# Visually Grounded Cross-Lingual Transfer Learning

**Fangyu Liu**
University of Waterloo
fangyu.liu@uwaterloo.ca

**Rémi Lebret**
EPFL
remi.lebret@epfl.ch

**Karl Aberer**
EPFL
karl.aberer@epfl.ch

**Abstract.** We explore transfer learning in a multilingual setting under the context of bidirectional text-image retrieval and *Visual Semantic Embeddings* (VSE). VSE bridges natural language and vision by jointly optimizing and aligning semantic embeddings (from texts) and visual embeddings (from images). While distributional semantics exist in all natural languages, current VSE models only address the monolingual scenario. We study multilingual VSE. As vision is universal, multilingual texts would be grounded by consistent visual signals extracted from images. We introduce two basic recipes for training multilingual model and transporting knowledge across languages. On top of the two recipes, we propose a language space transformation layer embedded inside neural networks, addressing transfer learning under continuous word embeddings. We then further enhance the recipes from the perspective of Multitask Learning (MTL), offering insights for multilingual training.

**Introduction.** The semantic space of human knowledge is formed through interaction with a rich environment and grounded by concrete, real-world human senses. Among them there is language, which is a revelation and representation of the reality and whose coding is based on reality. Likewise, visual representation functions in the same way. There widely exists neat and orderly mappings between concepts in language and concepts in vision. Based on this observation, we use vision as the shared modality to ground concepts across languages. As this work researches on how Cross-Lingual Transfer Learning (CLTL) works as model interacts with visual modality (specifically, digital images), we call it Visually Grounded Cross-Lingual Transfer Learning (VGCLTL).

Specifically, we follow the line of work named VSE who deals with the task of text-image retrieval (Frome et al., 2013; Kiros et al., 2015; Vendrov et al., 2016; Wang et al., 2018; Faghri et al., 2018). We jointly model both language and vision with a two-branch framework where one branch is an encoder for image and the other for text. Image and text are both encoded into vector representations to be compared with some similarity metric in a joint space. The model aims to retrieve the best matching texts for an image or vice versa. We adopt a triplet ranking loss to cluster positive pairs and push negative pairs away from each other. While conducting the cross-modal matching task, we try to capture the language-invariant knowledge grounded by visual proof and transport them among languages.

**Two Recipes.** We validate the transfer learning idea with two basic configurations:

(a) *one for all*: building a language-agnostic model. In experiments, model trained on both language `A` and `B` outperforms models trained on monolingual data in both language `A` and `B` text-image retrieval;

(b) *all for one*: building a language-deterministic model but still benefiting from multilingual data. Starting with model weights trained on language `A`, better retrieval performance on language `B` is achieved by substituting only the word embeddings and finetuning.

Both *one for all* and *all for one* are vague and general frameworks which could be further filled with more concrete technical details. We then focus on (1) enabling the two recipes to work with continuous word embeddings; (2) investigating various learning schemes in MTL as multilingual learning fits into settings in MTL well.

**Transfer Learning Under Continuous Word Embeddings.** Multilingual Word Embedding (MWE) is a prerequisite for multilingual NLP. All popular ready-to-use aligned MWEs today treat each word as a whole and assigns distinct vectors to them (Smith et al., 2017; Lample et al., 2018). They ignore morphological information and do not handle out-of-vocabulary words, which could be essential for both languages with rich vocabulary and real-world (open-vocabulary) applications. Though there has been many efforts encoding morpheme into word vectors (Lazaridou et al., 2013; Luong et al., 2013; Cotterell and Schütze, 2015; Sennrich et al., 2016; Bojanowski et al., 2017), they are however monolingual and unaligned. We thus introduce

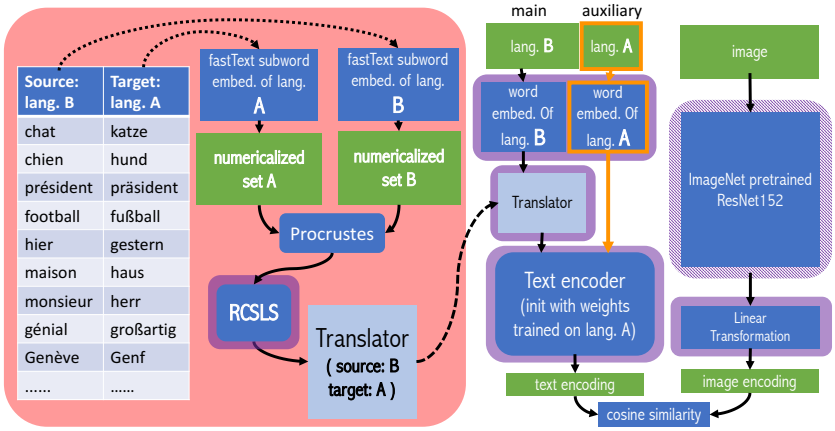


Figure 1: *Translator* is a trainable linear transformation embedded between word embeddings and text encoder. Its initial weights is produced following procedures on LHS of the figure. On RHS, though our task of interest is training a text-image retrieval model for language `B`, we add an auxiliary task (ie. also retrieve images for language `A`) as regularization.

a *Translator* layer to get the best of both worlds (multilinguality & morphology). It is a linear layer, embeded between word embeddings and text encoder, whose weights are obtained from the task of Bilingual Lexicon Induction (BLI). It enables online multilingual continuous word vector computations.

Explicitly, we first perform a Proscrustes Analysis between two sets of bilingual word vectors to obtain a transformation $T$. Then we refine $T$ with a word retrieval criterion (named RCSLS) introduced by (Joulin et al., 2018; Lample et al., 2018). During the refinement, $T$ incorporates the prior that the graph of word matching ought to be locally bipartite, addressing potential hubness problem (Dinu et al., 2014). By using the refined matrix weights to initialize a linear layer embedded between word embeddings and text encoders, the two recipes introduced above could work with continuous word embeddings and compute universal word representations on-the-fly. The general pipeline is visualized in Figure 1.

We extensively examined *Translator* in both high-resource and low-resource conditions, finding it very helpful under all conditions. We also observed that even when *Translator*s are randomly initialized, by freezing other parts of the model (only train *Translator*) for a few epochs and then finetuning the whole model, with certain amounts of data, the model automatically learns a comparable transformation to perform equivalently as *Translator*s initialized with BLI weights.

**Multilingual as Multitask.** At last, we borrow ideas from MTL literature to improve multilingual training. We (a) use a self-paced active sampling strategy to ease the gaps of different levels of learning difficulty in different languages; (b) train the model with an auxiliary language as a regularization in transfer learning to avoid overfitting.

(a) Some tasks are intrinsically harder than others. In MTL, it is common that some tasks are lagged behind and the joint optimization fails for different tasks are in very different stages of learning. This is also true for multilingual learning. For $i$-th language, we use $\frac{e^{-r_i/\tau}}{\sum_{c=1}^{k} e^{-r_c/\tau}}$, where $\tau$ is a preset scalar (we used $\tau = 0.1$); $r_i$ is Recall@$K$ from last epoch (we used $K = 10$), to compute portion of training data it gets in the next epoch. We find our active sampling strategy encourages faster convergence while maintaining model performance.

(b) In *all for one*, it's not necessary to learn without forgetting as we only care about one language(task). However, long have there been claims in MTL that a model could benefit from maintaining its performance on task `A` when aiming to do good in task `B`. This is rather intuitive in our problem: imagine `B` is a low-resource language, it is likely that model quickly overfits to a small training set in language `B` and forgets the useful representations learned on `A`. We expect that this joint optimization regularizes model from forgetting knowledge learned on `A`. A sketch diagram can be seen on RHS of Figure 1. In experiments, auxiliary training significantly improved model performance in low-resource condition as we expected.

**Conclusion.** We introduced two recipes for VG-CLTL, addressing both language-agnostic and language-deterministic scenarios, achieving better model performance by utilizing multilingual data. We proposed a *Translator* layer with weights obtained from the task of BLI who embeds a learnable language space transformation for continuous word vectors into neural networks. And finally we investigated two strategies (inspired by MTL) for multilingual training. By experimenting with different amounts of data points, we found that the above methods enhance both multilingual and monolingual models by a large margin, especially in low resource conditions.

## References


P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *ICLR workshop track*.

F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. *EMNLP (short paper)*.

R. Kiros, R. Salakhutdinov, and R. S. Zemel. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics (TACL)*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1517–1526.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *ICLR*.

I. Vendrov, R. Kiros, S. Fidler, and R/ Urtasun. 2016. Order-embeddings of images and language. *ICLR*.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE TPAMI*.