NewsSeek-NOVA at MediaEval 2021: Context-enriched Multimodal Transformers For News Images Re-matching

Cláudio Bartolomeu, Rui Nóbrega, David Semedo
NOVA LINCS, NOVA School of Science and Technology, Lisbon, Portugal
c.bartolomeu@campus.fct.unl.pt,{rui.nobrega,df.semedo}@fct.unl.pt

ABSTRACT

In this paper, we present our participation in the NewsImages task where we address the complex challenge of connecting images to news text. We leverage transformer-based multimodal models to jointly attend to different contextual news elements when performing predictions, and transfer learning to improve the performance. Our experiments demonstrate that the models benefit from jointly attending to context-enriched samples, supporting our hypothesis. We also extracted rich insights on the principles underlying the connection between images and news text.

1 INTRODUCTION

News articles are rich multimodal pieces that aim to inform users in a concise and accurate manner. These are often composed by a title, a headline and a body of text. To better convey the topic and events being covered, journalists use images as illustrations. Providing visual elements helps the news reader visualizing the event and have a better sense of what happened [11]. Connecting news text and images is a complex endeavour as it goes beyond matching what we see in an image (visual concepts) to words. Instead, it is often explained by a combination of journalistic criteria combining authenticity, topic semantic relevance and aesthetics [9, 10].

In this paper we present our approach to the NewsImages task [6], which asks researchers to re-match news images to articles, towards devising a systematic approach that captures the intricacies of how the two modalities are connected, in a journalistic perspective.

Multimodal Transformer-based architectures [2, 8, 14] have demonstrated to be highly effective at modeling image and text semantics. These can be a) encoder-based, such as LXMERT [14], which is composed by an object relationship, language and cross-modality encoders, or b) decoder-based (hence generative) like VL-T5/BART [2], which adopts a single decoder architecture to tackle multiple visio-linguistic tasks in a generative manner. We will investigate how suited these self-attention models are to news content.

Particularly, in the developed models we followed the LXMERT architecture to exploit different ways to enrich these models context. The idea is to provide complementary views of the two modalities, and leverage the model's capability of jointly attending to different news elements when performing a prediction. Then, through transfer-learning, we were able to significantly improve the performance on the NewsImages task dataset. The results confirm the importance of providing extra context, in order to bridge the semantic gap between images and news text. Namely, our best-performing variant, which achieved an MRR@100 of 9.31%, is the one that has access to more complementary views of the news piece.

2 METHODOLOGY

The connection between news and images goes beyond visual concept matching [1, 3, 20]. In this scenario, not only the challenges of image-text matching are inherited [4], but also the underlying journalistic subjectivity that stems from trade-off aspects, such as, aesthetics or authenticity. Thus, we focused on adopting a model capable of considering multiple views of news articles, to predict if an image matches a news article. Accordingly, we hypothesize that by leveraging on self-attention models, specifically the Transformer [17], and by providing extra contextual information, we allow the model to jointly reason over multiple data views and learn the relationships between text and images directly from data.

News pieces are multimodal documents composed by title, text body (as a set of paragraphs) and images that are used throughout the news piece to illustrate specific paragraphs, providing extra context. We also find several named entities, such as persons or locations, that are crucial that define its topic and scope. The challenge is on jointly using all these information to match news to images. We observed the following challenges: a) the topic of a news article cannot always be extracted from the images, b) the news title is highly concise and lacks context (e.g. "63-year-old pedestrian succumbs to his injuries"), c) to correctly capture the news context it is important to consider the mentioned entities, as well as the news central topic, and finally, d) deal with subjectivity, evidenced by situations where multiple images could actually be used, and the pattern is dependent on the journalist preference.

Data Pre-processing and Protocol. The dataset is comprised by news articles, composed by title, text snippet (in German) and an image. Since most multimodal pre-trained models were trained in English, and assuming that we do not lose information in the translation process, we used a combination of the Google's API and OPUS-MT [15], available in HuggingFace [18], to translate texts. For each news article, we extracted entities from the Title and Text Snippet using Spacy [5]. For images, we used a Faster R-CNN [12], trained on Visual Genome [1, 7], and extract a total of 36 region embeddings per image. This will allow the model to attend individually to specific parts of an image. We split the development set (7530 samples) by using 500 samples for validation, 1000 for testing and the remaining for training.

Approach. We tackled the previously discussed challenges by adopting a multimodal transformer, LXMERT [14], and learning enriched multimodal representations of news pieces. In particular, we trained the model end-to-end, by optimizing all its loss functions except the visual question-answering one. It jointly learns internal data representations and optimizes for matching images to news texts, by scoring individual (image, news text) pairs.

Exploiting News Context. We investigated different ways to provide extra context to the model. The first baseline takes as input

MediaEval'21, December 13-15 2021, Online

news title + snippet, and the extracted image regions. Then, since entities play a major role in news, inspired by [19] which extends masked language modeling to account for coarse-grain (n-gram) information, we force the model to pay special attention to entities. Namely, we added a separate masked language modeling loss, with increased masking probability for entity tokens.

Faces-entity context. We noticed that a large portion of images contain faces of persons, and these are then mentioned in the news piece. To support this new input, we added an extra projection layer to LXMERT, mapping face features internal representations. As a result, the visual sub-network is augmented with face embeddings, such that the model will be able to jointly reason over image regions, faces, news text and entities, and eventually learn the relations between faces and entities. Faces were extracted from all images using MTCNN [21], and face recognition embeddings from FaceNet [13] were used as features.

Transfer Learning. Given the reduced size of the NewsImages task dataset, we resorted to pre-training, to improve model's representations. Namely, we performed pre-training using the NYTimes800k [16] dataset, which comprises 440k news articles (100 times bigger). Then, we fine-tuned the model using the development dataset. This allows the model to be exposed to a greater number of different news pieces, therefore improving its representations and better capture cross-modality relationships. In NYTimes800k, each article can have multiples images. Moreover, in addition to the title and news text body, each image also contains a caption. Since these captions are describing an individual image, sometimes they do not reflect the news article main topic. Thus, to have rich context, we considered the headline, a snippet and the image caption.

3 RESULTS AND DISCUSSION

In all experiments we pre-trained LXMERT using NYTimes800k - with different combinations of the article's headline, snippet and image caption - and fine-tuned it on the task's dataset (development split). For the task dataset, we fixed the language input to always use the article's title and text. Table 1 describes our runs results. We choose our best runs based on the results in our test split.

Run 1 - NT-CS + ME-TS. In this experiment we used NYTimes800k articles' snippet (S) and image caption (C) during pre-training. From the task dataset we used the articles' title (T) and snippet (S). Compared to our baseline without pre-training, we noticed an improvement of ≈ 47.1% in MRR@100, on our test split. This shows the importance of transfer-learning to improve the model performance.

Run 2 - NT-CS + ME-TSE. For this run, we used NYTimes800k articles' snippet (S) and image caption (C). From the task dataset we used the title (T), snippet (S) and named entities (E). We noticed improvements in R@50 and R@100, while worsening the other metrics, what can be due to not using entities in pre-training.

Run 3 - NT-CHS + ME-TS. In this run we wanted to assess the impact of considering the headline (H) in pre-training, together with news snippet (S) and image caption (C). In this scenario we have a better alignment between elements used in pre-training vs. fine-tuning. We observed an improvement in R@5, and a deterioration of R@10, R@50 and R@100.

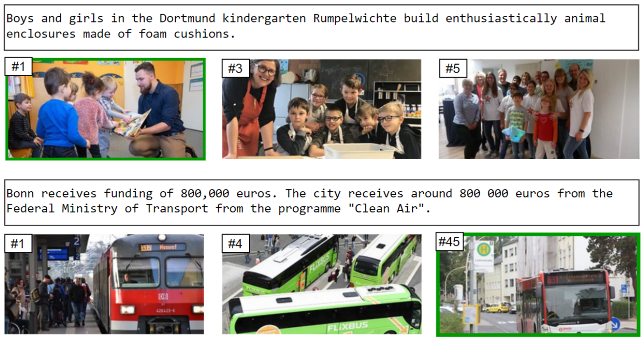
Run 4 - NT-CSEF + ME-TSEF. In this last run, we used our augmented LXMERT architecture to incorporate both face features (F) and entities (E). This experiment held the best results in MRR@100, R@10 and R@50, what corroborates with our experiments, in which it achieved the best overall results. This proves that allowing the model to jointly attend to faces and entities, better leverages the context to establish the connection between images and news text.

Run 5 - RRF. During our experiments, we observed that different configurations obtained better results at different recall thresholds (K value in R@K). Thus, in our last run, we used Reciprocal Rank Fusion to merge the first, second and forth experiments' ranks. This held the best results for all metrics.

Table 1: Runs results in MRR@100 and Recall at K (R@K).

Run	MRR	R@5	R@10	R@50	R@100
1-NT-CS + ME-TS	0.0922	0.1248	0.1927	0.4433	0.5990
2-NT-CS + ME-TSE	0.0877	0.1232	0.1922	0.4439	0.6068
3-NT-CHS + ME-TS	0.0931	0.1274	0.1906	0.4381	0.5875
4-NT-CSEF + ME-TSEF	0.0931	0.1269	0.2052	0.4611	0.6057
5-RRF (1 + 2 + 4)	**0.1043**	**0.1467**	**0.2183**	**0.4789**	**0.6277**

Figure 1: Models' predictions inspection example.


Boys and girls in the Dortmund kindergarten Rumpelwichte build enthusiastically animal enclosures made of foam cushions.

Bonn receives funding of 800,000 euros. The city receives around 800 000 euros from the Federal Ministry of Transport from the programme "Clean Air".

3.1 Models' Predictions Inspection

To understand our model decisions, we inspect in Figure 1 two sample predictions (using Run #4). In the top row, the model succeeds, and in the bottom row it ranks the correct image in position 45. Both examples illustrate the inherent subjectivity of the task, as semantically, any of the shown images seem to match the article's text. Notwithstanding, we can see that in general, the model captures the complex relations between modalities.

4 CONCLUSIONS AND FUTURE WORK

In this work we proposed a set of context-enriched variants, of a multimodal transformer model, to address the task of news rematching. These alternate between the type of context (textual and visual) provided to the model, and used to learn the connection between images and text. We confirmed that going beyond news title and a small snippet is crucial. Despite our promising results, we posit that there are essentially two key challenges that follow: a) learn how different news entities are related and how they are visually materialized, and b) dealing with inherent journalistic subjectivity when opting for a specific image.

Acknowledgments This work has been partially funded by the iFetch project, Ref. 45920, co-financed by ERDF, COMPETE 2020, NORTE 2020 and FCT under CMU Portugal, and by the FCT project NOVA LINCS Ref. (UIDB/04516/2020).

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Marina Meila and Tong Zhang (Eds.), Vol. 139. PMLR, 1931–1942. https://proceedings.mlr.press/v139/cho21a.html

[3] 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

[4] Yan Gong, Georgina Cosma, and Hui Fang. 2021. On the Limitations of Visual-Semantic Embedding Networks for Image-to-Text Information Retrieval. *Journal of Imaging* 7, 8 (2021). https://doi.org/10.3390/jimaging7080125

[5] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). https://spacy.io/

[6] Benjamin Kille, Andreas Lommatzsch, Özlem Özgöbek, Mehdi Elahi, and Duc-Tien Dang-Nguyen. News Images in MediaEval 2021. In *Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021*.

[7] 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 Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision* 123, 1 (may 2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf

[9] Gonçalo Marcelino, Ricardo Pinto, and João Magalhães. 2018. Ranking News-Quality Multimedia. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. Association for Computing Machinery, New York, NY, USA, 10–18. https://doi.org/10.1145/3206025.3206053

[10] Gonçalo Marcelino, David Semedo, André Mourão, Saverio Blasi, João Magalhães, and Marta Mrak. 2021. *Assisting News Media Editors with Cohesive Visual Storylines*. Association for Computing Machinery, New York, NY, USA, 3257–3265. https://doi.org/10.1145/3474085.3475476

[11] Nelleke Oostdijk, Hans van Halteren, Erkan Başar, and Martha Larson. 2020. The Connection between the Text and Images of News Articles: New Insights for Multimedia Analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4343–4351. https://aclanthology.org/2020.lrec-1.535

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823.

[14] Hao Tan and Mohit Bansal. 2019. LXMert: Learning cross-modality encoder representations from transformers. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (2019), 5100–5111. https://doi.org/10.18653/v1/d19-1514 arXiv:1908.07490

[15] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*. Lisbon, Portugal.

[16] Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020), 13032–13042. https://doi.org/10.1109/CVPR42600.2020.01305 arXiv:2004.08070

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017). arXiv:cs.CL/1706.03762

[18] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. 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, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[19] Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-Gram: Pre-Training with Explicitly N-Gram Masked Language Modeling for Natural Language Understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 1702–1715. https://doi.org/10.18653/v1/2021.naacl-main.136

[20] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Francis Bach and David Blei (Eds.), Vol. 37. PMLR, Lille, France, 2048–2057. https://proceedings.mlr.press/v37/xuc15.html

[21] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342