, U.S.A.
{y_z37, ag1548, lpd18, jtesic}@txstate.edu

ABSTRACT

In this paper, we describe our multi-view approach to news image re-matching to text for the news articles run submission. The feature pool consists of provided features, baseline text, and image features using pre-trained and domain-adapted modeling and contextual features for the news and image article. We have evaluated multiple modeling approaches for the features and employed a deep multi-level encoding network to predict a probability-like matching score of images for a news article. Our best results are the ensemble of proposed models, and we found the URL for the image and related images provides the most discriminative context in this pairing task.

1 INTRODUCTION

Online news articles are multimodal; the textual content of an article is often accompanied by an image. The image is important for illustrating the content of the text, but also attracting readers' attention. Existing research generally assumes a simple relationship between images and text e.g. image captioning is often assumed to a brief textual description of an image. In contrast, when images accompany news articles, the relationship becomes less clear. In this research, we employ a state-of-the-art method to build models to describe the connection between the textual content of articles and the images that accompany them. We evaluate our proposed model on the benchmark dataset derived from four months of webserver log files of a German news publisher. The performance of the proposed model is measured by image matching precision such as MRR and Mean Recall at different depths.

2 RELATED WORK

Recent work utilizes deep neural networks to capture the visual-semantic similarity between image and text. Wang et al. [2] and Faghri et al. [3] map the image and the full sentence into a common vector space and compute the similarity between the global representations. Fine-tuned version of the approach uses a range of embedded information in news and images e.g. extracted named entity and image features [6] or news image caption with named entities [7]. Semantic concept learning [4] and region relationship reasoning [5] approaches showed to improve the discriminative ability of the unified embeddings.

MediaEval'21, December 13-15 2021, Online

3 APPROACHES

3.1 *Image-Text Matching via Categorization* The Image-Text Matching via Categorization portion of the primary model involved finding a correlation between an image output and a text-processing output. Previous work emphasized matching an image to the category via the URL by which the image was downloaded. The image-processing model involved retraining a portion of ResNet50, and the model could be fed any image and predict in which category the model thought the image fit best—the final output being a 70-dimension array. This final output could then be correlated via Cosine Similarity to the similar output of the text-classification model, and the text-classification model was also trained on categories provided by the article's URL. For our text-processing model, we combined the title and text of the article into one input. The text was standardized by lowercasing all text, and all punctuation was removed. The text sequence length was also trimmed to 250 tokens. After each data item had its output predicted, every title/text prediction was cosine-similarity compared to the image-prediction output to arrive at a final value that would describe the entire model's predictions as to which article each image best matched. These scores were then combined with the rest of the whole model with varying weights to find better combinations of individual-component-model weights.

3.2 *Face-Name Matching* In many image-article pairs, the publisher uses a portrait of person who is mentioned in the article as the news image. Therefore, we may re-match images and texts by matching the names within the texts and the faces within the images. The Stanford Named Entity Recognizer (NER) [8] was employed for the person's name recognition. By performing the NER on the 7530 given news headlines, we find that at least 24% of them include the person's name. We used open-source face detection frameworks, deepface [9] and Google FaceNet, to detect and represent the face as a 128-dims vector. During testing, we encoded the face from the image and aggregate the number of matched faces connected to the person mentioned in the news headline. The image and the person matched if the cosine-distance between two vectors was less than 0.4. The matching score of images were calculated by multiplying the similarity score with the total matched.

3.3 *Image-Text Fusion with Image Captioning* Based on the hypothesis that the description of a new image is semantically like the matched news title, we first adopted an image captioning model [10] pre-trained with COCO dataset for image caption generation, and then calculate the similarity score between the generated image captions and the given news headlines. The pre-trained image captioning model has three main components: 1. *Image Feature Extractor*: the image caption model uses

ResNet101, a convolutional neural network (CNN) that is 101 layers deep for feature extraction; 2. *Transformer encoder*: the extracted image features are then passed to a Transformer-based encoder that generates a new representation of the inputs; 3. *Transformer Decoder*: this component takes the encoder output and the text data sequence as inputs and tries to learn to generate the caption; 4. *Text Similarity :* we employ Word Mover's Distance (WMD) to compare the similarity between image captions and articles titles. The WMD algorithm uses normalized Bag-of-Words and word embeddings to calculate the distance between documents and sentences. The wmdSimilarity is simply the negative wmd between the image caption and the article title.

3.4 *Impact of using Image URLs* The image URL and text URL pairs demonstrate some explicit relationships between them. More specifically, an image and a text may be matched if their URLs contain one or more common tokens. The pipeline in Figure 1 has been proposed and implemented.

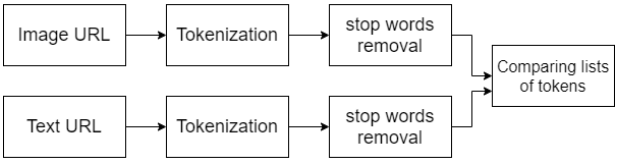

Figure 1: Usage of Image URL

A pair of URLs are matched if they have at least one common token. The ranking of images is sorted by the number of total matched. It is noted that this part of work is only for the research interest, but not part of the submission.

3.5 *Metric Learning* We used metric learning to match images with their corresponding articles. Image features were extracted using EfficientNetB0 model, where we took out the fully connected output layer and added a pooling layer, a dropout layer, and a dense layer to map high dimensional image features into 70-dimensional space. Text features learning pipeline has six layers: vectorization, embedding, dropout, pooling, dropout, and a dense layer, and produces a 70-dimensional text feature vector. Next, we use a triplet loss model for metric learning. The triplets consist of the anchor sample (image feature), positive sample (text feature that corresponds to the anchor image), and negative sample (text feature that is dissimilar to the anchor image). The triplet loss model learns similar representations for the samples we defined as positive and distant representations for samples we defined as negative. We used to learn model weights to compute the feature vectors of all the test set images and text. Trained model did not output any representation for the queries. Due to lack of experience using the triplet loss model, feature vectors, and time constraints, we were unable to make the triplet loss model work. So, we decided to discard the model.

4 RESULTS AND ANALYSIS

Data The MediaEval 2021 Image-Text Re-Matching benchmark provides four batches of data which consist of the headline and a text snippet of German news articles and their accompanying images. The first three batches are used for training, and the last one is used for testing. We split the training dataset into the actual training set and validation set. The training set included 5135 records, while the validation set included 2384 records.

Result Our proposed approach uses an ensemble design, so our submissions are combined results from three or four models. According to the evaluation results (as shown in Table 1) on the both training set and test set, the performance of categorization and captioning are similar, and results from Face-Name matching can be used to optimize the ranking of the acquired image list.

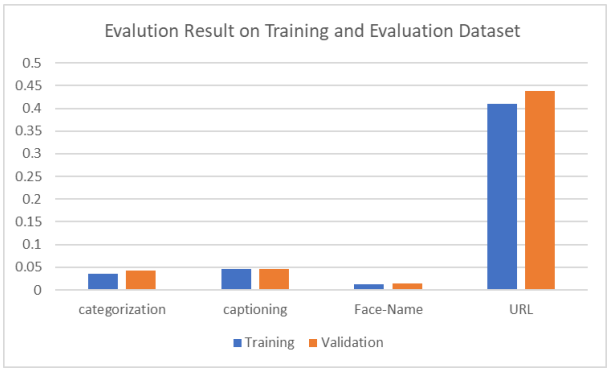

Figure 2: Evaluation Result on Train and Validation Dataset

Run1 combines three different methods. Equal weights are assigned to the categorization-based method, and a combination of face-name matching and image captioning-based methods. The ranking of a candidate image in Run1 is:

$R_{\text{Run1}} = 0.5R_{\text{Categorization}} + 0.5\,(R_{\text{Face}} + R_{\text{caption}})$

Run2 combines all proposed methods. The first three models are ensembled using the same approach as in Run1. This ensembled model is used for creating the initial top 100 image list. Then we append the result, which is generated from the URL matching-based method, to the end of the top 100 image list.

Run3 is like Run2. The only difference is that we append the result of last method to the head of the top 100 image list.

Since the image URL is an artificial feature, the results from Run1 and Run3 are not included in the result.

Table 1: **Results from different Runs on Test Dataset**

Run	MRR@100	MR@5	MR@5	MR@5	MR@5
1	0,00668	0,00836	0,01097	0,02977	0,05274
2	0,01147	0,00836	0,01097	0,03029	0,49347
3	0,28788	0,3718	0,4094	0,46684	0,49347

5 CONCLUSIONS

In this work, we explored the relationship between the textual and visual (images) content of news articles and built a few deep learning-based models which can calculate ranking of candidate images for a given news article. From our experiments, it was clear that an image-text relation-based model can be used for news image re-matching prediction, but it performed poorly, while the usage of text features and the image URL showed an improved performance. In the future, we plan to incorporate metric learning in our model. We also plan to conduct the image-text matching experiment with improved features like news image captions with embedded named entities or metadata. Furthermore, the ensemble may be extended by the application of techniques such as bagging, boosting, and stacking.

REFERENCES

[1] H.Diao, Y.Zhang, L.Ma and H.Lu. *Similarity Reasoning and Filtration for Image-Text Matching*, 36–44. In the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), 2021

[2] Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5005-5013. 2016.

[3] Faghri, Fartash, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. "Vse++: Improving visual-semantic embeddings with hard negatives." arXiv preprint arXiv:1707.05612 (2017).

[4] Huang, Yan, Qi Wu, Chunfeng Song, and Liang Wang. "Learning semantic concepts and order for image and sentence matching." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6163-6171. 2018.

[5] Li, Kunpeng, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. "Visual semantic reasoning for image-text matching." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4654-4662. 2019.

[6] Nguyen-Quang, T., Nguyen, T. D. H., Nguyen-Ho, T. L., Duong, A. K., Hoang-Xuan, N., Nguyen-Truong, V. T., ... & Tran, M. T. (2020). HCMUS at MediaEval 2020: Image-Text Fusion for Automatic News-Images Re-Matching.

[7] Z. Yumeng, Y. Jing, G. Shuo and L. Limin, "News Image-Text Matching With News Knowledge Graph," in IEEE Access, vol. 9, pp. 108017-108027, 2021, doi: 10.1109/ACCESS.2021.3093650.

[8] 2021, Stanford Named Entity Recognizer (NER). https://nlp.stanford.edu/software/CRF-NER.html

[9] 2021, DeepFace – The Most Popular Open Source Facial Recognition Library. https://viso.ai/computer-vision/deepface/

[10] 2021, An image caption code base. https://github.com/ruotianluo/ImageCaptioning.pytorch#generate-image-captions

[11] 2021, A guide to transfer learning with Keras using ResNet50 https://medium.com/@kenneth.ca95/a-guide-to-transfer-learning-with-keras-using-resnet50-a81a4a28084b