Visual Topic Modelling for NewsImage Task at MediaEval 2021

Pivovarova, Lidia

MediaEval
2021

Pivovarova , L & Zosa , E 2021 , Visual Topic Modelling for NewsImage Task at MediaEval 2021 . in Working Notes Proceedings of the MediaEval 2021 Workshop . MediaEval , The MediaEval 2021 , 13/12/2021 .

http://hdl.handle.net/10138/339048

Visual Topic Modelling for NewsImage Task at MediaEval 2021

Lidia Pivovarova, Elaine Zosa
University of Helsinki, Finland
first.last@helsinki.fi

ABSTRACT

We present the Visual Topic Model (VTM)—a model able to generate a topic distribution for an image, without using any text during inference. The model is applied to an image-text matching task at MediaEval 2021. Though results for this specific task are negative (the model works worse than a baseline), we demonstrate that VTM produces meaningful results and can be used in other applications.

1 INTRODUCTION

We present a novel approach for Visual Topic Modelling (VTM), i.e. assigning to an image a topic distribution, where 2-3 topics are the most probable ones. A topic is represented as a list of words, so an image is labeled with a set of predefined keywords.

VTM is an extension of Contextualized Topic Models (CTM) [1]. For training it requires pairs of images and texts. During inference, it takes as an input only an image. Thus, the model is capable of assigning topics to an image without any textual description.

In this paper, we apply VTM for MediaEval 2021 NewsImage Task 1, i.e. matching news articles with corresponding images [4]. Our approach consists of training two *aligned* topic models: one model takes as an input text, another takes as an input image, both produce as an output a topic distribution from the common set of topics. During training, we use aligned texts and images and train models in such a way that they have a similar output distributions. During inference, to find images corresponding to a given text, we apply visual and text models independently and then sort images by topic distribution similarity to the text topic distribution. Since each topic can be represented as a set of keywords, results are interpretable.

To train aligned visual and text topic models we use *knowledge distillation* [3], i.e. first train a *teacher* and then train a *student* model that should produce an output similar to those produced by the teacher.

Our experiments with text to image matching produced *negative results*: a solution based on VTM works worse than a baseline, based on cosine similarity between out-of-the-box text and image embeddings [5]. Nevertheless, we believe that topic modelling for images can have many other applications. It can also be possible to improve the current solution with hyperparameter tuning or by using a larger training set.

2 METHOD

VTM is an extension of CTM [1]. CTM is a family of neural topic models that is trained to take as an input, text embeddings and to produce as an output the bag-of-words reconstruction. The model trains an inference network to estimate the parameters of the topic distribution of the input. During inference this topic distribution is used as the model output to describe texts unseen during training.

Thus, to train a model, each input instance has two parts: text embeddings and bag-of-words representation (BoW). Our main contribution is that we replace text embeddings with visual embeddings and demonstrate that they can be used to train a topic model. The ZeroShot CTM model uses the BoW representation only to compute loss, i.e. this information is not needed during inference time. Since we have a training set that consists of aligned text and image pairs we can use the texts to produce the BoW representation and use it to train a model.

To obtain image embeddings we use CLIP—a pretrained model that produces text and image embeddings in the same space [5]. CLIP representations for text and image are already aligned. However, this is not a requirement for VTM: in our preliminary experiments we used ViT [2] for image and German BERT for texts (https://huggingface.co/bert-base-german-cased). The results obtained using non-aligned embeddings were only slightly worse than those with CLIP embeddings. Topic models converge to similar results because they use the same BoW to compute loss; alignment of embeddings simplifies this process but is not necessary.

This basic procedure, i.e. training image and text models independently, produces similar but not aligned topic models. Topics could be slightly different and even similar topics are organized in different (random) order. To increase similarity between text and image models we use knowledge distillation. In this approach a student model uses a different input than a teacher—e.g. image instead of text—but should produce the same result.

CTM uses a sum of two losses: reconstruction loss and divergence loss. The reconstruction loss ensures that the reconstructed BoW representation is not far from the true one. The divergence loss, measured as KL-divergence between *priors* and *posteriors*, ensures a diversity property, that is desired for any topic model: only few words have large probabilities for a given topic and only few topics have high probabilities for a given document.

In the knowledge distillation approach we leave the reconstruction loss intact but replace divergence loss with KL-divergence with regards to the *teacher output*. The assumption here is that since a teacher model is already trained to be diverse and a student model is trained to mimic the teacher, the student does not need priors. Experiments supported this assumption.

We use knowledge distillation in two versions: *joint model* and *text-teacher*. In the joint approach we first train a joint model that takes as an input a concatenation of text and image embeddings, then train two student models for image and text separately. In the second approach, we first train a text model and then train an image model as a student.

MediaEval'21, December 13-15 2021, Online

Table 1: Results

Model	Correct in Top100	MRR@100	Recall@5	Recall@10	Recall@50	Recall@100
baseline (CLIP)	1225	0.169	0.22	0.30	0.53	0.64
joint 120 topics	767	0.043	0.06	0.09	0.26	0.40
joint 60 topics	698	0.030	0.04	0.07	0.24	0.36
text teacher 120 topics	816	0.042	0.05	0.09	0.30	0.43
text teacher 60 topics	757	0.037	0.05	0.08	0.26	0.39

(a) CLIP 1st

(b) VTM 1st

(c) CLIP 2nd

(d) VTM 2nd

Figure 1: Images, most close to the story about the trial of Anna Semenova according to the baseline (a,c) and VTM (b,d) models.

We try 60 and 120 topics with both joint and text-teacher approaches. Preliminary experiments showed that the more topics are used the higher is the model performance in text-image matching.

As a baseline, we use raw cosine similarities between CLIP embeddings, without any domain adaptation for the text. We use an implementation provided as a part of Sentence Bert package (https://www.sbert.net/examples/applications/image-search/README.html).

3 RESULTS

The results are presented in Table 1. As can be seen from the table, the best results are obtained with CLIP embeddings, that are used without any fine-tuning to the training set. They are able to find the correct image in 1225 cases out of 1915 and has a Mean Reciprocal Rank (MRR) of 0.17. The best VTM model finds correct image in 816 cases out of 1915 and yields an MRR of 0.03.

These results to some extent correspond to our previous experiments, where we showed that topic modelling does not work well for document linking [6]. The probable explanation for that might be that topic modelling produces a sparse representation of the data. While CLIP embeddings are continuous vectors and could represent an almost infinite amount of information, in topic modelling dimensions are not independent due to the diversity requirement described above. It can be seen from Table 1 that models that have more topics yield better performance.

Another interesting observation is that models that use the text model as a teacher for a visual model work better than joint models. This is an unexpected result, since one would expect that a model that has access to full information could serve as a better teacher. It is possible that text bears less noise: a text model uses the same text for contextual and BoW representation, while an image could be completely irrelevant to a corresponding article.

The fact that embeddings and topic modelling work on different principles is illustrated in Figure 1, where we reproduce images found by the model for the text about the Anna Semenova trial. CLIP model finds photos of Anna Semenova, probably due to the huge text and image base used to train the embeddings. VTM returns images with a statue of Themis, a personification of Justice, which represent the text *topic* rather than specific facts. Though according to our results, CLIP embeddings outperform VTM, the ability to illustrate text topic might be a desirable property for some applications, as well as topic interpretability.

Our code is available at https://github.com/lmphcs/media_eval_vctm.

ACKNOWLEDGMENTS

This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

REFERENCES

[1] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1676–1683. https://www.aclweb.org/anthology/2021.eacl-main.143

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and others. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[3] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

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

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. *Learning Transferable Visual Models From Natural Language Supervision*. Technical Report.

[6] Elaine Zosa, Mark Granroth-Wilding, and Lidia Pivovarova. A Comparison of Unsupervised Methods for Ad hoc Cross-Lingual Document Retrieval. In *LREC 2020 Language Resources and Evaluation Conference 11–16 May 2020*. 32.