Multimodal Heterogeneous Transfer Learning for Multilingual Image-Text Classification

Andrea Pedrotti[1,*], Alejandro Moreo[1] and Fabrizio Sebastiani[1]

[1]Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche, via G. Moruzzi 1, 56124 Pisa, Italia

Abstract

The Multilingual Image-Text Classification (MITC) task is a specific instance of the Image-Text Classification (ITC) task, where each item to be classified consists of a visual representation *and* a textual description written in one of several possible languages. In this paper we propose MM-gFun, an extension of the gFun learning architecture originally developed for cross-lingual text classification. We extend its original text-only implementation to handle perceptual modalities.

Keywords

Heterogeneous Transfer Learning, Multilingual Classification, Multimodal Classification,

1. Introduction

Transfer learning involves using a labeled training set $\text{Tr}_{\mathcal{S}}^{L}$ from a *source* domain $\mathcal{S}$ to make predictions for unlabeled data in a *target* domain $\mathcal{T}$, which is related but different from $\mathcal{S}$. The goal is to leverage information from the source domain to improve performance in the target domain. Transfer learning can be categorized either as *homogeneous* (when the feature spaces are the same, also known as *domain adaptation* in NLP) or *heterogeneous* (when the feature spaces are different and non-overlapping). According to the broad framework of Heterogeneous Transfer Learning (HTL), heterogeneous different languages and different perceptual modalities can be regarded as non-overlapping feature spaces describing the same object, with each one complementing the other.

A specific task dealing with both perceptual inputs such as images and multilingual texts is Multilingual Image-Text Classification (MITC). In Image-Text Classification (ITC), an item consisting of both a textual description and a visual representation (e.g., an image) must be assigned a binary vector of length $|\mathscr{C}|$, where each element indicates whether the corresponding class is a correct label for the given item. Furthermore, in the MITC scenario, items come with textual descriptions written in one out of a finite set of languages. MITC tasks can be solved by independently relying on each modality (e.g. only textual or only visual), however it is reasonable to assume this choice to be sub-optimal: in each scenario, useful information is discarded by ignoring the other modality. Moreover, textual descriptions available in different languages may contain crucial information not encoded in other ones.

In this work, we approach the task of MITC through the lens of HTL, where the features spaces involved are those of the textual and the visual inputs. The textual modality is further subdivided into non-overlapping feature spaces, each characterized by a specific language. Under this formulation we develop MM-gFun, an extension of gFun integrating multilingual data with the heterogeneous modality of textual data.

Discovery Science - Late Breaking Contributions 2024

*Corresponding author.

andrea.pedrotti@isti.cnr.it (A. Pedrotti); alejandro.moreo@isti.cnr.it (A. Moreo); fabrizio.sebastiani@isti.cnr.it (F. Sebastiani)

0000-0002-2322-7043 (A. Pedrotti); 0000-0002-0377-1025 (A. Moreo); 0000-0003-4221-6427 (F. Sebastiani)

2. Extending gFun to the Visual Domain

Generalized Funnelling (gFun) [1] is an architecture for HTL especially developed for Cross-Lingual Text Classification (CLTC) tasks, in which the model is trained to assign labels to documents written in one of a finite set of languages according to a shared classification scheme. The model is a two-tier architecture in which the first layer projects documents written in different languages to the same shared space where a metaclassifier (i.e., the second layer) subsequently takes care of the classification step. In this way, the metaclassifier can be trained on the whole collection of training documents, without being forced to operate on a language-specific subset of the data. Specifically, the first-tier is equipped with a set of View Generating Functions (VGFs) each designed to mine different types of information encoded in heterogeneous views of the same object.

VGFs are language-dependent functions that map (monolingual) documents into language-independent vectorial representations (i.e., views) aligned across languages. Since each view is aligned across languages, they can be aggregated into a single representation that is also aligned across languages, and which can be thus fed to the meta-classifier. In gFun, for each language a set of VGFs is instantiated to leverage different types of information that can be brought to bear on the training process. In [1], unimodal VGFs are based on posterior probabilities, MUSE embeddings, WCE embeddings, and multilingual BERT.

While the original gFun was developed for a unimodal setting (i.e., involving only textual data), in this article we extend it to be deployed in a multimodal setting involving images as well as text written in different languages. This is achieved by augmenting the set of VGFs originally proposed with a Visual VGFs. This module is designed to explicitly mine the correlations between the images and the target classes, which remained out of reach to the other VGFs.

The training of MM-gFun is a two-phase training. In the first phase, all the VGFs of the first-tier are tuned to produce the posterior probabilities representations of images and texts. In the second-phase, output representations of the first-tier are aggregated according to a pre-defined policy and subsequently used to train the meta-classifier. By leveraging the shared (modality-agnostic) space of the posterior probabilities, the meta-classifier can be trained on the entire collection of items, be they images or text, and independently of the textual description language.

To encode the image-class correlations, we derive image representations from a Vision Transformer (ViT). o deal with the continuous input of images, the ViT divides the image into fixed-size patches. The patches are then flattened into vectors and linearly projected to a higher-dimensional space, serving as input *visual tokens* for the transformer. Similarly to transformer model for texts, positional embeddings are added to the visual tokens to account for the relative positions of patches.

3. Experiments

Dataset We base our experiments on GLAMI-1M [2], a publicly available multilingual image-text classification dataset. [1] and the largest multilingual image-text classification dataset providing standardized class labels. , which makes it suitable for MITC tasks. It is a collection of fashion products, gathered from an online catalog of fashion goods and accessories. Each item is represented by an image and a textual description. Textual descriptions are written in one of 13 different languages. Each item is labelled with a category created from the GLAMI category tree, resulting in a codeframe of 191 classes. Images are provided with a standard resolution of 228 by 298 pixels. The class distribution exhibits a long tail, with a few frequent classes and many infrequent ones. Similarly, the language distribution is skewed towards a few dominant languages. The dataset is already partitioned into a training set of 1,000,000 items and a test set of 116,004 items.

From the original GLAMI-1M, we extract a smaller version with a controlled distribution of languages, both for the training and in test sets. For each language, we sample 15,000 documents in order to represent higher-resourced (or "source", in the context of HTL) languages ($\mathscr{L}_{src}$ = {*Bulgarian (bg), Czech*

[1] https://github.com/glami/glami-1m

(cz), Hungarian (hu), Lithuanian (lt), Slovak (sk), Turkish (tr)}), and we sample 5,000 documents in order to represent lower-resourced languages ($\mathscr{L}_{trg}$ = {*Estonian (ee), Spanish (es), Greek (gr), Croatian (hr), Latvian (lv), Romanian (ro), Slovenian (sl)*}). Note that for for Greek, Spanish, Romanian, and Latvian languages there are fewer training examples (respectively: 3405, 2434, 3533, and 4184); in these cases we simply take them all. The final reduced dataset consists of 100,000 training items and 58,553 test items. The class distribution in the reduced version closely resembles the one of the original dataset after the sampling. This reduced version facilitates a controlled analysis of the ability to transfer knowledge from higher-resourced to lower-resourced languages.

Training Details We initialize the textual VGFs of MM-GFUN from pre-trained models mDeBERTa-v3. The textual input is processed via the relative SentencePiece tokenizer, with a maximum sequence length of 32 tokens to ease the comparison with results obtained in [2]. For the visual VGF, we initialize the component from CLIP visual encoder. Visual images are center cropped to square features of 224 pixels. All the learners in the first-tier and the metaclassifier are optimized to minimize the cross entropy loss, with AdamW and cosine annealing learning rate scheduler. We set the learning rate to 0.0001 in all our experiments, and train all of the models for a maximum of 25 epochs, with a early-stopping set to 5 epochs without any increase on the accuracy on the validation set.

Results We report the results that we have obtained using the original GLAMI-1M dataset in its entirety as well as those obtained in the reduced setting, where we balance the number of documents across the different languages. In order to allow for a direct comparison with [2], we employ accuracy as our evaluation measure, as well as accuracy at top 5. Table 1 reports the results we have obtained on the test set (denoted by the "Full-" prefix). Results are grouped according to the input modalities that the methods can process. In the first group, we report unimodal models for the visual modality; in the second one, unimodal models for the (multilingual) textual modality; in the third one, we group all of the multimodal models.

Table 1
The results for EmbraceNet and ResNeXt are taken from [2]. The best results for the full dataset are highlighted in bold, while the best results for the few-shot setting are highlighted in gray.

Model	Text	Image	Full-*Acc*@1	Full-*Acc*@5	Few-Acc_{μ}	Few-Acc_{M}
ResNeXt-50	✗	✓	63.10	93.50	58.95	57.94
CLIP-ViT	✗	✓	71.25	95.79	N/A	N/A
XLM-roBERTa	✓	✗	83.44	**97.87**	72.35	71.50
mDeBERTa	✓	✗	83.17	97.83	72.40	71.70
AltCLIP [3]	✓	✓	70.55	95.53	48.61	47.68
CLIP-M [4]	✓	✓	68.77	95.75	47.71	47.35
LLaVA-NeXT [5]	✓	✓	N/A	N/A	46.55	72.13
EmbraceNet [6]	✓	✗	59.30	84.00	N/A	N/A
	✗	✓	68.50	94.80	N/A	N/A
	✓	✓	69.70	94.00	N/A	N/A
MM-GFUN	✓	✗	83.42	97.61	71.43	72.26
	✗	✓	51.12	83.65	47.25	48.35
	✓	✓	**83.68**	97.73	75.05	74.26

The results indicate that multilingual textual information encodes a stronger signal for the classification task. Indeed, multilingual models achieve strong performance, with top-1 and top-5 accuracy around 83%, significantly outperforming vision-only models, which achieve top-1 of 63 and 71%, respectively. Multimodal baselines such as AltCLIP, m-CLIP, and EmbraceNet show similar performance, but they do not measure up in terms of accuracy when compared with the text-based multilingual models. The proposed method MM-GFUN achieves the best results in terms of top-1 accuracy when leveraging both textual and visual inputs. This, along with the fact that MM-GFUN improves upon its individual components (CLIP-ViT and mDeBERTa), demonstrates the ability of the method of leveraging both

modalities. Consistently with our results on the full dataset, we observe that multilingual models outperform visual ones also in the few-shot setting (denoted by the "Few-" prefix), in terms of micro- and macro-averaged accuracy across languages. However, in this data-scarce scenario, all our multimodal baselines fall short in terms of performance when compared to MM-GFUN and CLIP-ViT, by also exhibiting a significant drop in accuracy with respect to the data-rich scenario with respect to the other approaches. Here, MM-GFUN achieves good results, improving over the unimodal textual-component by no less than 2.5 points and by 16 points with respect to the visual component.

4. Conclusion

In this paper, we have presented MM-GFUN, an architecture for Heterogeneous Transfer Learning (HTL) in the task of multilingual image-text classification. This architecture is an extension of the unimodal multilingual GFUN. By augmenting the set of Visual Grounding Functions (VGFs) with a specific module to represent input images, we demonstrated improved performance on multilingual image-text classification tasks. Our approach achieves superior performance compared to MM-GFUN's internal components when trained independently. We validated our hypothesis through experiments on GLAMI-1M, a multilingual and multimodal dataset of fashion product images with textual descriptions in 13 different languages.

Acknowledgments

This work has been supported by the FAIR and SoBigData.it projects, funded by the Italian Ministry of University and Research under the NextGenerationEU program. The authors' opinions do not necessarily reflect those of the funding agencies.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check.

References

[1] A. Moreo, A. Pedrotti, F. Sebastiani, Generalized funnelling: Ensemble learning and heterogeneous document embeddings for cross-lingual text classification, ACM Transactions on Information Systems 41 (2023) 36:1–36:37. doi:10.1145/3544104.

[2] V. Kosar, A. Hoskovec, M. Šulc, R. Bartyzal, Glami-1m: A multilingual image-text fashion dataset, in: British Machine Vision Conference, 2022.

[3] Z. Chen, G. Liu, B. Zhang, Q. Yang, L. Wu, Altclip: Altering the language encoder in CLIP for extended language capabilities, in: Findings of the Association for Computational Linguistics: ACL, Association for Computational Linguistics, 2023, pp. 8666–8682. doi:10.18653/V1/2023.FINDINGS-ACL.552.

[4] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, 2020, pp. 4512–4525. doi:10.18653/V1/2020.EMNLP-MAIN.365.

[5] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.

[6] J. Choi, J. Lee, Embracenet: A robust deep learning architecture for multimodal classification, Inf. Fusion 51 (2019) 259–270. doi:10.1016/J.INFFUS.2019.02.010.