Language-Agnostic Visual-Semantic Embeddings

Jônatas Wehrmann[1] **and Rodrigo C. Barros**[1]

[1] Escola Politécnica
Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681 - Partenon, Porto Alegre - RS, 90619-900

jonatas.wehrmann@pucrs.edu.br, rodrigo.barros@pucrs.br

Abstract. *We propose a framework for training language-invariant cross-modal retrieval models. We introduce four novel text encoding approaches, as well as a character-based word-embedding approach, allowing the model to project similar words across languages into the same word-embedding space. In addition, by performing cross-modal retrieval at the character level, the storage requirements for a text encoder decrease substantially, allowing for lighter and more scalable retrieval architectures. The proposed language-invariant textual encoder based on characters is virtually unaffected in terms of storage requirements when novel languages are added to the system. Contributions include new methods for building character-level-based word-embeddings, an improved loss function, and a novel cross-language alignment module that not only makes the architecture language-invariant, but also presents better predictive performance. Moreover, we introduce a module called* ADAPT*, which is responsible for providing query-aware visual representations that generate large improvements in terms of recall for four widely-used large-scale image-text datasets. We show that our models outperform the current state-of-the-art all scenarios. This thesis can serve as a new path on retrieval research, now allowing for the effective use of captions in multiple-language scenarios.*

1. Introduction

Neural Networks are a particular kind of Machine Learning (ML) approach that excel in learning representations from raw data. They have shown remarkable results specially for feature extraction from images and text when trained in very large datasets. The network learns to extract relevant features according to the provided data to solve the task at hand. They can be seen as fully-differentiable end-to-end computational graphs, allowing for data-driven training of complete models, thus automatically learning to preprocess, extract features, and provide predictions using the learned functions on unseen data. With neural networks, one can learn multimodal models that can be used in many tasks, such as Image Captioning, VQA, Text-to-Image Generation, Visually-grounded Translation, and Image Search via textual queries.

A *multimodal model* is a broad term *per se*, denoting roughly any model trained over more than a single modality (e.g., images, videos, text, and audio). However, in my thesis we have defined such a term as a model specifically trained for aligning images and textual descriptions. The task of Image-Text alignment is also referred to in the literature as Multimodal Retrieval, Cross-Modal Retrieval, Image-Text Matching, and Bidirectional Alignment. Those models are mainly used to handle two tasks: (i) Image

Retrieval *via* textual queries; and (ii) Image Annotation, which consists in finding proper textual descriptions for a given image.

A typical framework for training Image-Text Alignment models is the use of neural networks to extract high-level features of both images and captions. Those features are then projected onto the same shared space, the so-called multimodal embedding space (or also visual-semantic embedding space). A pairwise loss function is employed to approximate similar image-text pairs, while making uncorrelated ones to be far from each other in that same space. There has been an extensive amount of research dedicated in exploring several aspects of those models, providing novelties on the image encoding function, similarity computation, loss function, and text encoding. Even though the research area of Multimodal Retrieval did attract great attention from the scientific community in recent years, most of the work has focused on training single-language models, which makes the application of those models only feasible for the English language.

For training both image/caption encoding neural networks, one needs a large set of labeled images with textual descriptions. For that matter, a major issue when training retrieval models is the scarcity of high-quality labeled data. There are very few image-text datasets that are labeled in languages other than English. Therefore, the research question we answer is: "*how to develop a language-agnostic multimodal retrieval system without necessarily requiring labeled image-text datasets for all languages?*".

2. Proposed Approaches

We provide a cross-language training framework that allows text encoders to present language-invariant behavior. Such an approach, namely CLMR (depicted in Figure 1), is designed so one can leverage a large multimodal dataset, such as MS COCO, and adapt the learned representations across other languages using a separated aligned cross-lingual corpus. CLMR works by projecting images, captions, and cross-language sentences into the same embedding space, which makes it possible to retrieve and annotate images in distinct languages using the same model. A key aspect on the proposed architecture is the use of the same text encoder to process all the textual information that is provided, while enforcing similar-content cross-lingual sentences to preset high similarity values in the learned embedding space. Note that the proposed approach does not require the use of machine translation at any level. Machine translation models can be used to achieve language-agnostic behavior, though state-of-the-art models are quite heavy, slow, and complex, while off-the-shelf solutions often come with the extra burden of having to process large datasets.

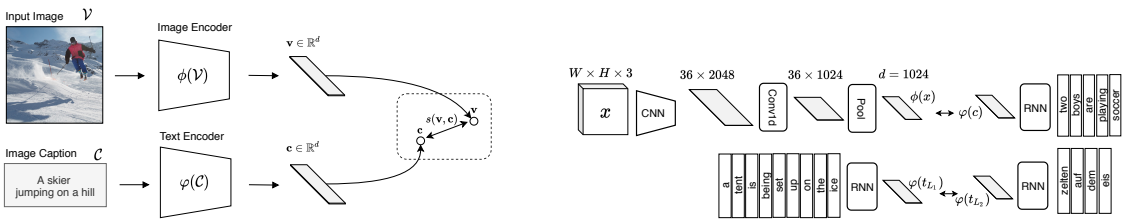

Figure 1. Overview of the language-agnostic training framework.

One of or hypothesis was that character-level convolutional networks would be capable of learning multi-lingual representations, while being able to learn proper pre-

processing steps and present a behavior of robustness to noise. In addition, considering that character-level inputs present a rather reduced semantic gap when compared to raw image pixels, we believed that shallow networks might suffice to learn proper textual representations for queries and short sentences (*tweets*, for instance). We validated those assumptions by introducing CharCNN, CHAIN-VSE, and LIWE, which are character-level approaches trained without any preprocessing step. We showed that they can be used as fast and efficient text encoders, and we evaluated their performance for both multimodal retrieval and text classification.

The proposed cross-lingual framework, CLMR, can take roughly any image and text encoding approach, as well as a diverse set of similarity functions, while being very stable and easy-to-train even in cross-language cases (qualitative results are shown in Figure 3). The core component that makes it all possible is our novel loss function, namely the Exponentially Weighted Ranking Loss. It leverages stable gradient signals averaged from all contrastive samples in a mini-batch during the early stages of training, while exponentially increasing the weight for the hard-contrastive pairs as soon as the model learns more robust multimodal representations. Such a loss function provides a large impact on retrieval performance when compared to the standard approach from [Kiros et al. 2014], and also allows for training light and efficient character-level models. Indeed, the hard-contrastive-based loss function [Faghri et al. 2017] does not provide enough gradient signal when training with character-level text encoders and with cross-language data. Hence, the proposed approach presents itself as the best of both worlds, with the stability seen in [Kiros et al. 2014] while outperforming the results found in [Faghri et al. 2017], and thus having the potential to benefit virtually any retrieval model based on the training of a common latent space.

Figure 2. Image Retrieval examples with a language-agnostic model (LIWE).

We also provided contributions regarding the image encoding and similarity functions. Most notably, we have investigated the fact that the use of more coherent multimodal representations can be helpful as the model becomes capable of learning fine-grained correlations. For instance, attention can use visual features to compute textual features and *vice-versa* with the aid of co-attention layers, which helped to achieve better predictive performance. Nonetheless, such an upgrade comes with a cost: it is much slower during both training and inference times due to the need of computing distinct features for each image-text pair. Hence, we introduced ADAPT, which is a method that improves the embedded representation of instances from modality a based on the global information of modality b. ADAPT is designed to modify intermediate features (word-level or region-wise projections) by using parameters predicted by the vector representation from the other modality. Such a feature adaptation procedure works as a filtering strategy. For instance, we can use visual-based features in order to filter the most impor-

tant hidden-state dimensions of captions to build a better textual embedding. We show that such an approach, despite being quite faster during both training and test times, is capable of outperforming attention-based strategies, specially in the Image Retrieval task.

We provided extensive experimental analysis regarding each major choice of the proposed neural architectures. In those ablation studies, we demonstrated the importance of particular modules used in our models, and also insights behind their functionality. Moreover, we presented several qualitative examples that clearly show the ability of the proposed models to retrieve and annotate images across distinct languages. Finally, we also clearly demonstrated the existence of semantic regularities (see Figure 3) that can be found in the learned multimodal embedding space. Such results provided strong evidence that the learned space does present semantic structural organization in which simple vector arithmetics is capable of achieving semantically-relevant retrieval results.

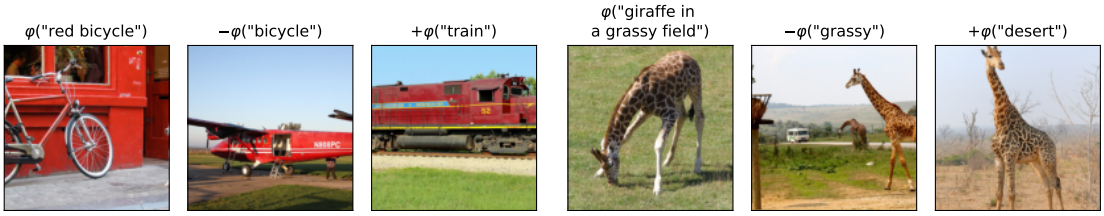

Figure 3. Example of embedding-space regularities found in our models.

3. Summary of Contributions

My thesis generated more than 20 full papers accepted for publication. Several of them were published in world-class top-tier Machine Learning, Artificial Intelligence, and Computer Vision venues, such as ICML, CVPR, ICCV, and AAAI, which are the most prestigious and important conferences in their respective areas. In the following paragraphs, we present a list of the main contributions of the thesis, also indicating a list of selected papers. I invite the reviewer to check the full paper list in my Google Scholar page[1], since I could not show it here due to space constraints.

To the best of our knowledge, we have proposed the first character-level neural network architecture for language-agnostic sentiment analysis and multimodal retrieval, namely CharCNN. We have presented this novelty in several papers [Wehrmann et al. 2017, Becker et al. 2017, Wehrmann et al. 2018c], which explore different aspects of the main architecture. We have demonstrated that it can outperform state-of-the-art approaches, though requiring up to $1,000\times$ fewer parameters than LSTMs and GRUs. We have also demonstrated that it can learn multi-task functions to perform language-detection and sentiment analysis altogether.

We have also proposed a novel inception-inspired text encoder named CHAIN-VSE for efficient multimodal retrieval [Wehrmann and Barros 2018]. That work was accepted in the CVPR 2018 main conference, which is the conference with highest H-index in computer science as of today. In fact, CVPR currently is the fifth scientific venue in terms of H-index considering all knowledge areas. In that work, we have shown that ar-

[1] `https://scholar.google.com.br/citations?user=cUOwsGMAAAAJ&hl=pt-BR`

chitectures like CHAIN-VSE can automatically learn preprocessing steps and outperform LSTM networks while being much more robust to scenarios with noisy textual queries.

Our third contribution is a novel character-level-based word-embedding generation approach, namely LIWE. It has the ability to project correlated words across languages into vectors that are close to each other in the semantic space. In our ICCV 2019 paper [Wehrmann et al. 2019], we have demonstrated that LIWE presents superior language-agnostic multimodal retrieval performance than state-of-the-art approaches that rely on word-embedding vectors (for details see Chapters 3 and 6 of the thesis). We also provide an overall language-agnostic framework for training retrieval models, alongside a new loss function, the Exponentially Weighted Pairwise Loss Function (denoted $\mathcal{J}_{WE}$). We provided in the thesis an extensive set of experiments to validate our assumptions on the proposed loss function, making it clear that it does benefit training by stabilizing gradients and allowing the generation of more accurate models. The novel loss function allows improving up to 20% recall values when compared to the previous state-of-the-art. It also allows training models when the default loss function diverges. Moreover, for running our experiments we have generated novel training, validation, and test splits for the YJCaptions dataset, given that it did not have default splits for retrieval. For more details, the reviewer is invited to read Chapter 4 of the thesis.

Our fourth contribution was the introduction of an efficient embedding adaptation module (ADAPT) for generating query-aware image representations [Wehrmann et al. 2020]. ADAPT has been published in the leading Artificial Intelligence conference, namely AAAI 2020. We analyze the impact of ADAPT on retrieval models, and show that they provide a large improvement over state-of-the-art cross-attention-based architectures. Those results are thoroughly analyzed in the thesis, generating important insights regarding the workings of ADAPT. Moreover, we provide a method for visualizing focal-points that are given more weight by ADAPT for the generation of the final image representation (Chapter 5 of the thesis).

Finally, the following are contributions that were proposed on side projects that are somewhat related to my thesis. We have discovered that a hierarchical penalty applied in multimodal retrieval can benefit Hierarchical Multi-Label Classification, and that approach was published in ICML 2018, the leading conference on Machine Learning research. We have proposed self-attentive text encoders for very fast multimodal retrieval [Wehrmann et al. 2018b] (in WACV, Qualis A1). We have observed that self-attentive encoders also work for multi-label classification of movie synopses [Wehrmann et al. 2018a], which was selected as one of the top-3 best student papers at FLAIRS 2018. We have contributed for better understanding the importance of each design component within VQA models [Kolling et al. 2020]. We have also demonstrated that a sentence interpolation strategy, loosely inspired by the *fovea* module from ADAPT, does benefit text-to-image synthesis models [Souza et al. 2020].

3.1. Impact

The impact of this work is far beyond the area of image-text retrieval alone. It brings solutions to a plethora of tasks and areas:

1. Tasks related to text encoding can benefit from the proposed architectures and representation strategies. That includes text classification, summarization, translation

and others;

2. Models that leverage from attention mechanisms could potentially use the proposed ADAPT technique for faster computation of attention maps. Those tasks include image captioning, text-to-image-synthesis, VQA, etc;
3. The proposed loss function has the potential to benefit virtually *any* content-based retrieval system: document retrieval, image retrieval, sentence-to-text matching, code search, matching medical reports to exams, fashion retrieval, etc;
4. Democratization of deep learning research: by using the proposed language-agnostic framework, researchers can train models in any language leveraging from large datasets annotated in English.

Such impact is visible through the ≈ 500 citations that the papers generated by this thesis received in the past couple of years. Some of those are high-quality studies that employ our methods to solve many real-world problems (Covid-19 detection in X-Ray images, protein function recognition), that expand our ideas (extreme compression of neural networks), and that also use them as benchmark approaches (several studies on language-agnostic learning have now used our results as the state-of-the-art baseline).

References

Becker, W., Wehrmann, J., Cagnini, H. E. L., and Barros, R. C. (2017). An efficient deep neural architecture for multilingual sentiment analysis in twitter. In *FLAIRS*.

Faghri, F., Fleet, D. J., Kiros, J. R., and Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, pages 1–13.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Multimodal neural language models. In *Proceedings of the International Conference on Machine Learning*, pages 595–603.

Kolling, C., Wehrmann, J., and Barros, R. C. (2020). Component analysis for visual question answering architectures. In *IJCNN*, pages 1–8.

Souza, D. M., Wehrmann, J., and Ruiz, D. D. (2020). Efficient neural architecture for text-to-image synthesis. In *IJCNN*, pages 1–8.

Wehrmann, J. and Barros, R. C. (2018). Bidirectional retrieval made simple. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 7718–7726.

Wehrmann, J., Becker, W., Cagnini, H. E., and Barros, R. C. (2017). A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In *IJCNN*, pages 2384–2391. IEEE.

Wehrmann, J., Kolling, C., and C Barros, R. (2020). Adaptive cross-modal embeddings for image-text alignment. In *AAAI*, volume 34, pages 12313–12320.

Wehrmann, J., Lopes, M. A., and Barros, R. C. (2018a). Self-attention for synopsis-based multi-label movie genre classification. In *FLAIRS*, pages 236–242.

Wehrmann, J., Lopes, M. A., More, M. D., and Barros, R. C. (2018b). Fast self-attentive multimodal retrieval. In *WACV*, pages 1871–1878.

Wehrmann, J., Mattjie, A., and Barros, R. C. (2018c). Order embeddings and character-level convolutions for multimodal alignment. *PRL*, 102:15–22.

Wehrmann, J., Souza, D. M., Lopes, M. A., and Barros, R. C. (2019). Language-agnostic visual-semantic embeddings. In *ICCV*, pages 5804–5813.