Image-Text Rematching for News Items using Optimized Embeddings and CNNs in MediaEval NewsImages 2021

Tom Sühr, Ajay Madhavanr, Nasim Jamshidi Avanaki, René Berk, Andreas Lommatzsch
Technische Universität Berlin
Berlin, Germany
{tom.suehr,jamshidiavanaki,ajay.m.ravichandran,rene.m.berk,andreas.lommatzsch}@campus.tu-berlin.de

ABSTRACT

Finding a matching image for a news article is a core problem in the creation of traditional and online newspapers. The task of image-text matching has thus become a vibrant research area in computer science. The performance of state-of-the-art image retrieval systems on various benchmarks is excellent. However, they all rely on datasets with a detailed textual description of the images or on very large training collections. In this work, we optimize image-text matching algorithms for a small dataset based on the data of a single newspaper. Our optimized processing pipeline and the computed configurations reach precise results. The evaluation results obtained in the MediaEval NewsImages benchmark significantly outperforming the algorithms from previous years.

1 INTRODUCTION

The process of selecting images for news articles in the multimedia industry is crucial. Images play a significant role of the storytelling process. They are used to attract the user's attention, thus achieving a high number of clicks or high average dwell time per user. However, finding a good image that matches the news article in a single picture is a hard task. Automating this task can provide beneficial effects in different areas, e.g. leveraging the efficiency of publishing articles, saving costs and human resources. Finding a relationship between a text and an image is a problem that is researched in the field of recommender systems. Several papers exist that achieved good results, but most works rely on huge generic data collections. In this paper we develop models for a specific newspaper that has its own image database, a different journalistic style and a significantly smaller amount of data. We evaluate our models using the data provided in the MediaEval 2021 NewsImages Challenge. A detailed description of the dataset and the evaluation metrics are discussed in the Task Overview paper [11].

Our approach is inspired by recent works in the domain of text and image encoding as well as advanced Image-Text Matching methods. We analyzed commonly used CNNs (pretrained on ImageNet [6]) for the image encoding, such as ResNet [7], VGG [8], and DenseNet [9]. For the efficient encoding of texts and their contexts, the use of text embeddings has shown promising results [2, 13, 18]. Recent image-text matching algorithms are usually based on two branches for extraction of image and text representations, for which then the computed representations are aligned for both modalities in a joint semantic space [1, 3, 15, 20]. Critical aspects are the size of the dataset and its' features, the specific vocabulary of the domain as well as the models for transforming the textual and visual data.

In this work we research the degree to which the textual and visual contents of a news article are related. Our developed model should be able to recommend a ranked list of related images, for a given text input. We analyze, whether state of the art image-text matching architectures like VSE work for a small and homogeneous dataset from just one newspaper. Furthermore, we research which adaptations are needed to improve the performance in the MediaEval NewsImages scenario.

The rest of this paper is organized as follows: Sec. 2 explains our approach and the implementation. In Sec. 3 we present the performance results and discuss the specific strengths of the models. Finally, we summarize our work and discuss extensions in Sec. 5.

2 APPROACH

Our approach follows the general architecture of *Visual Semantic Embeddings* [5]. The core idea of this architecture is to embed both, text input and image input, into a joint embedding space. In this joint embedding, matching text-image pairs can then be done based on distance or similarity measures such as cosine similarity. Thus, the challenge of this approach is to learn such a joint embedding and to extract those features which characterize image and text pairs best. Fig. 1 shows our architecture and the components.

Image Encoding. The image encoding consists of three steps: (i) preprocessing, (ii) feature extraction and (iii) linear mapping into the joint embedding size. In the preprocessing, we normalize the RGB values of the pixels and resize the images to 250 pixels. In the second step, the preprocessed image are fed into a pretrained CNN (VGG19 [16]).

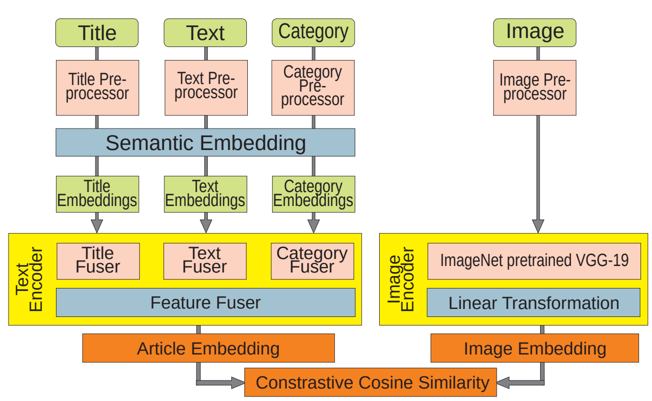

Figure 1: Our system architecture.

MediaEval'21, 13-15 December 2021, Online

Text Encoding. The text encoding stands in the center of this work. One special feature of news image retrieval is that more than one textual input might exist. In the NewsImages task the article *title*, the *snippet* and the article *category* are provided. We employ three preprocessing steps for each textual input. We apply stop words removal and stemming (using NLTK). In order to get the same number of word vectors for each input, we picked a constant length and cropped or extended the input to that length. Subsequently, we vectorized the text and compute a semantic embedding [4, 13, 14]. Due to the limited amount of data, we test pretrained embeddings.

First Fusion Layer: The task of the first fusion layer is to reduce the three matrices to three vector representations. Embedding each textual input on a word level, yields three matrices of the sizes: $(a = 5, w)$ for the title input, $(b = 25, w)$ for the text input and $(c = 1, w)$ for the category input; the word embedding size w is 300.

Stacking and Second Fusion Layer: Receiving three inputs of size $(1, w)$ for title, text and category, the next step is to fuse all three representations and transform them in the size of the joint embedding space of the size $(1, d)$. In order to achieve that, we stack all three input representations of $(1, w)$ which yields one vector of size $(1, 3w)$. Another fully connected layer of size $(3w, d)$ then maps the stacked representations to the size of the joint embedding space $(1, d)$.

Contrastive Loss. A multitude of loss functions exist to train the joint embedding space of article and image embedding. The loss function should ensure with the learned model that the similarity between an article and the true matching image is higher than the similarity to other images and vice versa; the use of a *margin-based contrastive loss* fulfills these requirements [3, 10, 12, 12, 13]. For the image embedding x_i and the article embedding x_t we first define the similarity measure as the inner product of both vectors: $s(i,t) = \langle x_i, x_t \rangle : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$. In our implementation we use the L2-normalized vectors x_i, x_t for computing the similarity.

3 EXPERIMENTS AND RESULTS

We tested different configurations focusing on finding optimal embeddings and hyperparameters.

The experimental results on the Mediaeval test set of size 3022 are shown in Table 1. The experiments reveal that the linear layer for the dimension reduction of the textual inputs outperforms adaptive max pooling in all compositions with a margin of almost 10% in the settings **D** and **C**. However, the adaptive max pooling component performed extremely well in most recent works. The reason for that seems to be the difference between pretrained word embeddings and fine-tuned word embeddings. The adaptive max pooling can consider positions in the textual input. The fully connected layer on the other hand is better suited for the pretrained embeddings because it will learn an average importance of the different positions.

This suggests that the linear layer instead of adaptive max pooling is more adaptive to the word embedding.

In addition, we find the models **B** and **D** differ in the performance whereas the models only differ in the used data for learning the embeddings. While model B with the word embedding trained on wiki achieves higher recall at position 5 and 10, the same model with our word embedding trained on German news article data, performs better at recall at 50 and 100. The wiki-based embedding has a more fine-grained differentiation between words. Thus, given a word and a slight modification, the wiki embedding is able to produce two significantly different representations. Furthermore, the vocabulary of the wiki embedding is much larger than the vocabulary of our custom embedding. The custom embedding performs better if we look at a large interval of the ranking (r@50, r@100) because it is better suited to embed news article words. In summary, the custom embedding provides a better representation of the articles compared to the embeddings computed based on the wiki corpus. However, when fine grained differentiations between words are relevant, the wiki-based embedding performs better.

Model	r@5	r@10	r@50	r@100
A: Word Embeddings MaxPool + wiki	1.93%	3.76%	12.59%	19.37%
B: Word Embeddings Linear + wiki	**4.49%**	**7.26%**	20.99%	31.91%
C: Word Embeddings MaxPool + custom	2.92%	4.60%	14.36%	24.86%
D: Word Embeddings Linear + custom	3.97%	7.10%	**21.57%**	**33.26%**
E: Word/Subw. Emb. Linear + wiki	2.56%	4.70 %	16.19 %	26.68%

Table 1: The evaluation results obtained for the evaluation set for the analyzed models.

4 CONTRIBUTIONS

In this work we made the following contributions: First, we showed that state of the art architectures perform significantly worse on a small, non-descriptive and homogeneous dataset. Secondly, we showed that the performance of embeddings trained on large corpora such as Wikipedia, improve the performance in the top 10 retrieved images while tailored embeddings (to a specific style of a newspaper) improve the top 100 performance. Thirdly, we provide our code[1]. For future Mediaeval participants and other researchers for benchmarking purposes and to build upon.

5 CONCLUSION

We have investigated how to adapt state of the art image-text matching systems to a small, homogeneous and specific dataset. We analyzed existing and well performing image-text matching systems like VSE, identified components which do not work well with our dataset, and systematically tested possible substitutions for them. Our experiment show that the non-viability of components like the trainable word embeddings have impacts on the viability of other components, e.g. the adaptive max pooling. We further showed that we can successfully substitute these components in an easy way and achieve reasonable performance on our data. Future work could investigate other substitutions for the identified components, e.g. optimizing the word embeddings with respect to the loss. Furthermore, future projects could research other configurations or even inputs for the image encoding layer as well as investigating fairness aspects. It might be that our strategy works well for political articles but not for sports articles. Thus, analyzing and incorporating fairness aspects of matching and ranking [17, 19] could normalize the performance of our model across various article subjects.

[1] https://github.com/tsuehr/News-text-image-matching

REFERENCES

[1] Yanbei Chen and Loris Bazzani. 2020. Learning joint visual semantic matching embeddings for language-guided retrieval. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. Springer, 136–152.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

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

[4] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. (2018). https://github.com/fartashf/vsepp

[5] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. (2013).

[6] 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.

[7] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2333–2338.

[8] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. 2017. Learning robust visual-semantic embeddings. In Proceedings of the IEEE International Conference on Computer Vision. 3571–3580.

[9] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, and others. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics 5 (2017), 339–351.

[10] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3128–3137.

[11] Bennamin Kille, Andreas Lommatzsch, Özlem Özgöbek, Mehdi Elahi, and Duc-Tien Dang-Nguyen. 2021. News Images in MediaEval 2021. In Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2021. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2882/

[12] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014).

[13] Fangyu Liu, Rémi Lebret, Didier Orel, Philippe Sordet, and Karl Aberer. 2020. Upgrading the Newsroom: An Automated Image Selection System for News Articles. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16, 3 (2020), 1–28.

[14] Fangyu Liu, Rongtian Ye, Xun Wang, and Shuaipeng Li. 2020. HAL: Improved text-image matching by mitigating visual semantic hubs. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 11563–11571.

[15] Lin Ma, Wenhao Jiang, Zequn Jie, Yu-Gang Jiang, and Wei Liu. 2019. Matching image and sentence with multi-faceted representations. IEEE Transactions on Circuits and Systems for Video Technology 30, 7 (2019), 2250–2261.

[16] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[17] Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty. 2019. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 3082–3092.

[18] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. 2020. Learning Dual Semantic Relations with Graph Attention for Image-Text Matching. IEEE Transactions on Circuits and Systems for Video Technology (2020).

[19] Meike Zehlike, Tom Sühr, Carlos Castillo, and Ivan Kitanovski. 2020. Fairsearch: A tool for fairness in ranked search results. In Companion Proceedings of the Web Conference 2020. 172–175.

[20] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV). 686–701.