Visually Grounded Cross-Lingual Transfer Learning

Fangyu Liu [1], **Rémi Lebret** [2], **Karl Aberer** [2]
[1] University of Waterloo, Waterloo, ON, Canada
[2] École polytechnique fédérale de Lausanne, Lausanne, Switzerland

1 Overview

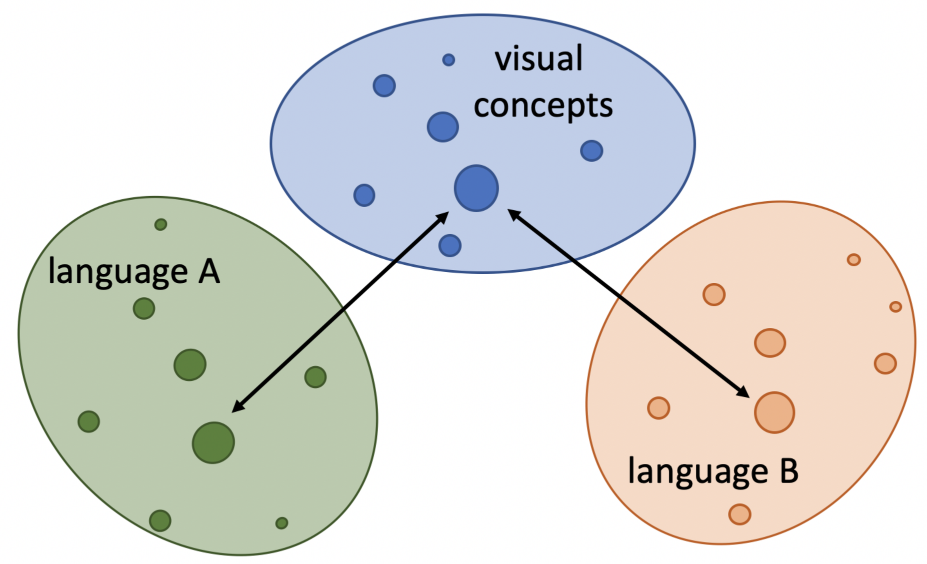

Figure 1: Grounding multilingual concepts with vision as the shared modality.

The Task: Bidirectional Text-Image Retrieval. Given an image, the model retrieves the most descriptive caption; or given a caption, the model selects the most descriptive image.

The Basic Model: Visual-Semantic Embeddings (VSE). VSE bridges language and vision by jointly optimizing and aligning semantic embeddings (from texts) and visual embeddings (from images), aiming that texts/images with similar semantics are close to each other in the embedding space.

Our Idea: Grounding Multilingual Concepts with Vision. As vision is universal, multilingual texts would be grounded by consistent visual signals extracted from images which helps to transport knowledge across languages. We propose a language space transformation embedded inside neural networks, addressing transfer learning under continuous word embeddings.

2 Model Details

First, we train a language transformation matrix M called TRANSLATOR by applying SVD and RCSLS [2]. Then, we embed M in the pipeline of standard VSE training.

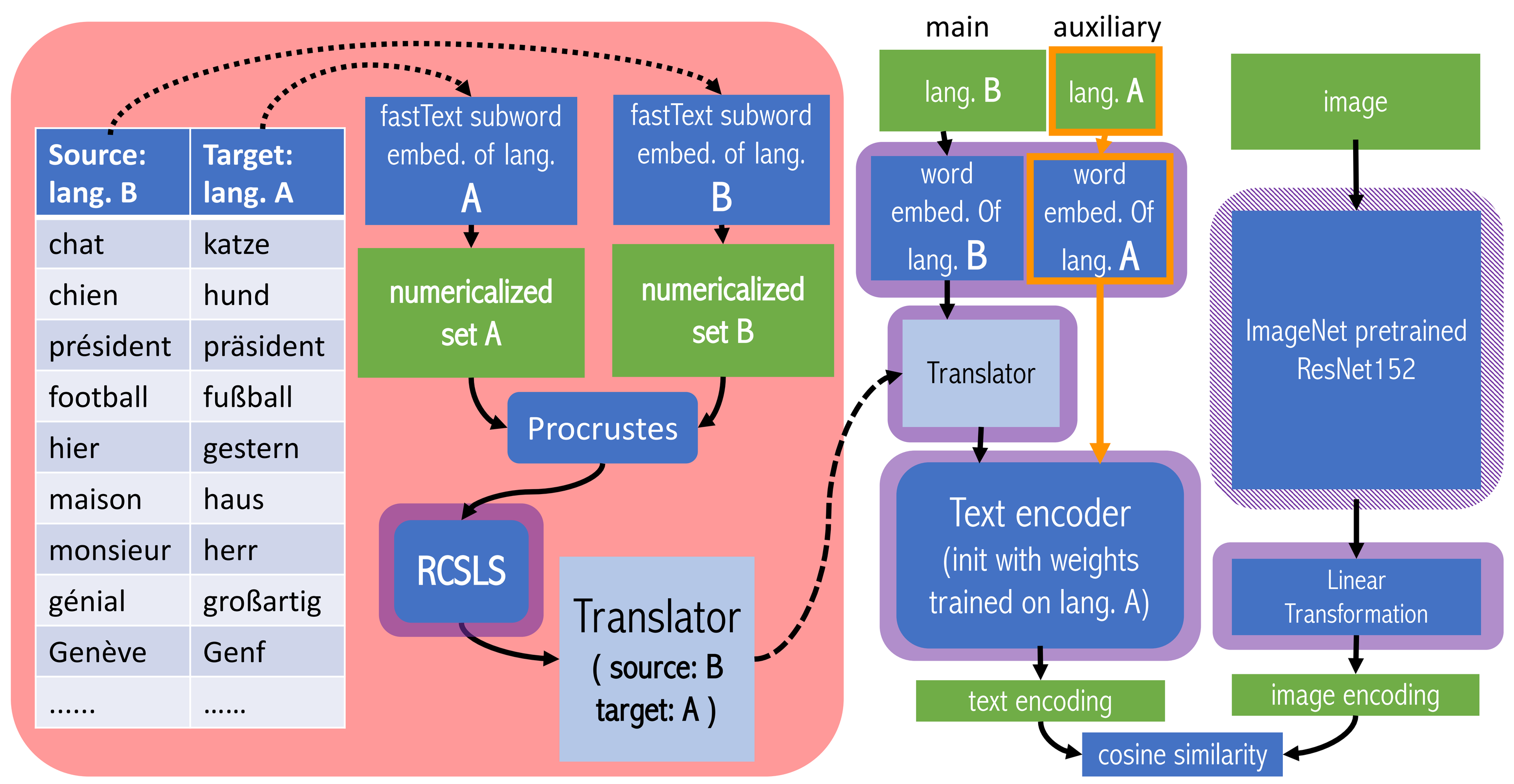

Figure 2: Overview of our proposed method.

SVD: $\arg\min_M \|MT - I\|_2^2$

RCSLS:

$$\min_M \frac{1}{n}\sum_{i=1}^{n}\left(-2a_i^\top M^\top b_i + r(Ma_i, B) + r(b_i, A)\right)$$

where $r(x, Y) := \frac{1}{k}\sum_{y\in \mathrm{kNN}(x,Y)} x^\top y$.

Training.

$$s(i,t) = \langle \frac{f(M\cdot t)}{\|f(M\cdot t)\|_2}, \frac{g(i)}{\|g(i)\|_2}\rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

$$\min_\theta \sum_{i\in I}\sum_{\bar{t}\in T\setminus\{t\}} \max\{0, \alpha - s(i,t) + s(i,\bar{t})\} + \sum_{t\in T}\sum_{\bar{i}\in I\setminus\{i\}} \max\{0, \alpha - s(t,i) + s(t,\bar{i})\}$$

3 Results

Dataset. We use a self-collected very large-scale news image-caption dataset containing 350,204 `de` and 178,270 `fr` samples.

Three Configurations. To demonstrate how *Translator* functions exactly, we experiment three protocols on the text branch:

- *FS*: `fr` subword embeddings [1] + text encoder (randomly initialized);
- *T1*: `fr` subword embeddings [1] + *Translator* (randomly initialized) + text encoder (initialized with `de` weights);
- *T2*: `fr` subword embeddings [1] + *Translator* (initialized with SVD+RCSLS) + text encoder (initialized with `de` weights).

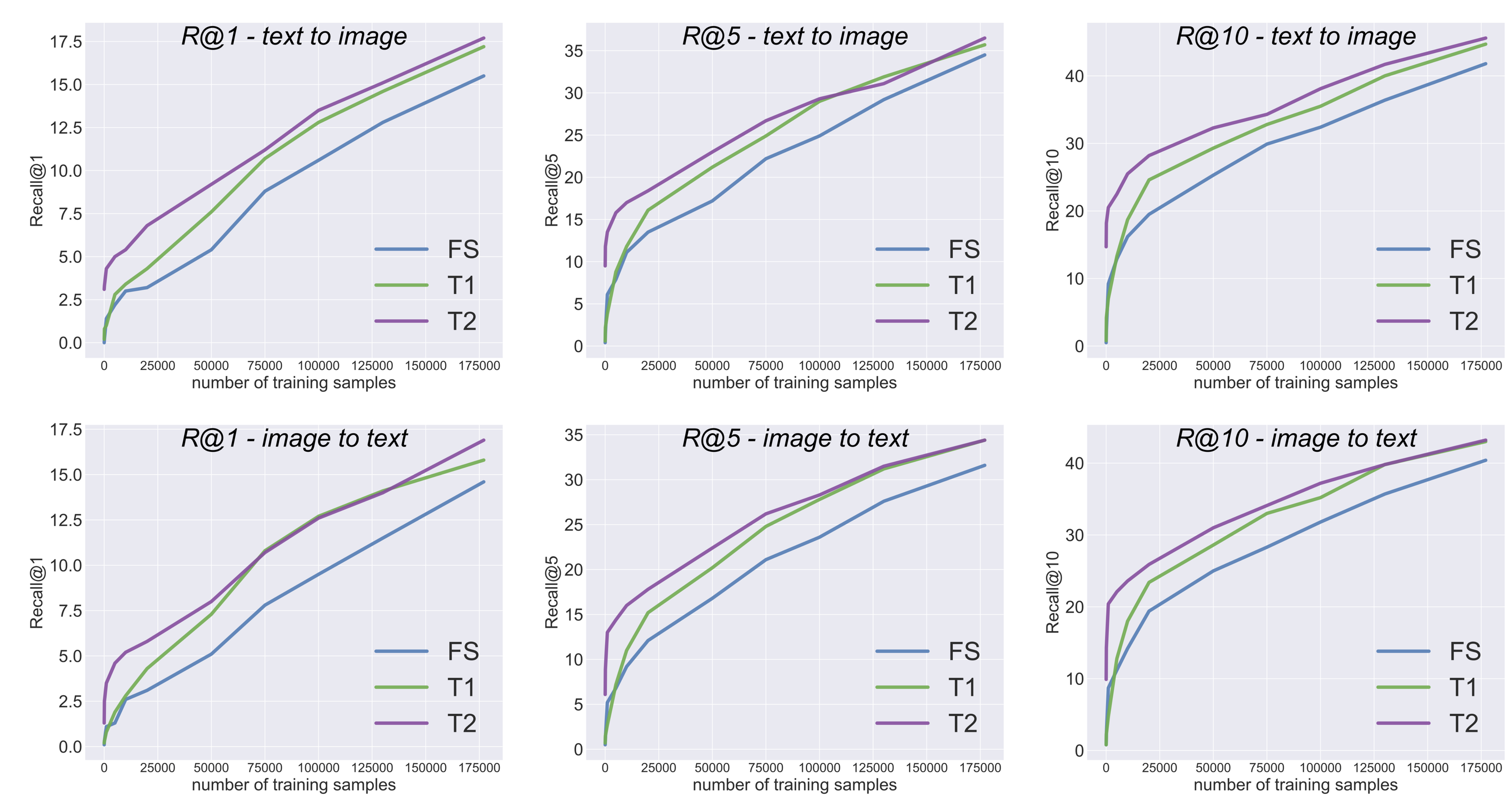

Figure 3: Plotting recalls (y axis) against number of `fr` training examples (x axis). First row is text→image R@1, R@5, R@10 respectively; second row is image→text R@1, R@5, R@10.

References

[1] Piotr Bojanowski et al. Enriching word vectors with subword information. *TACL*, 2017.

[2] Armand Joulin et al. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *EMNLP 2018*.