

# Cross-modal Networks and Dual Softmax Operation for MediaEval NewsImages 2022

Damianos Galanopoulos<sup>1,\*</sup>, Vasileios Mezaris<sup>1</sup>

<sup>1</sup>Information Technologies Institute / Centre for Research & Technology Hellas, Thessaloniki, Greece

## Abstract

Matching images to articles is challenging and can be considered a special version of the cross-media retrieval problem. This working note paper presents our solution for the MediaEval NewsImages benchmarking task. We investigated the performance of two cross-modal networks, a pre-trained network and a trainable one, the latter originally developed for text-video retrieval tasks and adapted to the NewsImages task. Moreover, we utilize a method for revising the similarities produced by either one of the cross-modal networks, i.e., a dual softmax operation, to improve our solutions' performance. We report the official results for our submitted runs and additional experiments we conducted to evaluate our runs internally.

## 1. Introduction

In this paper, we deal with the text-to-image retrieval task adapted for the needs of the MediaEval 2022 NewsImages task [1]. As Internet speed increases, news sites publish multimedia content in their online news article. Images and videos are important to better convey the message the textual article wants to convey to readers. So, associating news articles with multimedia content is crucial for several research tasks such as cross-modal retrieval and disinformation detection. To deal with image retrieval using textual articles as input queries, we utilize two cross-modal networks, a pre-trained one (CLIP [2]) and a trainable one, the  $T \times V$  model [3]. Moreover, similarly to [3], we adopt a dual-softmax operation to recalculate the initially computed article-image similarities, an approach that leads to improved performance.

## 2. Related Work

Text-image association is a challenging task that has gained a lot of interest in recent years. The task has been extensively examined in the multimedia research community e.g. see [4] [5], and there is consensus that the evolution of deep learning methods has boosted performance. Indicative relevant methods include [6], where an object detector is pre-trained to encode images and visual objects on images and a cross-modal model is trained to associate visual and textual features; and [7], where a context-aware attention network is proposed that focuses on important regions within images to extract possible correlations between image regions and words.

NewsImages is a relatively new and highly specific task, and limited research has been done on it. Focusing on the previous year's NewsImages participations, HCMUS [8] proposed a solution based on the power of the pre-trained model CLIP [2] along with sophisticated text

---


*MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online*

\*Corresponding author.

✉ dgalanop@iti.gr (D. Galanopoulos); bmezaris@iti.gr (V. Mezaris)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

preprocessing, which achieved the best performance. In [9] a visual topic model was proposed to align topics illustrated on images with textual topics using knowledge distillation training.

### 3. Approach

#### 3.1. Data pre-processing

We preprocess both training and testing textual data in order to fully exploit our approach’s power. First, we use a language detector of the *lingua* python package to detect the article’s language. Then we use a translator model from Hugging Face Transformers package [10] to translate German articles (title and text) into English.

#### 3.2. Pre-trained model: CLIP

We utilize an open-source implementation of CLIP [2], the openCLIP [11], as our pre-trained model. To obtain text and image feature representations, we use the ViT-H/14 pre-trained model. For a given article, in order to retrieve the most relevant images from the test set, we calculate the cosine similarity between the article’s title (or article’s text) CLIP embedding and the embeddings of all test images. Then the top-100 most relevant images are selected in a ranked list, from the most relevant to the least relevant image.

#### 3.3. Trainable model: $T \times V$

In parallel to CLIP, we examined a modification of the  $T \times V$  model [3] adapted to deal with images instead of videos. The  $T \times V$  model utilizes textual and visual features and encodes them into multiple joint feature spaces. In these spaces, instances from different modalities (e.g., textual snippets, images, etc.) are directly comparable; thus, their similarity can be calculated. In contrast to the original version of  $T \times V$ , here we treat the image as a special video version that consists of only one frame. Moreover, we use only one textual and image feature (obtained from the openCLIP ViT-H/14 pre-trained model) as the initial representation instead of multiple ones. In essence, in this way we try to adapt the pre-trained CLIP representations specifically to the NewsImages task.

Since the NewsImages-provided training datasets are relatively small, we first utilize a large dataset that contains news articles, images, and captions, the NYTimes800k dataset [12], to pre-train our  $T \times V$  model. Subsequently, we merge the NewsImages-provided training datasets and we split this overall dataset in a 80-20% manner to finetune our model. We use the 80% portion of the dataset to train the model and the remaining 20% to validate the performance of our approach for selecting the best possible model.

#### 3.4. Dual-softmax similarity revision

In order to improve the performance of our method we utilize a similarity revision approach at the retrieval stage, both for CLIP and  $T \times V$ . We calculate the similarities between all images from the test set and all testing articles, resulting in a similarity matrix  $\mathbf{Z} \in \mathcal{R}^{C \times D}$ , where  $C$  is the number of the testing article queries and  $D$  the number of test images. To revise the calculated similarities, we apply two cross-dimension softmax operations (one row-wise:  $dim = 0$ , and one column-wise:  $dim = 1$ ) as follows:

$$\mathbf{Z}^* = \text{Softmax}(\mathbf{Z}, dim = 0) \odot \text{Softmax}(\mathbf{Z}, dim = 1)$$

where  $\odot$  denotes the Hadamard product.

## 4. Submitted Runs and Results

We submitted five runs for each testing dataset (TW, RT, RSS), as detailed below:

- **Run #1** (iti\_certh\_clip\_run\_1): This uses the text and image CLIP embeddings and calculates the cosine similarity between the embedding of an article and all images. Then for each article, the 100 most relevant images are selected.
- **Run #2** (iti\_certh\_clip\_ds\_run\_2): As **Run #1**, additionally using the dual softmax (DS) revision method to recalculate the article-images similarities.
- **Run #3** (iti\_certh\_TxV\_run\_3): We train the  $T \times V$  model using a merged dataset consisting of the 80% of the three provided training datasets. We use this trained model to calculate the three testing datasets'  $T \times V$  article title and images embeddings. Finally, we use the cosine similarity to compute the similarities between a testing article and all images and the 100 most relevant images are selected.
- **Run #4** (iti\_certh\_TxV\_ds\_run\_4): Similarly to **Run #3**, additionally using dual softmax revision to revise the computed similarities.
- **Run #5** (iti\_certh\_TxV\_text\_ds\_run\_5): Similarly to **Run #4** but using the full text of the articles instead of just the title that was used in all the above runs.

We present the official results on three testing datasets and results from the internal experiments we conducted in order to evaluate our methods and select our final runs. The Recall@K, where  $K = 5, 10, 50, 100$  and the Mean Reciprocal Rank (MRR) are used as evaluation metrics.

Table 1.A presents the results on the three testing datasets evaluated officially by the task organizers. Run #2 (CLIP + DS) performs the best on all datasets in MRR terms and on RSS and RT in Recall@K terms, while on the TW dataset the results are mixed. The dual softmax operation is beneficial for the raw CLIP embeddings, but it has limited effect on our trainable solutions ( $T \times V$ ). Moreover, Run #5 ( $T \times V$  using articles' full text) achieves lower scores than the other runs on the RSS and RT datasets, but on the TW dataset performs comparably to Runs #3 and #4.

The above official results contrast with the findings of our internal experiments, conducted prior to the release of the official results. Table 1.B presents our internal results on the 20% of the provided training dataset (using the remaining 80% for training and validation where necessary). We conducted these experiments to select our best-trained models and examine our runs' performance. From these preliminary experiments, we had concluded that Runs #3 and #4 constantly outperform the rest of the runs in every dataset, i.e. our training step seemed to be beneficial for performance.

The distribution diversity between the task's official training and testing datasets could explain the contrast between the official results and our findings. Our experiments were conducted on an 80-20% split of the official training set, so our internal-experiments test set is closely related to our training set, and this is beneficial for our experiments. Contrarily, the official test set is probably more different, as it was collected at a much later time than the training set; in this case the original CLIP model, which was trained on much larger and more diverse datasets, is more suitable to address this task.

## 5. Conclusion

In this work we proposed a solution for the MediaEval NewsImages task using state-of-the-art text and image representations calculated from a pre-trained cross-modal network, a task-adapted trainable cross-modal network and a similarity revision approach. We concluded from

**Table 1**

Evaluation results for the three testing datasets (RSS, RT and TW) for the five submitted runs.

A. Official evaluation results for the five submitted runs.

		R@5	R@10	R@50	R@100	MRR
RSS	Run #1	0.61000	0.68267	0.82067	0.86533	0.49013
	Run #2	<b>0.62133</b>	<b>0.69333</b>	<b>0.82667</b>	<b>0.87400</b>	<b>0.49800</b>
	Run #3	0.59933	0.68267	0.80800	0.86067	0.47901
	Run #4	0.60067	0.68267	0.80800	0.85667	0.47664
	Run #5	0.59267	0.68000	0.81267	0.85400	0.46889
RT	Run #1	0.42733	0.52267	0.71800	0.80667	0.30875
	Run #2	<b>0.46200</b>	<b>0.54667</b>	<b>0.75533</b>	<b>0.83400</b>	<b>0.33370</b>
	Run #3	0.43933	0.53733	0.74667	0.82533	0.31131
	Run #4	0.43933	0.53733	0.75200	0.82333	0.31039
	Run #5	0.37267	0.46200	0.65667	0.72667	0.27638
TW	Run #1	0.65667	0.72867	0.86000	0.91200	0.54209
	Run #2	<b>0.66267</b>	0.73200	0.86333	0.91133	<b>0.54645</b>
	Run #3	0.66133	0.74133	<b>0.86600</b>	0.90933	0.53554
	Run #4	0.65333	0.74133	0.86467	0.91000	0.53268
	Run #5	0.65733	<b>0.74600</b>	0.86333	<b>0.91267</b>	0.53920

B. Results on a random 80-20% split (training-testing) of the training dataset.

RSS	Run #1	0.8125	0.8375	0.9250	1.000	0.7590
	Run #2	0.8125	0.8375	0.9500	<b>1.000</b>	0.7730
	Run #3	<b>0.8750</b>	0.9250	<b>0.9875</b>	<b>1.000</b>	0.8014
	Run #4	<b>0.8750</b>	<b>0.9375</b>	<b>0.9875</b>	<b>1.000</b>	<b>0.8147</b>
	Run #5	0.8125	0.8875	<b>0.9875</b>	<b>1.000</b>	0.7672
RT	Run #1	0.6234	0.7247	0.8649	0.9377	0.4800
	Run #2	0.6390	<b>0.7559</b>	0.8857	0.9325	0.5040
	Run #3	<b>0.6546</b>	0.7377	<b>0.9039</b>	0.9455	<b>0.5065</b>
	Run #4	0.6442	0.7507	<b>0.9039</b>	0.9481	0.5045
	Run #5	0.6364	0.7325	0.8987	<b>0.9507</b>	0.4885
TW	Run #1	0.6826	0.7630	0.8957	0.9444	0.5540
	Run #2	0.6882	0.7664	0.9070	<b>0.9501</b>	0.5610
	Run #3	0.7030	<b>0.7721</b>	0.9036	0.9431	0.5695
	Run #4	<b>0.7075</b>	<b>0.7721</b>	<b>0.9116</b>	0.9422	<b>0.5748</b>
	Run #5	0.6916	0.7710	0.9048	0.9410	0.5678

the official evaluation results that the utilization of cutting-edge models trained on huge-scale datasets (i.e. CLIP) performs better compared to our cross-modal network that is trained on a quite small but task-specific dataset. Moreover, our proposed DS similarity revision approach was shown to improve the performance.

In our future work we will aim to improve textual pre-processing, combine more text-video and text-image retrieval methods and introduce explainable AI methods in order to achieve improved results and to better understand which model components influence the most the results.

**Acknowledgements** This work was supported by the EU Horizon 2020 programme under grant agreement H2020-101021866 CRiTERIA.

## References

- [1] B. Kille, A. Lommatzsch, O. Ozgobek, M. Elahi, D.-T. Dang-Nguyen, News Images in MediaEval 2022, in: Proceedings of the MediaEval 2022 Workshop, 2023.
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, , et al., Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the 38th Int. Conf. on Machine Learning (ICML), 2021.
- [3] D. Galanopoulos, V. Mezaris, Are all combinations equal? Combining textual and visual features with multiple space learning for text-based video retrieval, in: Computer Vision – ECCVW 2022, Springer, 2022.
- [4] N. Borah, U. Baruah, Image Retrieval Using Neural Networks for Word Image Spotting—A Review, Machine Learning in Information and Communication Technology (2023) 243–268.
- [5] K. Ueki, Survey of Visual-Semantic Embedding Methods for Zero-Shot Image Retrieval, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 628–634.
- [6] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, VinVL: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.
- [7] Q. Zhang, Z. Lei, Z. Zhang, S. Z. Li, Context-aware attention network for image-text retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3536–3545.
- [8] T. Cao, N. Ngô, T. D. Le, T. Huynh, N. T. Nguyen, H. Nguyen, M. Tran, HCMUS at MediaEval 2021: Fine-tuning CLIP for Automatic News-Images Re-Matching, in: Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021, volume 3181 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.
- [9] L. Pivovarova, E. Zosa, Visual Topic Modelling for NewsImage Task at MediaEval 2021, in: Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021, volume 3181 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.
- [10] T. Wolf, L. Debut, V. Sanh, et al., Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45.
- [11] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, L. Schmidt, OpenCLIP, 2021. URL: <https://doi.org/10.5281/zenodo.7439141>.
- [12] A. Tran, A. Mathews, L. Xie, Transform and Tell: Entity-Aware News Image Captioning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.