Methods for Text-Image-Rematching using Pair-wise Similarity and Canonical Similarity Analysis

Kani Abdul, Kiran Kiran, Max Rudat, Alexandros Vasileiou, Andreas Lommatzsch
Technische Universität Berlin, Germany
{kani.abdul,k.kiran,rudat,a.vasileiou,andreas.lommatzsch}@campus.tu-berlin.de

ABSTRACT

Matching images to text plays an important role in cross-media retrieval and research has proven this to be an underestimated challenge. This problem is addressed by the MediaEval 2021 NewsImages Challenge with the goal to gain more insights into the real-world relationship of news articles and images. We develop models for re-establishing the connection of a news article to its corresponding image using datasets of a German news publisher ("task 1"). Our approaches follow the idea of pairwise similarity learning and are optimized by algorithmic hill climbing. Additionally, we employ Canonical Correlation Analysis as an approach using joint embedding learning. The evaluation shows that our approaches produce good results for the underlying image-text rematching task, yet require further optimization to yield stable prediction performance.

1 INTRODUCTION

Multimedia content is accompanying our everyday life. News articles are one form of multimedia that are characterized by textual content accompanied by imagery. The assumption that a simple relationship underlies this connection has frequently turned out to be oversimplified in research [2]. The MediaEval 2021 NewsImages task aims at addressing this challenge by investigating the real-world relationship of news and images. The challenge provides a dataset consisting of three batches training data and one batch for the evaluation. The performance of the participants' algorithms is evaluated on a test batch. The evaluation metrics Recall@k and Mean Reciprocal Rank are used [3].

The challenge of image-text retrieval has been addressed broadly in research around multimedia analysis [8]. Deep image-text matching serves as one frequently used approach for this scenario. Zhang and Lu [9] classify the main approaches based on deep learning into two categories: pairwise similarity learning and joint embedding learning. For pairwise similarity learning, the main idea is to learn a similarity network for predicting the score of image-text pairs [8]. As for the other category of joint embedding learning, a joint latent space is defined in that the vectors of texts and images can be compared directly. The typically used learning methods belonging to this category are canonical correlation analysis (CCA) and bi-directional ranking loss [9].

Based on the existing methods, we develop two text-image matching strategies optimized for the specific requirements of the NewsImages task. In Sec. 2 we explain the preprocessing and the steps of our strategies in detail. Sec. 3 presented the evaluation results. Finally, the overall findings are discussed in Sec. 4.

MediaEval'21, 13-15 December 2021, Online

2 APPROACH

We develop two approaches addressing the text-image rematching task. This section discusses the steps of our approaches.

Data Preprocessing. We preprocess the provided data for efficiently computing similarity scores. Firstly, we translate the image labels (computed by VGG-19 trained on ImageNet) from English to German using Google Translate [7]. We decided to translate the image labels instead of the article snippets due to the smaller volume to translate. In addition, we enhanced the dataset by extracting the item category (e.g. 'koeln', 'panorama' 'wirtschaft' 'politik') from the article URL. In the next step, we normalize the dataset by removing stop words, punctuation marks, spaces, special characters, and digits. After removing the above tokens, we employ *part-of-speech (POS)* tagging to identify the nouns in the dataset. Finally, we perform *Morphological Processing* ("lemmatization") for creating the dataset.

In addition to the standard preprocessing, we integrate Open-de-WordNet [6] enabling us to consider synonyms when computing the similarity score. Moreover, we implement an outlier detection and removal strategy based on the Z-score for the terms derived from the textual description. If a word's z-score is larger than 3.0 (3 standard deviations away from the mean), then it is considered an outlier and gets removed from the dataset. Adding these two derives fields to our dataset, enables us to research, whether additional preprocessing improves the performance.

Pairwise similarity learning & algorithmic hill climbing. Our first approach follows the idea of pairwise similarity learning. The objective is to compute a similarity score for each image-article pair to be used to construct 1-to-1 matches. We implement this using spaCy similarity from the natural language processing (nlp) module SpaCy [4]. SpaCy offers two methods to find the similarity between words: one based on context-sensitive tensors and another one based on word vectors [1]. We utilize the later method and generate a similarity matrix containing the similarity scores for each image-text pair. Our pre-processed dataset gives different options for setting the input for computing the similarity scores. For example, we tested to include only the words within the article text or only the words within the article title. Analogously for the images, we have data from 10 different labeler configurations, each generating at least 2 image labels with different label probabilities.

For computing the best parameter configuration, we make use of algorithmic hill climbing. Starting with an (arbitrary) initial configuration, the solution is incrementally adapted [5]. We implement this by first initializing the parameter (e.g. set labelprobability to 0.0,

number of considered labels per image to nine and the remaining parameters to `False`). Then we go iteratively through the parameter space and compute for each parameter the (locally) optimal value. For the optimization of each parameter, we randomly select 1,000 samples.

The parameter configurations are evaluated using matrices containing pairwise similarity scores with the columns representing the image IDs, the rows the article IDs and the correct matches being located along the diagonal of the matrix. The score is computed using the variables *Row counter*, *Column counter*, and *Total counter*.

For each generated parameter configuration, the performance is evaluated as follows: if the similarity score for the diagonal value is higher than the other scores within its row and column, then the index of the total counter is increased by 1. If this condition is not met, then it is checked whether the diagonal value is higher than the other scores within its row and the row counter index increased by 1 dependently. If both conditions are not met, the column counter index is increased by 1. Once the values of all the 3 counting variables are set, the performance is calculated by comparing the values of the row and column counters: if the row counter is larger than the column counter, then the row counter is divided by the number of pairs (n) and returned. Otherwise the column value divided by the number of pairs (n) is returned as the final score and evaluated by our hill climbing algorithm.

Canonical Correlation Analysis. As a second approach, we apply the SKLEARN Canonical Correlation Analysis. On the preprocessed dataset, the spaCy implementation of *Word2Vec* is used for computing a vector representation (having 300 dimensions) of the dataset. We use a random sample of size 1,500 data points, utilizing the article text and image labels columns. Then we split this set into train set (2/3) and validation set (1/3). Initial tests of CCA showed a poor performance; that is why, we adapted the method. We applied *Kernel PCA (kPCA)* to transform the data through a *Radial Basis Function (RBF)* expansion, limited to utmost 700 generated data dimensions. The kPCA transformation is applied on both the train and test data to keep compatible and comparable dimensions. Then, we train a CCA instance on the train data set and evaluate the model on both the train and validation set. The evaluation is performed based on the predicted vector for each of the article text vectors. Since the CCA-predicted vectors are most likely not corresponding to actual word2vec image label transformations, we compare each CCA prediction to all of the word2vec image label vectors using the cosine similarity measure, thus constructing a similarity matrix. This allows us to make 1-1 article texts to image label mappings.

3 EVALUATION

The evaluation (by the task organizers) shows that out pairwise similarity-based approach reaches $Recall@100 = 0.21$, the CCA-based approach reaches $Recall@100 = 0.24$. Even though CCA outperforms the pairwise similarity-based approach in general, the recall score for k=5 and k=10 is higher with hill-climbing, indicating that low-level semantics are found more effectively using the straightforward method of pairwise similarity learning. The better evaluation scores for higher k observed for CCA show that CCA performs better with regard to high-level semantic similarity.

Table 1: The table shows the evaluation results. Both developed approaches outperform the baseline "random". Hill climbing provides better results than the CCA-based method for short result lists ($k < 50$). The CCA-based approach outperforms Hill climbing based on Recall@50 and Recall@100.

Strategy	CCA-based	Hill-climbing	Random
MRR@100	0.018	0.019	0.003
Recall@5	0.018	0.021	0.002
Recall@10	0.034	0.041	0.005
Recall@50	0.136	0.125	0.026
Recall@100	0.236	0.206	0.051

We analyze the parameter settings for maximizing the performance for the **pairwise similarity learning approach**. We find that the used configuration considers for each image the 8 labels with the highest score (no label probability score has been applied). This indicates that a detailed image description is crucial for the text-image rematching task. Furthermore, we find that considering the title in addition to the article snippet does not improve the performance. Moreover, our analysis shows that the replacement of words with the first word of their synsets as well as the removal of outliers and duplicates are not activated for the final evaluation.

Analyzing the parameters used by **CCA** we find, that considering lemmatized article texts and image labels does not yield the optimal results. Using Kernel PCA with a RBF kernel, the R2 score, which indicates how well the regression model fits the observed data, was greatly increased for the train set (achieving values up to 99.8% with k=100). The performance on the test set reached a R2 score of 38.5% (k=100); thus this method outperformed the hill-climbing optimized pairwise similarity learning approach, that reached an R2 score of 32.6%.

4 CONCLUSION

The evaluation shows, that both approaches yield robust results for the image-text re-matching. The results slightly outperform the best results from MediaEval NewsImages 2020. The CCA-based approach reaches a recall@100 score of 23.6% on the evaluation set; the hill climbing approach based on pairwise similarity learning yields a recall@100 score of 20.6%. Due to limited resources, we have tested only a restricted set of parameter configurations; we think that a further parameter optimization will improve the performance. For our CCA model, we observe a performance difference between training and testing set, indicating the presence of overfitting. The overfitting could be tackled by applying regularization or adding a dropout layer (eliminating features with a low impact). Furthermore, a penalty-based component could be used for boosting articles considering the margin-based error. The ranking of the top-k images for an article could then be optimized significantly.

Furthermore, an alternative image labeling component should be considered to getting a more detailed image description that could be matched with the article text. This is based on the observation that we observed a better performance when considering more images labels.

REFERENCES

[1] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020). https://doi.org/10.5281/zenodo.1212303

[2] Benjamin Kille, Andreas Lommatzsch, and Özlem Özgöbek. 2020. NewsImages: The role of images in online news. In *Proceedings of the MediaEval Benchmarking Initiative for Multimedia Evaluation 2020*. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2882/

[3] Benjamin 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/

[4] Fouad Omran and Christoph Treude. 2017. Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. (05 2017). https://doi.org/10.1109/MSR.2017.42

[5] Dimitris Papadias. 2000. Hill Climbing Algorithms for Content-Based Retrieval of Similar Configurations. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 240–247. https://doi.org/10.1145/345508.345587

[6] Melanie Siegel and Francis Bond. 2021. OdeNet: Compiling a German Wordnet from other Resources. In *Proceedings of the 11th Global Wordnet Conference (GWC 2021)*. 192–198. https://www.aclweb.org/anthology/2021.gwc-1.22

[7] Google Translator. 2020. (2020). https://pypi.org/project/googletrans/

[8] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020. Cross-Modal Attention With Semantic Consistence for Image–Text Matching. *IEEE Transactions on Neural Networks and Learning Systems* 31, 12 (2020), 5412–5425. https://doi.org/10.1109/TNNLS.2020.2967597

[9] Ying Zhang and Huchuan Lu. 2018. *Deep Cross-Modal Projection Learning for Image-Text Matching*. Springer International Publishing, Cham, 707–723.