Visual Bilingual Lexicon Induction with Transferred ConvNet Features

Douwe Kiela
Computer Laboratory
University of Cambridge
douwe.kiela@cl.cam.ac.uk

Ivan Vulić
Department of Computer Science
KU Leuven
ivan.vulic@cs.kuleuven.be

Stephen Clark
Computer Laboratory
University of Cambridge
stephen.clark@cl.cam.ac.uk

Abstract

This paper is concerned with the task of bilingual lexicon induction using image-based features. By applying features from a convolutional neural network (CNN), we obtain state-of-the-art performance on a standard dataset, obtaining a 79% relative improvement over previous work which uses bags of visual words based on SIFT features. The CNN image-based approach is also compared with state-of-the-art linguistic approaches to bilingual lexicon induction, even outperforming these for one of three language pairs on another standard dataset. Furthermore, we shed new light on the type of visual similarity metric to use for genuine similarity versus relatedness tasks, and experiment with using multiple layers from the same network in an attempt to improve performance.

1 Introduction

Bilingual lexicon induction is the task of finding words that share a common meaning across different languages. It plays an important role in a variety of tasks in information retrieval and natural language processing, including cross-lingual information retrieval (Lavrenko et al., 2002; Levow et al., 2005) and statistical machine translation (Och and Ney, 2003). Although parallel corpora have been used successfully for inducing bilingual lexicons for some languages (Och and Ney, 2003), these corpora are either too small or unavailable for many language pairs. Consequently, mono-lingual approaches that rely on comparable instead of parallel corpora have been developed (Fung and Yee, 1998; Koehn and Knight, 2002). These approaches work by mapping language pairs to a shared bilingual space and extracting lexical items from that space. Bergsma and Van Durme (2011) showed that this bilingual space need not be linguistic in nature: they used labeled images from the Web to obtain bilingual lexical translation pairs based on the visual features of corresponding images. Local features are computed using SIFT (Lowe, 2004) and color histograms (Deselaers et al., 2008) and aggregated as bags of visual words (BOVW) (Sivic and Zisserman, 2003) to get bilingual representations in a shared visual space. Their highest performance is obtained by combining these visual features with normalized edit distance, an orthographic similarity metric (Navarro, 2001).

There are several advantages to having a visual rather than a linguistic intermediate bilingual space: First, while labeled images are readily available for many languages through resources such as Google Images, language pairs that have sizeable comparable, let alone parallel, corpora are relatively scarce. Second, it has been found that meaning is often grounded in the perceptual system, and that the quality of semantic representations improves significantly when they are grounded in the visual modality (Silberer and Lapata, 2012; Bruni et al., 2014). Having an intermediate visual space means that words in different languages can be grounded in the same space. Third, it is natural to use vision as an intermediate: when we communicate with someone who does not speak our language, we often communicate by directly referring to our surroundings. Languages that are linguistically far apart will, by cognitive necessity, still refer to objects in the same visual space. While some approaches to bilingual lexicon induction rely on orthographic properties (Haghighi et al., 2008; Koehn and Knight, 2002) or properties of frequency distributions (Schafer and Yarowsky, 2002) that will work only for

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 148–158,
Lisbon, Portugal, 17-21 September 2015.

closely related languages, a visual space can work for any language, whether it's English or Chinese, Arabic or Icelandic, or all Greek to you.

It has recently been shown, however, that much better performance can be achieved on semantic similarity and relatedness tasks by using visual representations from deep convolutional neural networks (CNNs) instead of BOVW features (Kiela and Bottou, 2014). In this paper we apply such CNN-derived visual features to the task of bilingual lexicon induction. To obtain a translation of a word in a source language, we find the nearest neighbours from words in the target language, where words in both languages reside in a shared visual space made up of CNN-based features. Nearest neighbours are found by applying similarity metrics from both Kiela and Bottou (2014) and Bergsma and Van Durme (2011). In summary, the contributions of this paper are:

- We obtain a relative improvement of 79% over Bergsma and Van Durme (2011) on a standard dataset based on fifteen language pairs.
- We shed new light on the question of whether genuine similarity versus semantic relatedness tasks require different similarity metrics for optimal performance (Kiela and Bottou, 2014).
- We experiment with using different layers of the CNN and find that performance is not affected significantly in either case, obtaining a slight improvement for the relatedness task but no improvement for genuine similarity.
- Finally, we show that the visual approach outperforms the linguistic approaches on one of the three language pairs on a standard dataset. To our knowledge this is the first work to provide a comparison of visual and state-of-the-art linguistic approaches to bilingual lexicon induction.

2 Related Work

2.1 Bilingual Lexicon Learning

Bilingual lexicon learning is the task of automatically inducing word translations from raw data, and is an attractive alternative to the time-consuming and expensive process of manually building high-quality resources for a wide variety of language pairs and domains. Early approaches relied on limited and domain-restricted parallel data, and the induced lexicons were typically a by-product of word alignment models (Och and Ney, 2003). To alleviate the issue of low coverage, a large body of work has been dedicated to lexicon learning from more abundant and less restricted comparable data, e.g., (Fung and Yee, 1998; Rapp, 1999; Gaussier et al., 2004; Shezaf and Rappoport, 2010; Tamura et al., 2012). However, these models typically rely on the availability of bilingual seed lexicons to produce shared bilingual spaces, as well as large repositories of comparable data. Therefore, several approaches attempt to learn lexicons from large monolingual data sets in two languages (Koehn and Knight, 2002; Haghighi et al., 2008), but their performance again relies on language pair-dependent clues such as orthographic similarity. A further approach removed the requirement of seed lexicons, and induced lexicons using bilingual spaces spanned by multilingual probabilistic topic models (Vulić et al., 2011; Liu et al., 2013; Vulić and Moens, 2013b). However, these models require document alignments as initial bilingual signals.

In this work, following recent research in multi-modal semantics and image representation learning—in particular deep learning and convolutional neural networks—we test the ability of purely visual data to induce shared bilingual spaces and to consequently learn bilingual word correspondences in these spaces. By compiling images related to linguistic concepts given in different languages, the potentially prohibitive data requirements and language pair-dependence from prior work is removed.

2.2 Deep Convolutional Neural Networks

Deep convolutional neural networks (CNNs) have become extremely popular in the computer vision community. These networks currently provide state-of-the-art performance for a variety of key computer vision tasks such as object recognition (Razavian et al., 2014). They tend to be relatively deep, consisting of a number of rectified linear unit layers (Nair and Hinton, 2010) and a series of convolutional layers (Krizhevsky et al., 2012). Recently, such layers have been used in transfer learning techniques, where they are used as mid-level features in other computer vision tasks (Oquab et al., 2014). Although the idea of transferring CNN features is not new (Driancourt and Bottou, 1990), the simultaneous availability of

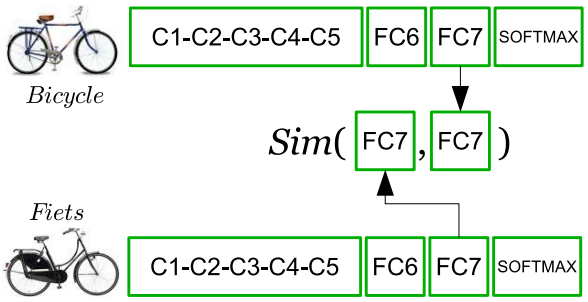

Figure 1: Illustration of calculating similarity between images from different languages.

massive amounts of data and cheap GPUs has led to considerable advances in computer vision, similar in scale to those witnessed with SIFT and HOG descriptors a decade ago (Razavian et al., 2014).

2.3 Multi-Modal Semantics

Multi-modal semantics is motivated by parallels with human concept acquisition. It has been found that semantic knowledge, from a very early age, relies heavily on perceptual information (Louwerse, 2008), and there exists substantial evidence that many concepts are grounded in the perceptual system (Barsalou, 2008). One way to accomplish such grounding is by combining linguistic representations with information from a perceptual modality, obtained from, e.g., property norming experiments (Silberer and Lapata, 2012; Silberer et al., 2013; Roller and Schulte im Walde, 2013; Hill and Korhonen, 2014) or extracting features from raw image data (Feng and Lapata, 2010; Leong and Mihalcea, 2011; Bruni et al., 2014; Kiela et al., 2014). Such multi-modal visual approaches often rely on local descriptors, such as SIFT (Lowe, 2004), SURF (Bay et al., 2008), or HOG (Dalal and Triggs, 2005), as well as pyramidal variants of these descriptors such as PHOW (Bosch et al., 2007). However, deep CNN features have recently been successfully transferred to multi-modal semantics (Kiela and Bottou, 2014; Shen et al., 2014). Deep learning techniques have also been successfully employed in cross-modal tasks (Frome et al., 2013; Socher et al., 2014; Lazaridou et al., 2014; Kiros et al., 2014). Other examples of multi-modal deep learning use restricted Boltzmann machines (Srivastava and Salakhutdinov, 2014) or auto-encoders (Wu et al., 2013; Silberer and Lapata, 2014).

3 A Purely Visual Approach to Bilingual Lexicon Learning

We assume that the best translation, or matching lexical item, of a word w_s (in the source language) is the word w_t (in the target language) that is the nearest cross-lingual neighbour to w_s in the bilingual visual space. Hence a similarity (or distance) score between lexical items from different languages is required. In this section, we describe: one, how to build image representations from sets of images associated with each lexical item, i.e. how to induce a shared bilingual visual space in which all lexical items are represented; and two, how to compute the similarity between lexical items using their visual representations in the shared bilingual space. We also describe the evaluation datasets and metrics we use.

To facilitate further research, we will make our code and data publicly available. Please see the following webpage: `http://www.cl.cam.ac.uk/~dk427/bli.html`.

3.1 Image Representations

We use Google Images to extract the top n ranked images for each lexical item in the evaluation datasets. It has been shown that images from Google yield higher quality representations than comparable sources such as Flickr (Bergsma and Goebel, 2011) and that Google-derived datasets are competitive with "hand prepared datasets" (Fergus et al., 2005). Google Images also has the advantage that it has full coverage and is multi-lingual, as opposed to other potential image sources such as ImageNet (Deng et al., 2009) or the ESP Game Dataset (von Ahn and Dabbish, 2004). For each Google search we specify the target language corresponding to the lexical item's language. Figure 2 gives some example images retrieved using the same query terms in different languages. For each image, we extract the pre-softmax layer of an AlexNet (Krizhevsky et al., 2012). The network contains a number of layers, starting with five convolutional layers, two fully connected layers and finally a softmax, and has been pre-trained on the ImageNet classification task using Caffe (Jia et al., 2014). See Figure 1 for a simple diagram illustrating the approach.

3.2 Visual Similarity

Suppose that, as part of the evaluation, the similarity between *bicycle* and *fiets* is required. Each of

the two words has n images associated with it – the top n as returned by Google image search, using *bicycle* and *fiets* as separate query terms. Hence to calculate the similarity, a measure is required which takes two sets of images as input. The standard approach in multi-modal semantics is to derive a single image representation for each word, e.g., by averaging the n images. An alternative is to take the pointwise maximum across the n image vector representations, also producing a single vector (Kiela and Bottou, 2014). Kiela and Bottou call these combined representations CNN-MEAN and CNN-MAX, respectively. Cosine is then used to calculate the similarity between the resulting pair of image vectors.

An alternative strategy, however, is to consider the similarities between individual images instead of their aggregated representations. Bergsma and Van Durme (2011) propose two similarity metrics based on this principle: taking the average of the maximum similarity scores (AVGMAX), or the maximum of the maximum similarity scores (MAXMAX) between associated images. Continuing with our example, for each of the n images for *bicycle*, the maximum similarity is found by searching over the n images for *fiets*. AVGMAX then takes the average of those n maximum similarites; MAXMAX takes the maximum. To avoid confusion, we will refer to the CNN-based models that use these metrics as CNN-AVGMAX and CNN-MAXMAX. Formally, these metrics are defined as in Table 1. We experiment with both kinds of MAX and find that they optimize for different kinds of similarity.

AVGMAX	$\frac{1}{n}\sum_{i_s \in \mathcal{I}(w_s)} \max_{i_t \in \mathcal{I}(w_t)} sim(i_s, i_t)$
MAXMAX	$\max_{i_s \in \mathcal{I}(w_s)} \max_{i_t \in \mathcal{I}(w_t)} sim(i_s, i_t)$
CNN-MEAN	$sim(\frac{1}{n}\sum_{i_s \in \mathcal{I}(w_s)} i_s, \frac{1}{n}\sum_{i_t \in \mathcal{I}(w_t)} i_t)$
CNN-MAX	$sim(\max' \mathcal{I}(w_s), \max' \mathcal{I}(w_t))$

Table 1: Visual similarity metrics between two sets of n images. $\mathcal{I}(w_s)$ represents the set of images for a given source word w_s, $\mathcal{I}(w_t)$ the set of images for a given target word w_t; $\max'$ takes a set of vectors and returns the single element-wise maximum vector.

3.3 Evaluations

Test Sets. Bergsma and Van Durme's primary evaluation dataset consists of a set of five hundred matching lexical items for fifteen language pairs, based on six languages. (The fifteen pairs results from all ways of pairing six languages). The data is publicly available online.[1] In order to get the five hundred lexical items, they first rank nouns by the conditional probability of them occurring in the pattern "{*image,photo,photograph,picture*} *of* {*a,an*} _____" in the web-scale Google N-gram corpus (Lin et al., 2010), and take the top five hundred words as their English lexicon. For each item in the English lexicon, they obtain corresponding items in the other languages—Spanish, Italian, French, German and Dutch—through Google Translate. We call this dataset BERGSMA500.

In addition to that dataset, we evaluate on a dataset constructed to measure the general performance of bilingual lexicon learning models from comparable Wikipedia data (Vulić and Moens, 2013a). The dataset comprises 1,000 nouns in three languages: Spanish (ES), Italian (IT), and Dutch (NL), along with their one-to-one gold-standard word translations in English (EN) compiled semi-automatically using Google Translate and manual annotators for each language. We call this dataset VULIC1000[2]. The test set is accompanied with comparable data for training, for the three language pairs ES/IT/NL-EN on which text-based models for bilingual lexicon induction were trained (Vulić and Moens, 2013a).

Given the way that the BERGSMA500 dataset was created, in particular the use of the pattern described above, it contains largely concrete linguistic concepts (since, eg, *image of a democracy* is unlikely to have a high corpus frequency). In contrast, VULIC1000 was designed to capture general bilingual word correspondences, and contains several highly abstract test examples, such as *entendimiento (understanding)* and *desigualdad (inequality)* in Spanish, or *scoperta (discovery)* and *cambiamento (change)* in Italian. Using the two evaluation datasets can potentially provide

[1] http://www.clsp.jhu.edu/~sbergsma/LexImg/

[2] http://people.cs.kuleuven.be/~ivan.vulic/software/

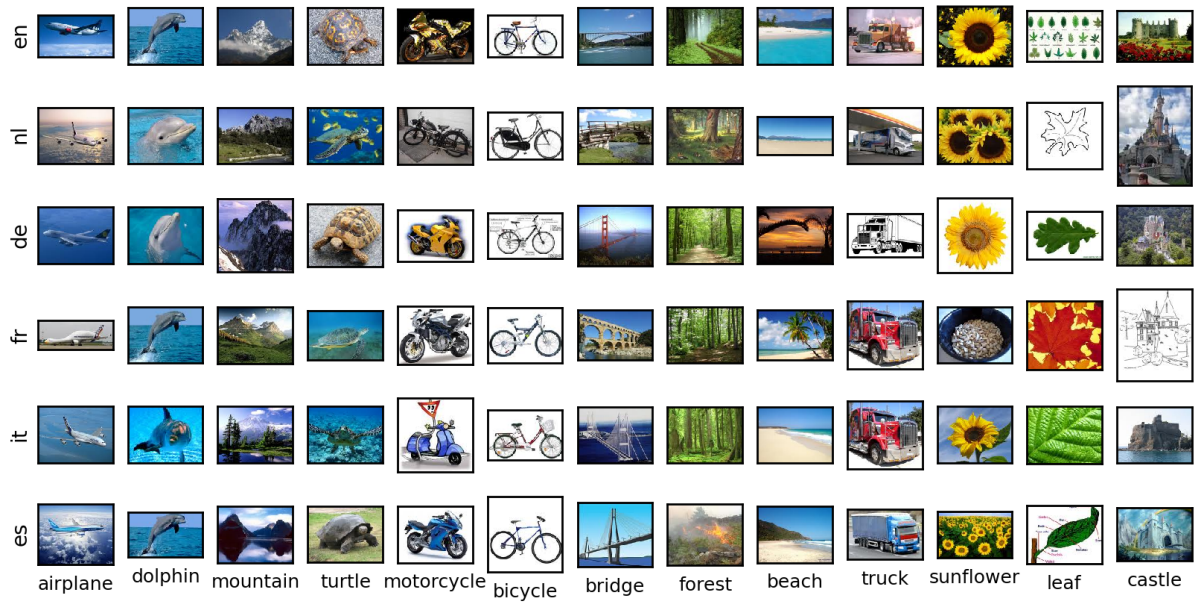

Figure 2: Example images for the languages in the Bergsma and Van Durme dataset.

Method	P@1	P@5	P@20	MRR
B&VD Visual-Only	31.1	41.4	53.7	0.367
B&VD Visual + NED	48.0	59.5	68.7	0.536
CNN-AVGMAX	**56.7**	**69.2**	**77.4**	**0.658**
CNN-MAXMAX	42.8	60.0	64.5	0.529
CNN-MEAN	50.5	62.7	71.1	0.586
CNN-MAX	51.4	64.9	74.8	0.608

Table 2: Performance on BERGSMA500 compared to Bergsma and Van Durme (B&VD).

some insight into how purely visual models for bilingual lexicon induction behave with respect to both abstract and concrete concepts.

Evaluation Metrics. We measure performance in a standard way using mean-reciprocal rank:

$$MRR = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{rank(w_s, w_t)} \quad (1)$$

where $rank(w_s, w_t)$ denotes the rank of the correct translation w_t (as provided in the gold standard) in the ranked list of translation candidates for w_s, and M is the number of test cases. We also use *precision at N* (P@N) (Gaussier et al., 2004; Tamura et al., 2012; Vulić and Moens, 2013a), which measures the proportion of test instances where the correct translation is within the top N highest ranked translations.

4 Results

We evaluate the four similarity metrics on the BERGSMA500 dataset and compare the results to the systems of Bergsma and Van Durme, who report results for the AVGMAX function, having concluded that it performs better than MAXMAX on English-Spanish translations. We report their best-performing visual-only system, which combines SIFT-based descriptors with color histograms, as well as their best-performing overall system, which combines the visual approach with normalized edit distance (NED). Results are averaged over fifteen language pairs.

The results can be seen in Table 2. Each of the CNN-based methods outperforms the B&VD systems. The best performing method overall, CNN-AVGMAX, provides a 79% relative improvement over the B&VD visual-only system on the MRR measure, and a 23% relative improvement over their best-performing approach, which includes non-visual information in the form of orthographic similarity. Moreover, their methods include a tuning parameter λ that governs the contributions of SIFT-based, color histogram and normalized edit distance similarity scores, whilst our approach does not require any parameter tuning.

4.1 Similarity and Relatedness

The results in Table 2 indicate that the per-image CNN-AVGMAX metric outperforms the

Language Pair	Method	P@1	P@5	P@10	P@20	MRR
ES ⇒ EN	BOOTSTRAP	57.7	74.7	80.9	84.8	0.652
	CNN-AVGMAX	41.9	54.6	59.1	65.6	0.485
	CNN-MAXMAX	34.9	47.4	53.7	58.5	0.414
	CNN-MEAN	35.4	48.5	51.7	55.8	0.416
	CNN-MAX	33.3	46.3	50.3	54.5	0.395
IT ⇒ EN	BOOTSTRAP	64.7	80.6	85.6	89.7	0.716
	CNN-AVGMAX	28.3	40.6	44.8	50.9	0.343
	CNN-MAXMAX	22.6	33.5	38.6	44.4	0.282
	CNN-MEAN	22.7	33.2	37.9	42.6	0.281
	CNN-MAX	21.3	32.7	36.8	41.5	0.269
NL ⇒ EN	BOOTSTRAP	20.6	35.7	43.4	51.3	0.277
	CNN-AVGMAX	38.4	48.5	53.7	58.6	0.435
	CNN-MAXMAX	30.8	42.6	47.8	52.9	0.367
	CNN-MEAN	32.3	42.3	46.5	50.1	0.373
	CNN-MAX	30.4	41.0	44.3	49.3	0.356

Table 4: Performance on VULIC1000 compared to the linguistic bootstrapping method of Vulić and Moens (2013b).

Method	MEN	SimLex-999
CNN-AVGMAX	0.56	0.34
CNN-MAXMAX	0.55	**0.36**
CNN-MEAN	**0.61**	0.32
CNN-MAX	0.60	0.27

Table 3: Spearman ρ_s correlation for the visual similarity metrics on a relatedness (MEN) and a genuine similarity (SimLex-999) dataset.

aggregated visual representation-based metrics of CNN-MEAN and CNN-MAX, despite the fact that Kiela and Bottou (2014) achieved optimal performance using the latter metrics on a well-known conceptual relatedness dataset. It has been noted before that there is a clear distinction between similarity and relatedness. This is one of the reasons that, for example, WordSim353 (Finkelstein et al., 2002) has been criticized: it gives high similarity scores to cases of genuine similarity as well as relatedness (Agirre et al., 2009; Hill et al., 2014). The MEN dataset (Bruni et al., 2014) that Kiela and Bottou (2014) evaluate on explicitly measures word relatedness. In contrast, the current lexicon learning task seems to require something else than relatedness: whilst a *chair* and *table* are semantically related, a translation for *chair* is not a good translation for *table*. For example, we want to make sure we translate *chair* to *stuhl* in German, and not to *tisch*. In other words, what we are interested in for this particular task is genuine similarity, rather than relatedness.

Thus, we can evaluate the quality of our similarity metrics by comparing their performance on similarity and relatedness tasks: if a metric performs well at measuring genuine similarity, this is indicative of its performance in the bilingual lexicon induction task. In order to examine this question further, we evaluate performance on the MEN dataset, which measures relatedness (Bruni et al., 2014), and the nouns-subset of the SimLex-999 dataset, which measures genuine similarity (Hill et al., 2014). For each pair in the dataset, we calculate the similarity score and report the Spearman ρ_s correlation, which measures how well the ranking of pairs given by the automatic system matches that according to the gold-standard human similarity scores. The results are reported in Table 3.

It is clear that the per-image similarity metrics perform better on genuine similarity, as measured by SimLex-999, than on relatedness, as measured by MEN. In fact, the "aggressive" CNN-MAXMAX method, which picks out a single pair of images to represent a linguistic pair, works best for SimLex-999, indicating how stringently it focuses on genuine similarity. For the aggregated visual representation-based metrics, we see the opposite effect: they perform better on the relatedness task. This sheds light on a question raised by Kiela and Bottou (2014), where they speculate

that certain errors are a result of whether their visual similarity metric measures genuine similarity on the one hand or relatedness on the other: we are better off using per-image visual metrics for genuine similarity, while aggregated visual representation-based metrics yield better performance on relatedness tasks.

4.2 Results on VULIC1000

This section compares our visual-only approach to linguistic approaches for bilingual lexicon induction. Since BERGSMA500 has not been evaluated with such methods, we evaluate on the VULIC1000 dataset (Vulić and Moens, 2013a). This dataset has been used to test the ability of bilingual lexicon induction models to learn translations from comparable data (see sect. 3.3). We do not necessarily expect visual methods to outperform linguistic ones, but it is instructive to see the comparison.

We compare our visual models against the current state-of-the-art lexicon induction model using comparable data (Vulić and Moens, 2013b). This model induces translations from comparable Wikipedia data in two steps: (1) It learns a set of highly reliable one-to-one translation pairs using a shared bilingual space obtained by applying the multilingual probabilistic topic modeling (MuPTM) framework (Mimno et al., 2009). (2) These highly reliable one-to-one translation pairs serve as dimensions of a word-based bilingual semantic space (Gaussier et al., 2004; Tamura et al., 2012). The model then bootstraps from the high-precision seed lexicon of translations and learns new dimensions of the bilingual space until convergence. This model, which we call BOOTSTRAP, obtains the current best results on the evaluation dataset. For more details about the bootstrapping model and its comparison against other approaches, we refer to Vulić and Moens (2013b).

Table 4 shows the results for the language pairs in the VULIC1000 dataset. Of the four similarity metrics, CNN-AVGMAX again performs best, as it did for BERGSMA500. The linguistic BOOTSTRAP method outperforms our visual approach for two of the three language pairs, but, for the NL-EN language pair, the visual methods in fact perform better. This can be explained by the observation that Vulić and Moens's NL-EN training data for the BOOTSTRAP model is less abundant (2-3 times fewer Wikipedia articles) and of lower quality than the data for their ES-EN and IT-EN models. We view these results as highly encouraging: while purely visual methods cannot yet reach the peak performance of linguistic approaches that are trained on sufficient amounts of high-quality text data, they outperform linguistic state-of-the-art methods when there is less or lower quality text data available —which one might reasonably expect to be the default scenario.

Method	**FC7**	**FC6+FC7**	**POOL5+ FC6+FC7**
MEN			
CNN-AVGMAX	0.56	0.57	0.57
CNN-MAXMAX	0.55	0.55	0.56
CNN-MEAN	0.61	0.61	0.61
CNN-MAX	0.60	0.62	0.61
SimLex-999			
CNN-AVGMAX	0.34	0.33	0.31
CNN-MAXMAX	0.36	0.35	0.34
CNN-MEAN	0.32	0.32	0.31
CNN-MAX	0.27	0.26	0.26

Table 5: Spearman ρ_s correlation for the visual similarity metrics on a relatedness (MEN) and a genuine similarity (SimLex-999) dataset using more than one layer from the CNN.

4.3 Adding CNN Layers

The AlexNet (Krizhevsky et al., 2012) from which our image representations are extracted contains a number of layers. Kiela and Bottou (2014) only use the fully connected pre-softmax layer (which we call FC7) for their image representations. It has been found, however, that other layers in the network, especially the preceding fully connected (FC6) and fifth convolutional max pooling (POOL5) layers, also have good properties for usage in transfer learning (Girshick et al., 2014; Yosinski et al., 2014). Hence we performed a (very) preliminary investigation of whether performance increases with the use of additional layers.

In light of our findings concerning the difference between genuine similarity and relatedness, this also gives rise to the question of whether the additional layers might be useful for similarity or relatedness, or both. We hypothesize that the nature of the task matters here: if we are only concerned with genuine similarity, layer FC7 is likely to contain all the necessary information to judge whether two images are similar or not, since

Dataset	Language	Image dispersion
BERGSMA500	EN	0.640 (σ=0.074)
	ES	0.639 (σ=0.072)
	IT	0.646 (σ=0.071)
	FR	0.647 (σ=0.072)
	DE	0.642 (σ=0.072)
	NL	0.645 (σ=0.074)
VULIC1000	EN	0.705 (σ=0.095)
	ES	0.694 (σ=0.092)
	IT	0.725 (σ=0.078)
	NL	0.716 (σ=0.080)

Table 6: Average image dispersion for the datasets, by language.

the network has been trained for object recognition. If, however, we are interested in relatedness, related properties may just as well be encoded deeper in the network, so in the layers preceding FC7 rather than in FC7 itself.

We combined CNN layers with each other by concatenating the normalized layers. For the bilingual lexicon induction tasks, we found that performance did not signficantly increase, which is consistent with our hypothesis (since bilingual lexicon induction requires genuine similarity rather than relatedness, and so only requires FC7). We then tested on the MEN dataset (Bruni et al., 2014) for relatedness and the nouns subset of the SimLex-999 dataset (Hill et al., 2014) for genuine similarity. The results can be found in Table 5.

The results appear to indicate that adding such additional information does not have a clear effect for genuine similarity, but may lead to a small performance increase for relatedness. This could explain why we did not see increased performance on the bilingual lexicon induction task with additional layers. However, the increase in performance on the relatedness task is relatively minor, and further investigation is required into the utility of the additional layers for relatedness tasks.

5 Discussion

A possible explanation for the difference in performance between languages and datasets is that some words are more concrete than others: a visual representation for *elephant* is likely to be of higher quality than one for *happiness*. Visual representations in multi-modal models have been found to perform much better for concrete than abstract concepts (Kiela et al., 2014).

Although concreteness ratings are available for (some) English words, this is not the case for other languages, so in order to examine the concreteness of the datasets we use a substitute method that has been shown to closely mirror how abstract a concept is: image dispersion (Kiela et al., 2014). The image dispersion d of a concept word w is defined as the average pairwise cosine distance between all the image representations $\{i_1 \dots i_n\}$ in the set of images for a given word:

$$d(w) = \frac{2}{n(n-1)} \sum_{i<j\leq n} 1 - \frac{i_j \cdot i_k}{|i_j||i_k|} \quad (2)$$

The average image dispersions for the two datasets, broken down by language, are shown in Table 6. BERGSMA500 has a lower average image dispersion score in general, and thus is more concrete than VULIC1000. It also has less variance. This may explain why we score higher, in absolute terms, on that dataset than on the more abstract one.

When examining individual languages in the datasets, we note that the worst performing language on VULIC1000 is Italian, which is also the most abstract dataset, with the highest average image dispersion score and the lowest variance.

There is some evidence that abstract concepts are also perceptually grounded (Lakoff and Johnson, 1999), but in a more complex way, since abstract concepts express more varied situations (Barsalou and Wiemer-Hastings, 2005). Using an image resource like Google Images that has full coverage for almost any word, means that we can retrieve what we might call "associated" images (such as images of voters for words like *democracy*) as opposed to "extensional" images (such as images of cats for *cat*). This explains why we still obtain good performance on the more abstract VULIC1000 dataset, in some cases outperforming linguistic methods: even abstract concepts can have a clear visual representation, albeit of the associated rather than extensional kind.

However, abstract concepts are overall more likely to yield noisier image sets. Thus, one way to improve results would be to take a multi-modal approach, where we also include linguistic information, if available, especially for abstract concepts.

6 Conclusions and Future Work

We have presented a novel approach to bilingual lexicon induction that uses convolutional neural

network-derived visual features. Using only such visual features, we outperform existing visual and orthographic systems, and even a state-of-the-art linguistic approach for one language, on standard bilingual lexicon induction tasks. In doing so, we have shed new light on which visual similarity metric to use for similarity or relatedness tasks, and have experimented with using multiple layers from a CNN. The beauty of the current approach is that it is completely language agnostic and closely mirrors how humans would perform bilingual lexicon induction: by referring to the external world.

Acknowledgments

DK is supported by EPSRC grant EP/I037512/1. IV is supported by the PARIS project (IWT-SBO 110067) and the PDM Kort postdoctoral fellowship from KU Leuven. SC is supported by ERC Starting Grant DisCoTex (306920) and EPSRC grant EP/I037512/1. We thank Marco Baroni for useful feedback and the anonymous reviewers for their helpful comments.

References

Eneko Agirre, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *NAACL*, pages 19–27.

Lawrence W. Barsalou and Katja Wiemer-Hastings. 2005. Situating abstract concepts. In *Grounding cognition: The role of perception and action in memory, language, and thought*, pages 129–163.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359.

Shane Bergsma and Randy Goebel. 2011. Using visual information to predict lexical preference. In *RANLP*, pages 399–405.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI*, pages 1764–1769.

Anna Bosch, Andrew Zisserman, and Xavier Muñoz. 2007. Image classification using random forests and ferns. In *ICCV*, pages 1–8.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artifical Intelligence Research*, 49:1–47.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Thomas Deselaers, Daniel Keysers, and Hermann Ney. 2008. Features for image retrieval: An experimental comparison. *Information Retrieval*, 11(2):77–107.

Xavier Driancourt and Léon Bottou. 1990. TDNN-extracted features. In *Neuro Nimes 90*.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *NAACL*, pages 91–99.

Robert Fergus, Fei-Fei Li, Pietro Perona, and Andrew Zisserman. 2005. Learning object categories from Google's image search. In *ICCV*, pages 1816–1823.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116—131.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *ACL*, pages 414–420.

Éric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *ACL*, pages 526–533.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, pages 771–779.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *EMNLP*, pages 255–265.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR*, abs/1408.3456.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pages 36–45.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *ACL*, pages 835–841.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Multimodal neural language models. In *ICML*, pages 595–603.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ULA'02 Workshop*, pages 9–16.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114.

George Lakoff and Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to Western thought*.

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *SIGIR*, pages 175–182.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *ACL*, pages 1403–1414.

Chee Wee Leong and Rada Mihalcea. 2011. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *IJCNLP*, pages 1403–1407.

Gina-Anne Levow, Douglas Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*, 41:523 – 547, 2005/05//.

Dekang Lin, Kenneth Ward Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New tools for Web-scale N-grams. In *LREC*, pages 2221–2227.

Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. 2013. Topic models + word alignment = A flexible framework for extracting bilingual dictionary from comparable corpus. In *CoNLL*, pages 212–221.

Max M. Louwerse. 2008. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 59(1):617–645.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

David M. Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*, pages 880–889.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *ACL*.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *EMNLP*, pages 1146–1157.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *CoNLL*, pages 1–7.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for Web search. In *WWW*, pages 373–374.

Daphna Shezaf and Ari Rappoport. 2010. Bilingual lexicon generation using non-aligned signatures. In *ACL*, pages 98–107.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *EMNLP*, pages 1423–1433.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *ACL*, pages 721–732.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *ACL*, pages 572–582.

Josef Sivic and Andrew Zisserman. 2003. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of ACL*, 2:207–218.

Nitish Srivastava and Ruslan Salakhutdinov. 2014. Multimodal learning with deep Boltzmann machines. *Journal of Machine Learning Research*, 15(1):2949–2980.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *EMNLP*, pages 24–36.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *CHI*, pages 319–326.

Ivan Vulić and Marie-Francine Moens. 2013a. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *NAACL*, pages 106–116.

Ivan Vulić and Marie-Francine Moens. 2013b. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *EMNLP*, pages 1613–1624.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *ACL*, pages 479–484.

Pengcheng Wu, Steven C. H. Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. 2013. Online multimodal deep similarity learning with application to image retrieval. In *ACM Multimedia*, pages 153–162.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328.