

HCMUS at MediaEval 2020: Image-Text Fusion for Automatic News-Images Re-Matching

Thuc Nguyen-Quang^{*1,3}, Tuan-Duy H. Nguyen^{*1,3}, Thang-Long Nguyen-Ho^{*1,3}, Anh-Kiet Duong^{*1,3},
Nhat Hoang-Xuan^{*1,3}, Vinh-Thuyen Nguyen-Truong^{*1,3}, Hai-Dang Nguyen^{1,3}, Minh-Triet Tran^{1,2,3}

¹University of Science, VNU-HCM, ²John von Neumann Institute, VNU-HCM

³Vietnam National University, Ho Chi Minh city, Vietnam

{nqthuc, nhthuy, ntvthuyen}@apcs.vn, {nhtlong, hxnhath, nhdang}@selab.hcmus.edu.vn, 18120046@student.hcmus.edu.vn,
tmtriet@fit.hcmus.edu.vn

ABSTRACT

Matching text and images based on their semantics has an important role in cross-media retrieval. Especially, in terms of news, text and images connection is highly ambiguous. In the context of MediaEval 2020 Challenge, we propose three multi-modal methods for mapping text and images of news articles to the shared space in order to perform efficient cross-retrieval. Our methods show systemic improvement and validate our hypotheses, while the best-performed method reaches a recall@100 score of 0.2064.

1 INTRODUCTION

News articles represent a complex class of multimedia, whose textual content and accompanying images might not be explicitly related [25]. Existing research in multimedia and recommendation system domains mostly investigate image-text pairs with simple relationships, e.g., image captions that literally describe components of the images [16]. To address this, the MediaEval 2020 NewsImages Task calls for researchers to investigate the real-world relationship of news text and images in more depth, in order to understand its implications for journalism and news recommendation systems [19].

Our team at HCMUS responds to this call by addressing the Image-Text Re-Matching task. Particularly, given a set of image-text pairs in the wild, the task requires us to correctly re-assign images to their decoupled articles, with the aim to understand the implication of journalism in choosing illustrative images.

Our methods mainly concern fusing cross-modal embeddings for automatic matching. We experimented with a range of embedded information, including simple set intersection, deep neural features, and knowledge-graph-enhanced neural features. We combine such features in various ways for various experiments. Finally, we obtain our best result with the ensemble of experimented methods.

2 METHODS

2.1 Metric Learning

The primary idea of this baseline method is using metric learning to project embeddings of image-text pairs to bases of significant similarity. Particularly, we use two approaches to embed image features: *global context embedding* and *local context embedding*. In the first approach, we use the EfficientNet [30], a SOTA classification architecture, to extract features of the image before taking the flatten output features. Our motivation in the latter approach is to harness critical local information from the extracted global context. Thus, we use the bottom-up-attention model [3] to extract the top- k objects based on their confidence score, before passing them over to a self-attention sequential model. For both routines, we employ BERT [12] language model to embed textual content, then project the textual and image

embeddings to the same dimension. Finally, we train our Triplet Loss [15] model with positive and negative pairs from a hard sample miner.

2.2 Image-Text Matching via Categorization

In this method, we train two gradient boosting decision trees [18], one for categorizing images, and the other for categorizing articles. The target categories are ['nrw', 'kultur', 'region', 'panorama', 'sport', 'wirtschaft', 'koeln', 'ratgeber', 'politik', 'unknown'], which are deduced from URLs in the train set.

We use features extracted for images and text to train the decision tree. To augment the data, we use VGG16, InceptionResNetV2, MobileNetV2, EfficientNetB1-7, Xception, ResNet152V2, NASNetLarge, DenseNet201 [10, 14, 17, 27–30, 32] for images, while using pre-trained BERT models [2, 8, 9, 11], and pretrained ELECTRA models [1, 9] to extract contextual features.

We presume that images and articles of the same category might have some relations. Moreover, the rank of matching categories also affects ranking. For example, an image-text pair sharing a 3rd-ranked category might be less relevant than the pair sharing a 1st-ranked category. Hence, instead of using Jaccard similarity, we propose an iterative ranking method that takes into account the order of matched categories. At the k -th iteration, our method first finds top- k categories for each image and top- k categories for each article. Then for each article, we create a list of candidate images whose top- k categories intersect that of the article. This list of candidates at the k -th iteration is concatenated to the final list. Finally, the remaining images that are not candidates are kept in their order and concatenated to the end of the final list.

2.3 Graph-based Face-Name Matching

Based on our observation, in a lot of instances, the publisher uses a portrait of somebody mentioned in the text. We build the *face-name graph* to represent the relation between the name and the face.

Person name extraction: To automatically extract people's name from the text, we use *entity-fishing* [23] – an open-source high-performance entity recognition and disambiguation tool. It relies on Random Forest and Gradient Tree Boosting to recognize named entities, in our case people's names, and link them against Wikidata entities using their word embeddings and Wikidata entities' embeddings.

Face encoding: We use face recognition open-source library [13] to detect and represent the face as 128-dims vector. The tool uses a pre-trained model from the dlib-models repository [20] and chooses ResNet as the backbone for face feature extraction.

Using the train set, we connect each person mentioned in the articles with features extracted from accompanying faces. During testing, we encode the face from the image and aggregate the number of matched faces connected to the people mentioned in the text. Two faces are matched if $L2$ -distance between two vectors less than 0.6. The ranking of images is sorted by the total matched.

Table 1: Submission result

Method	Acc.	Recall@100	MRR@100
TripletLocal	0.0000	0.0248	0.0012
TripletGlobal	0.0002	0.0238	0.0013
Group-Face&Cap	0.0194	0.1322	0.0237
KG-Fusion	0.0051	0.1667	0.0164
Ensemble	0.0075	0.2064	0.0222

2.4 Image-Text Fusion with Image Captioning and Contextual Embeddings

Based on the hypothesis that the description of the image is semantically similar to the title, we build an image captioning model which is inspired by the tutorial Image captioning with visual attention[31]. The model has three main parts:

- **Image feature extractor:** We use EfficientNet[30] for feature extraction. The feature has the shape (8, 8, 2048)
- **Feature encoder:** The features pass through fully connected giving a vector 256-dims.
- **Decoder:** To generate the caption, we use Bahdanau attention[4] and GRU to predict the next word.

We merge the train set with Flickr and COCO for training. We use *fuzzywuzzy* ratio and *partial* ratio string matching to compare captions and articles title. To represent the caption and the title as a vector, we use RoBERTa and doc2vec[22] enwiki_dbow, apnews_dbow. Then, we calculate the similarity of two vectors by cosine similarity. The final score is calculated by:

$$S_{total} = S_{wiki} + S_{apnews} + S_{RoBERTa} + (1 - D_{fuzzy}) + (1 - D_{partial})$$

where S_{wiki} , S_{apnews} , $S_{RoBERTa}$ are cosine similarity of two vectors generated by enwiki_dbow, apnews_dbow, RoBERTa, and D_{fuzzy} , $D_{partial}$ are fuzzywuzzy and partial ratios, respectively.

2.5 Image-Text Fusion with Knowledge Graph-based Contextual Embeddings

We observe that image-text pairs may not have any explicit relationships. Yet, such text-image pairs could still remotely related through layers of abstraction. For example, an article about violence could feature a stock photo of a gun barrel. Although such a stock photo does not literally illustrate the textual content, we understand that a gun conveys a sense of *threat*, which, in turn, is related to *violence*.

Thus, we consider exploiting knowledge graphs. On a knowledge graph, such as BabelNet [24], the concept node of *gun* is also remotely connected with *violence* through intermediate nodes. Thus, we hypothesize that the projection of the textual and imagery content of a news article onto a knowledge graph would be connected, and their embeddings, in turns, could be in close proximity.

To implement this projection, we use EWSER word sense disambiguator [6] to link textual entities from texts to their synsets in the WordNet subset of BabelNet. Then, the mean of accompanied SenSembERT+LMMS embeddings corresponds to these extracted synsets representing the texts. For the images, we first map images to the textual domain. To enhance the method by featuring abstract human-level concepts in the mapping, we decide to use TResNET-L with Asymmetric Loss (ASL) [5, 26] pre-trained on OpenImagesV6[21] to extract multi-label from images. Our decision is grounded since OpenImagesV6 features image-level labels conform with Freebase[7] knowledge graph with figurative labels, e.g., *festivals*, *sport*, *comedy*, etc., while TResNET-L with ASL is the state-of-the-art method for OpenImagesV6 multi-label benchmark. The

extracted lists of labels are also linked with synsets using EWSER, and the mean of these synset embedding vectors represent images.

We then train a canonical correlation analysis (CCA) module with the vector representation on the train set before using it to transform test set vectors. For relatedness measurement, for each test article, we rank all images in the test set using the L_2 -distance between the article vector and image vectors.

3 EXPERIMENTAL RESULTS

3.1 Data preprocessing

The MediaEval 2020 Image-Text Re-Matching benchmark releases three batches of data in total consists of the lede and titles of German news articles and their accompanying images. The first two are used for training, and the last one is used for testing.

For the sake of manual assertion, we decide to translate all the text to English using Google Translate and employ this translated text in our experiments. All data batches are cleaned automatically, with images crawled using the given URLs and pairs with *404 Not Found* URLs dropped from the train set.

3.2 Submissions

First, *TripletLocal* and *TripletGlobal* demonstrate respective methods in Section 2.1. In both submissions, we empirically choose $k = 30$ to embed images with top- k objects, then sort candidate images for each article by the similarity of their embedding to that of the article.

The *Group-Face&Cap* submission, meanwhile, combine three different methods. First, we matches image-article pairs using the method in Section 2.2 with $k = 5$. However, at each iteration, we sort the candidates by S_{total} score mentioned in 2.4. Finally, candidate images matched with the article through the method in Section 2.3 are prioritized to the top of the final result.

The *KG-Fusion* submission manifest the method described in Section 2.5. Specifically, the TResNet-L with ASL model used for multi-label extraction accepts a sigmoid threshold of 0.7, the EWSER disambiguator consumes chunks of 5 tokens, and the target decomposition of the CCA module has 64 components.

Finally, the *Ensemble* submission combines all described methods, weighting each models based on their efficiency. As such, the final ranking of a candidate image is:

$$R_{Ensemble} = w_1 R_{Caption} + w_2 R_{Triplet} + w_3 R_{Face} + w_4 R_{KG-Fusion}$$

where $R_{Ensemble}$, $R_{Caption}$, $R_{Triplet}$, R_{Face} , $R_{KG-Fusion}$ are ranks of the image produced by Group-Face&Cap, TripletGlobal, Face Matching, and KG-Fusion methods, respectively. Weighting factors are empirically chosen to be $w_1 = w_4 = 1$, $w_2 = 0.02$ and $w_3 = 0.25$.

4 CONCLUSION AND FUTURE WORKS

Although, our methods show poor *accuracy*, they systematically increase the performance on the *recall@100* metric. This fact validates our hypotheses that incorporating high-level semantics increase performance. Moreover, our methods yield consistent results, i.e., high-ranking images are of relevance to queried articles. Thus, they can still be useful for building news image recommendation systems as the news-images suitability is not injective in practice. The ensemble method's performance also suggests practical system builders to use multiple methods to handle different aspects of the complex image-text multimodal relation. In future works, we wish to investigate better fusion methods, consider a thorough ablation study for proposed methods, and enhance the dataset for thorough evaluation with information retrieval metrics like NDCG.

Acknowledgments: Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19.

REFERENCES

- [1] 2020. Model from <https://huggingface.co/german-nlp-group/electra-base-german-uncased>. (2020).
- [2] 2020. Model from <https://huggingface.co/T-Systems-onsite/bert-german-dbmz-uncased-sentence-stsb>. (2020).
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-Up and Top-Down Attention for Image Captioning and VQA. *CoRR* abs/1707.07998 (2017). arXiv:1707.07998 <http://arxiv.org/abs/1707.07998>
- [4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. (2016). arXiv:cs.CL/1409.0473
- [5] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2020. Asymmetric Loss For Multi-Label Classification. *arXiv preprint arXiv:2009.14119* (2020).
- [6] Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2854–2864.
- [7] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 1247–1250.
- [8] Malte Pietsch Tanay Soni Branden Chan, Timo Möller. 2020. Model from <https://huggingface.co/bert-base-german-cased>. (2020).
- [9] Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. (2020). arXiv:cs.CL/2010.10906
- [10] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.
- [11] dbmdz. 2020. Model from <https://huggingface.co/dbmdz/bert-base-german-uncased>. (2020).
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019). arXiv:cs.CL/1810.04805
- [13] Adam Geitgey. 2018. Face Recognition. (2018). https://github.com/ageitgey/face_recognition
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 630–645.
- [15] Elad Hoffer and Nir Ailon. 2018. Deep metric learning using Triplet network. (2018). arXiv:cs.LG/1412.6622
- [16] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)* 51, 6 (2019), 1–36.
- [17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
- [18] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., 3146–3154. <https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>
- [19] Benjamin Kille, Andreas Lommatzsch, and Özlem Özgöbek. 2020. News Images in MediaEval 2020. In *Proc. of the MediaEval 2020 Workshop. Online*.
- [20] Davis E. King. 2018. dlib-models. (2018). <https://github.com/davisking/dlib-models>
- [21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV* (2020).
- [22] Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. (2016). arXiv:cs.CL/1607.05368
- [23] Patrice Lopez. 2020. Entity Fishing. (2020). <https://github.com/kermitt2/entity-fishing>
- [24] Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. 216–225.
- [25] NHJ Oostdijk, H van Halteren, Erkan Basar, and Martha A Larson. 2020. The Connection between the Text and Images of News Articles: New Insights for Multimedia Analysis. (2020).
- [26] Tal Ridnik, Hussam Lawen, Asaf Noy, and Itamar Friedman. 2020. TRNet: High Performance GPU-Dedicated Architecture. *arXiv preprint arXiv:2003.13630* (2020).
- [27] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.
- [28] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [29] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261* (2016).
- [30] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2020). arXiv:cs.LG/1905.11946
- [31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. (2016). arXiv:cs.LG/1502.03044
- [32] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. 2018. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8697–8710.