

DL-TXST NewsImages: Contextual Feature Enrichment for Image-Text Rematching

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

ABSTRACT

In this paper, we describe our multiview approach to news image rematching to text for the news article 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 multilevel 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 that 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 to illustrate the content of the text and also to attract readers' attention. Existing research generally assumes a simple relationship between images and text, e.g. image captioning is often assumed to be a brief textual description of an image. On the contrary, when images accompany news articles, the relationship becomes less clear. In this research, we employ a state-of-the-art method that builds models to describe the connection between the textual content of articles and the images that accompany them. We evaluated our proposed model on the benchmark data set derived from four months from web server 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 entire sentence to a common vector space and compute the similarity between the global representations. The fine-tuned version of the approach uses a range of embedded information in news and images, for example, extracted named entities and image features [6] or the caption of the news image with named entities [7]. Semantic concept learning [4] and regional relationship reasoning [5] approaches were shown to improve the discriminative ability of unified embeddings.

3 MATCHERS

In this section we introduce the matchers we have tested for image-text rematching task, as illustrated in

3.1 *The Semantic Space Matcher* matches text and image embeddings in the semantic space using the cosine distance. Previous work emphasized matching an image to the category using the URL by which the image was downloaded [6]. We streamline the approach and fix the semantic space ahead. We refine the classification layers of ResNet50 to produce probability outputs for a fixed semantic space for 70 classes, creating a 70-dimensional array. We combine the title and the text of the article, normalize it, and feed it to a text classifier that produces a probability that the text will describe one of the 70 semantic classes. The probability of the category is a feature vector value for that category. The result is two sets: one containing the text feature vectors, one per text instance, and another one containing the image feature vector, one per instance, in the same feature space. Next, we match an image to the input text based on the minimal cosine distance between the said text feature vector and all image features.

3.2 *The Face-Name Matcher* correlates the names within the the articles with the faces within images using 128-dimensional image space embedding [7]. The Stanford Named Entity Recognizer (NER) [8] provided a named entity recognizer particularly for the extraction of person names: of the 7530 given news articles in the corpora, 24% of them included the person's name. We use open source face detection FaceNet to connect the person's name from the article to publicly available images and to create a 128-dimensional face vector [11]. We use Google DeepFace to detect the faces in the images, and encodes the detected faces to the same space [9]. The image is matched to the article based on a minimum cosine distance between the vectors for 24 % of the articles that contain the actual names. For articles that do not contain person names, image captioning is utilized.

3.3 *The Image Captioning Matcher* 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;

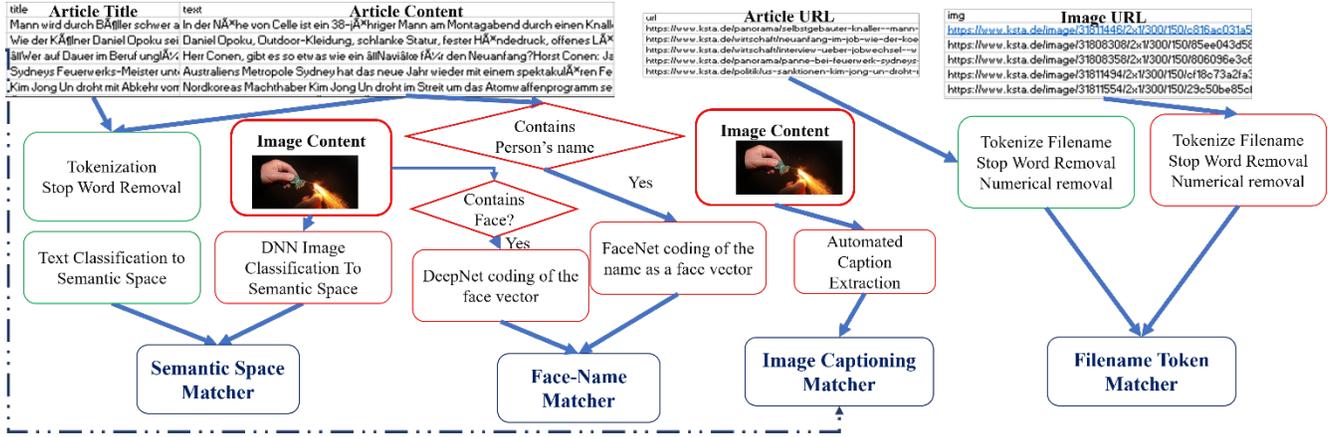


Figure 1 Data feeding and processing pipeline for four different matchers used in the benchmark.

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 article 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 title.

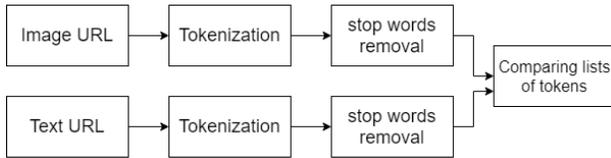


Figure 2 Usage of Image URL

3.4 *The Filename Token Matcher* The filenames of the images extracted from the URLs and the filenames of the articles extracted from the URLs encode semantic connection (as the names are likely crafted by humans). We propose tokenizing the filenames, and discover image-text match based on the number of overlapping tokens, as illustrated in Figure 1.

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 data set into the actual training set and the validation set. The training set included 5135 records, while the validation set included 2384 records. The findings of the training data are shown in Figure 2. Filename Token Matcher produces the best overall results on the training and validation dataset. Based on our findings, we have submitted 3 runs:

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 as follows:
 $RRun1 = 0.5R_{Categorization} + 0.5(R_{Face} + R_{caption})$

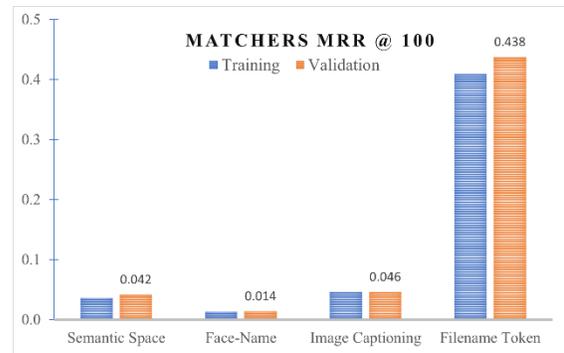


Figure 3 Matcher's MRR@100 during Training

Run2 combines all proposed methods. The first three models are assembled using the same approach as in Run1. This ensemble model is used to create the initial top 100 image list. Then we append the result, which is generated from the filename token matcher, to the end of the top 100 image list.
Run3 is like Run2. The only difference is that we append the result of the last method to the head of the top 100 image list.

Table 1: Results from different runs in the test data set

Run	MRR@100	R@5	R@10	R@50	R@100
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

Result Our proposed approach uses an ensemble design, so our submissions are combined results from three or four models.

5 CONCLUSIONS

The filename token matching method recovered almost 50% of the ground truth (R@100) in the test set, as outlined in Table 1. This is consistent with the findings of the training phase described in Figure 3. This experiment demonstrates the depiction gap in automated image-text correspondence. Human reasoning depicts the same piece of information in different modalities to compliment, not duplicate, the presented information. Semantic image-text connections are unconscious imprinted by humans in the filenames of images and articles.

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, TDH, 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, Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220.
- [10] 2021, R. Luo, G. Shakhnarovich, S. Cohen and B. Price, "Discriminability Objective for Training Descriptive Captions," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6964-6974, doi: 10.1109/CVPR.2018.00728.
- [11] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815-823, doi: 10.1109/CVPR.2015.7298682.