Deep Embedding-based Multimodal Matching for News Articles: Exploring the Effects of Transfer Learning & Data Augmentation

Martin Ludwig Zehetner
Technische Universität Berlin, Berlin, Germany
m.zehetner@tu-berlin.de

Mohamed Amine Dhiab
Technische Universität Berlin, Berlin, Germany
m.dhiab@campus.tu-berlin.de

ABSTRACT

Choosing the right combination of news article image and title is critical for attracting new users and converting latent interest into clicks. In the context of the MediaEval 2021 NewsImages Challenge, we therefore investigated the underlying multimodal matching problem between images and titles of news articles by employing deep embedding-based models to map and match images and text in the same semantic space. Additionally we explored the impact of transfer learning and paraphrasing-based data augmentation schemes on the task performance. We observed a clear improvement in performance using transfer learning approaches, but no consistent improvement using the data augmentation technique we selected. Our best model achieved a mean recall@100 of 0.3488.

1 INTRODUCTION

Online news articles commonly try to convey content in a distinctive and direct manner through combinations of expressive titles and images. Understanding the relationships between these news images and texts, e.g. news headlines, can therefore help to provide insight into a wide variety of different tasks in the news domain.

In this sense, the MediaEval 2021 NewsImages Re-Matching Task aims to advance this investigation by setting a task to re-match decoupled real-world news articles with the images used in said articles [8]. In particular, given a news article the corresponding article image is to be selected from the set of all images.

We attempted to further the understanding of these relationships and solve the task given by using multimodal embeddings, specifically embeddings that map images and texts into the same semantic "news" vector space. These embeddings are generated using variations of the Self-Attention Embeddings for Image-Text Matching (SAEM) model framework [18] and are intended to enable the computation of semantic similarity, with regard to the "news" domain, between texts and images using basic similarity metrics. In this context, we placed a special emphasis on investigating the effects of a transfer learning scheme, using information exploited from image caption data sets, and augmenting the given text data by applying a paraphrase-based approach.

2 RELATED WORK

Matching media objects of different modalities, e.g. text and images, is essential for various multimedia tasks. Learning a common space into which text and image feature vectors can be embedded and then compared in is a typical approach for such tasks [2, 7].

Wu et al. [18] present such an embedding-based SAEM model framework to solve a cross-media retrieval task posed by the image captioning data sets, i.e. images with associated descriptive sentences, Flickr30k [19] or MS-COCO [13]. The SAEM models can be understood as neural networks divided into two branches. The first of these branches takes feature vectors representing salient regions in images as inputs. These vectors and their intra-modal relations are then encoded using self-attention layers and final image embeddings are generated using an average pooling. In the second branch, text inputs are converted to continuous context-sensitive word embeddings using the encoder of a transformer initialized with a pre-trained BERT [3] architecture. The continuous representations are then fed into three distinct 1D Convolutional Neural Network (CNN) layers [9] and a fully connected layer subsequently generates the global text embeddings. Furthermore, to guarantee the mapping of similar images to texts, a weighted combination of a triplet loss [17] and angular loss [16] is used while training the models with image-text pairs. This allows computed embeddings to be directly compared using the cosine similarity measure.

3 APPROACH

3.1 Deep Embedding-based Multimodal Matching using SAEM

A central component of our approach is the generation of directly comparable text and image embeddings using variations of SAEM models. We chose to use the SAEM model framework as it allows easy customization of the input image feature sources and achieved good results in other cross-media retrieval tasks on the Flickr30k and MS-COCO image captioning data sets [18].

We decided to only use the images and article titles from the MediaEval data set [8] in our SAEM variations, due to superior performance in initial tests and restricted translation and paraphrasing resources. For the generation of image feature vectors representing the salient image regions, we used a bottom-up attention mechanism [1] consisting of a Faster R-CNN with ResNet-101 trained on Visual Genomes [11]. We set the dimensions of these feature vectors to (36, 2048). In addition we used the WordPiece tokenizer [3] to process the article titles to generate the initial text input.

It is important to point out that unlike in the introduction of the SAEM model framework [18], in which the focus was only on matching semantically identical images and texts, the focus in our approach lays rather on the somewhat more abstract task of mapping the similarity of image and text elements in the semantics of the online news article context, i.e. which image would be selected to match a given article title. This is to be achieved by means of training and fine-tuning on the provided MediaEval image-title pairs [8]. Concurrently, the training process is used to optimize

MediaEval'21, December 13-15 2021, Online

the hyperparameters, e.g. learning parameters, the use of transfer learning and the use of data augmentation. The trained models are then used to compute the images which best match the article titles based on the cosine similarity of the corresponding embeddings.

3.2 Transfer Learning

Transfer learning broadly describes the use of knowledge learned in one scenario to improve the training process or results in another [6]. Among the most commonly used strategies in the neural network field are various approaches using pre-trained models [14].

In our approach, due to the relatively small size of the MediaEval data set, we decided to investigate and try to exploit the effect of a direct pre-trained model strategy in our task scenario. To this end, we first train our SAEM variants on either the Flick30K or the MS-COCO data set. We then fine-tune the SAEM models using the best versions of the pre-trained models as the initial model states and then train the models using the news images and titles.

3.3 Data Preparation & Augmentation

In general, performance of neural networks can be influenced heavily by the size and quality of the training data sets [15]. In many real world tasks, relatively few data points are available for the training process, such as in our MediaEval task. A potential solution is data augmentation, i.e. the generation of new data points to increase training data diversity [4]. Therefore, we investigated the effects of an augmentation approach consisting of generating a new paraphrased title for each article title in the training set, where the image is mapped to both titles. By doing so, we doubled the amount of training data.

4 EXPERIMENT

4.1 Data Pre-Processing & Augmentation

Initial image feature vectors are generated following the procedure practiced in the SCAN [12] project. As the first step of text preprocessing, we translate the article titles into english using DeepL. For the text augmentation steps, we then used Quillbot to generate paraphrases for the article titles using the default mode with the highest possible abstraction level for the generated phrases [5].

4.2 Model Implementation

For the selected models we used the Adam [10] optimizer with an initial learning rate of 0.001 and a batch size of 64 while training. During pre-training, a decay rate of 0.1 was applied after every 10 epochs, but when training on the MediaEval data the decay was applied only after every 15 epochs. The dimension of the internal word embedding was set to 300 and the dimension of the final multimodal embedding was fixed to 256.

4.3 Experiment Protocol

Training & Evaluation of pre-trained Models. The training described below is performed separately for Flickr30K and MS-COCO. Firstly, the image-annotation pairs are split according to the public split [12]. The SAEM models are then trained for 30 epochs and after each epoch the ratio of recommendations in which at least one relevant image was ranked among the top 1, 5 and 10 is determined on the validation data. Afterwards, the model with the highest mean between these 3 metrics is selected as the best pre-trained model. In our experiment this best model was trained on MS-COCO.

Table 1: Results: Mean Recall@k of the submitted SAEM variants (*pre*: pre-trained, *para*: paraphrased titles used)

SAEM Variations	MR@5	MR@10	MR@50	MR@100
not pre & not para	0.04073	0.07050	0.20836	0.30966
not pre & para	0.04856	0.08512	0.21253	0.30287
pre & not para	**0.0700**	**0.1159**	**0.2585**	**0.3488**
pre & para	0.0653	0.1003	0.2381	0.3248

Training & Evaluation of MediaEval Models. The second phase of the experiment focuses on the comparison of the performance of the SAEM models with different configuration combinations. The "Initial Model State" M and the "Textual Input Type" T can be seen as variables of the configurations. M can represent the initialization of the SAEM model randomly or based on the best pre-trained model. T can represent the use of only translated article titles or the use of translated titles together with paraphrased titles. The data is then split, with the first two batches of the MediaEval data set representing the training data and the third batch used as validation data [8]. Afterwards, for each of the possible four (M, T) combinations, the following steps are performed. First the input data is augmented and processed according to the current T. Then the model is initialized according to the current M. Thereafter, the SAEM model is trained for 50 epochs using the pre-processed training data. Subsequently, the best models of the current configurations are selected by calculating $MRR@N$ for $N = 1, ..., 100$ for the matchings, according to cosine similarity, of the trained model on the validation data set. The four submission models represent the best performing models in each configuration.

5 RESULTS & FUTURE WORK

The performance of our four submitted models on the MediaEval test set [8] can be seen in Table 1. While no massive performance differences are present, clear performance gaps can be observed nevertheless. Our best performing variant is pre-trained but does not use the paraphrase-based data augmentation method, reflecting the general observed tendencies. As such, consistently better performance is observed for the pre-trained variants, while the use of our data augmentation approach shows no consistent performance improvements. Concluding, we recognize that our results indicate that exploiting learned information from similar task domains through transfer learning can be highly beneficial in news re-matching scenarios with small amounts of training data, such as the considered MediaEval 2021 NewsImages task.

In this sense, our observations suggest that further investigation regarding the exploitation of externally learned information may be worthwhile. In addition to more detailed analysis regarding the influence of information learned in less or more related tasks and with less or more available data, investigating the explicit addition of available contextual information, such as knowledge related to identities or locations, could allow for further valuable insights.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805

[4] Steven Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. *CoRR* abs/2105.03075 (2021).

[5] Tira Fitria. 2021. QuillBot as an online tool: Students' alternative in paraphrasing and rewriting of English writing. *Englisia: Journal of Language, Education, and Humanities* 9, 1 (2021), 183–196.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[7] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[8] Benjamin Kille, Andreas Lommatzsch, Özlem Özgöbek, Mehdi Elahi, and Duc-Tien Dang-Nguyen. 2021. News Images in MediaEval 2021. In *MediaEval 2021 Proceedings*. MediaEval.

[9] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *CoRR* abs/1408.5882 (2014). arXiv:1408.5882

[10] Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[11] 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 Fei-Fei Li. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *CoRR* abs/1602.07332 (2016). arXiv:1602.07332

[12] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[13] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312

[14] Leeja Mathew and Bindu. 2020. A Review of Natural Language Processing Techniques for Sentiment Analysis using Pre-trained Models. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. 340–345.

[15] Luis Perez and Jason Wang. 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *CoRR* abs/1712.04621 (2017). arXiv:1712.04621

[16] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. 2017. Deep Metric Learning with Angular Loss. *CoRR* abs/1708.01682 (2017). arXiv:1708.01682 http://arxiv.org/abs/1708.01682

[17] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2015. Learning Deep Structure-Preserving Image-Text Embeddings. *CoRR* abs/1511.06078 (2015). arXiv:1511.06078 http://arxiv.org/abs/1511.06078

[18] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2088–2096.

[19] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2 (2014), 67–78.