Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning

Angeliki Lazaridou Georgiana Dinu Marco Baroni
Center for Mind/Brain Sciences
University of Trento
{angeliki.lazaridou|georgiana.dinu|marco.baroni}@unitn.it

Abstract

Zero-shot methods in language, vision and other domains rely on a *cross-space mapping* function that projects vectors from the relevant feature space (e.g., visual-feature-based image representations) to a large semantic word space (induced in an unsupervised way from corpus data), where the entities of interest (e.g., objects images depict) are labeled with the words associated to the nearest neighbours of the mapped vectors. Zero-shot cross-space mapping methods hold great promise as a way to scale up annotation tasks well beyond the labels in the training data (e.g., recognizing objects that were never seen in training). However, the current performance of cross-space mapping functions is still quite low, so that the strategy is not yet usable in practical applications. In this paper, we explore some general properties, both theoretical and empirical, of the cross-space mapping function, and we build on them to propose better methods to estimate it. In this way, we attain large improvements over the state of the art, both in cross-linguistic (word translation) and cross-modal (image labeling) zero-shot experiments.

1 Introduction

In many supervised problems, the parameters of a classification function are estimated on $(\mathbf{x}, \mathrm{y})$ pairs, where $\mathbf{x}$ is a vector representing a training instance in some feature space, and y is the label assigned to the instance. For example, in image labeling $\mathbf{x}$ contains visual features extracted from a picture and y is the name of the object depicted in the picture (Grauman and Leibe, 2011). Since each label is treated as an unanalyzed primitive, this approach requires *ad-hoc* annotation for each label of interest, and it will not scale up to challenges where the potential label set is vast (for example, bilingual dictionary induction, where the label set corresponds to the full vocabulary of the target language).

Zero-shot methods (Palatucci et al., 2009) address the scalability problem by building on the observation that the labels of interest are often words (or longer linguistic expressions), which stand in a semantic similarity relation to each other. Moreover, distributional approaches allow us to estimate very large *semantic word spaces* in an efficient and unsupervised manner, using just unannotated text corpora as input (Turney and Pantel, 2010). Extensive evidence has shown that the similarity estimates obtained by representing words as vectors in such corpus-induced semantic spaces are extremely accurate (Baroni et al., 2014). Under the assumption that the domain of interest (e.g., objects in pictures, words in a source language) exhibits comparable similarity structure to that manifested in language, we can rephrase the learning task, from inducing multiple functions from the source feature space onto independent atomic labels, to that of estimating a single *cross-space mapping* function from vectors in the source feature space onto vectors for the corresponding word labels in distributional semantic space. The induced function can then also be applied to a data-point whose label was not used for training. The word corresponding to the nearest neighbour of the mapped vector in the latter space is used as the label of the data point. Zero-shot learning using distributional semantic spaces was originally proposed for brain signal decoding (Mitchell et al., 2008), but it has since been extensively applied in other domains, including image labeling (Frome et al., 2013; Lazaridou et al., 2014; Socher et al., 2013) and bilingual dictionary/phrase table induction (Dinu and Baroni, 2014; Mikolov et al.,

*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics
and the 7th International Joint Conference on Natural Language Processing*, pages 270–280,
Beijing, China, July 26-31, 2015.

2013a), the two applications we focus on here.

Effective zero-shot learning by cross-space mapping could get us through the manual annotation bottleneck that hampers many applications. However, in practice, the accuracy in label retrieval with current mapping methods is still too low for practical uses. In image labeling, when a search space of realistic size is considered, accuracy is just above 1% (which is still well above chance for large search spaces). In bilingual lexicon induction, accuracy reaches values around 30% (across words of varying frequency), which are definitely more encouraging, but still indicate that only 1 word in 3 will be translated correctly.

In this article, we look at some general properties of the linear cross-modal mapping function standardly used for zero-shot learning, in order to achieve a better understanding of its shortcomings, and improve its quality by devising methods to overcome them. First, when the mapping function is estimated with least-squares error techniques, we observe a systematic increase in *hubness* (Radovanović et al., 2010b), that is, in the tendency of some vectors ("hubs") to appear in the top neighbour lists of many test items. We connect hubness to least-squares estimation, and we show how it is greatly mitigated when the mapping function is estimated with a max-margin ranking loss instead. Still, switching to max-margin greatly improves accuracy in the cross-linguistic context, but not for vision-to-language mapping. In the cross-modal setting, we observe indeed a different problem, that we name (training instance) *pollution*: The neighbourhoods of mapped test items are "polluted" by the target vectors used in training. This suggests that cross-modal mapping suffers from overfitting issues, and consequently from poor generalization power. Taking inspiration from domain adaptation, which addresses similar generalization concerns, and self-learning, we propose a technique to augment the training data with automatically constructed examples that force the function to generalize better. Having shown the advantages of a ranking loss, our final contribution is the adaptation of some insights from the max-margin literature to our setting, in particular concerning the choice of negative examples. This leads to further accuracy improvements. We thus conclude the paper by reporting zero-shot performances in both cross-modal and cross-language settings that are well above the current state of the art. Table 1 provides a roadmap and summary of our results.

	cross-linguistic	cross-modal
former state of art	33.0	0.5
standard mapping	29.7	1.1
max-margin - §3	39.4	1.9
data augmentation - §4	NA	3.7
negative evidence - §5	40.2	5.6

Table 1: **Roadmap.** Proposed changes to cross-space mapping training and resulting percentage Precision @1 in our two experimental setups.

2 Experimental Setup

Cross-linguistic experiments In the cross-linguistic experiments, we learn a mapping from the semantic space of language A to the semantic space of language B, which can then be used for translating words outside the training set. Specifically, given the vector representation of a word in language A, we apply the mapping to obtain an estimate of the vector representation of its meaning in language B, returning the nearest neighbour of the mapped vector in the B space as candidate translation. We focus on translating from English to Italian and adopt the setup (word vectors, training and test data) of Dinu et al. (2015). For a set of 200K words, 300-dimensional vectors were built using the word2vec toolkit,[1] choosing the CBOW method.[2] CBOW, which learns to predict a target word from the ones surrounding it, produces state-of-the-art results in many linguistic tasks (Baroni et al., 2014). The word vectors were induced from corpora of 2.8 and 1.6 billion tokens, respectively, for English and Italian.[3] The train and test English-to-Italian translation pairs were extracted from a Europarl-derived dictionary (Tiedemann, 2012).[4] The 5K most frequent translation pairs were used for training, while the test set includes 1.5K English words equally split into 5 frequency bins. The search for the correct translation is performed in a semantic space of 200K

[1] https://code.google.com/p/word2vec/

[2] Other hyperparameters, which we adopted without further tuning, include a context window size of 5 words to either side of the target, setting the sub-sampling option to 1e-05 and estimating the probability of target words by negative sampling, drawing 10 samples from the noise distribution (Mikolov et al., 2013b).

[3] Corpus sources: http://wacky.sslmit.unibo.it, http://www.natcorp.ox.ac.uk

[4] http://opus.lingfil.uu.se/

Italian words.[5]

Cross-modal experiments In the cross-modal experiments, we induce a mapping from visual to linguistic space. Specifically, given an image, we apply the mapping to its visual vector representation to obtain an estimate of its representation in linguistic space, where the word associated to the nearest neighbour is retrieved as the image label. Similarly to translation pairs in the cross-linguistic setup, we create a list of "visual translation" pairs between images and their corresponding noun labels. Our starting point are the 5.1K labels in ImageNet (Deng et al., 2009) that occur at least 500 times in our English corpus and have concreteness score $\geq$5, according to Turney et al. (2011). For each label, we sample 100 pictures from its ImageNet entry, and associate each picture with the 4094-dimensional layer (fc7) at the top of the pre-trained convolutional neural network model of Krizhevsky et al. (2012), using the Caffe toolkit (Jia et al., 2014). The target word space is identical to the English space used in the cross-linguistic experiment. Finally, we use 75% of the labels (and the respective images) for training and the remaining 25% of the labels for testing.[6] From the 127.5K images corresponding to test labels, we sample 1K images as our test set. For zero-shot evaluation purposes, the search for the correct label is performed in the space of 5.1K possible labels, unless otherwise specified. However, when quantifying hubness and pollution, in order to have a setting comparable to that of cross-language mapping, we use the full set of 200K English words as search space.

Learning objectives We assume that we have cross-space "translation" pairs available for a set of $|Tr|$ items $(\mathbf{x}_i, \mathbf{y}_i) = \{\mathbf{x}_i \in \mathbb{R}^{d1}, \mathbf{y}_i \in \mathbb{R}^{d2}\}$. Moreover, following previous work, we assume that the mapping function is linear. For estimating its parameters $\mathbf{W} \in \mathbb{R}^{d1 \times d2}$, we consider two objectives. The first is L2-penalized least squares (**ridge**):

$$\hat{\mathbf{W}} = \underset{\mathbf{W} \in \mathbb{R}^{d1 \times d2}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\| + \lambda \|\mathbf{W}\|,$$

which has an analytical solution.

The second objective is a margin-based ranking loss (**max-margin**) similar in spirit to the one used in similar cross-modal experiments with WSABIE (Weston et al., 2011) and DeViSE (Frome et al., 2013). The loss for a given pair of training items $(\mathbf{x}_i, \mathbf{y}_i)$ and the corresponding mapping-based prediction $\hat{\mathbf{y}}_i = \mathbf{W}\mathbf{x}_i$ is defined as

$$\sum_{j \neq i}^{k} \max\{0, \gamma + dist(\hat{\mathbf{y}}_i, \mathbf{y}_i) - dist(\hat{\mathbf{y}}_i, \mathbf{y}_j)\},$$

where $dist$ is a distance measure, in our case the inverse cosine, and γ and k are tunable hyperparameters denoting the margin and the number of negative examples, respectively. Intuitively, the goal of the max-margin objective is to rank the correct translation $\mathbf{y}_i$ of $\mathbf{x}_i$ higher than any other possible translation $\mathbf{y}_j$. In theory, the summation in the equation could range over all possible labels, but in practice this is too expensive (e.g., in the cross-linguistic experiments the search space contains 200K candidate labels!), and it is usually computed over just a portion of the label space. In Weston et al. (2011), the authors propose an efficient way of selecting negative examples, in which they randomly sample, for each training item, labels from the complete set, and pick as negative sample the first label violating the margin. This guarantees that there will be exactly as many weight updates as training items. Another possibility is proposed in Mikolov et al. (2013b), where negative samples are picked from a non-item specific distribution (e.g., the uniform distribution).[7] For the experiments in Sections 3 and 4, we follow a more general setup in which the size of the margin and number of negative samples is tuned for each task. In this way, for a sufficiently large margin and number of negative samples, we increase the probability of performing a weight update per training item. We estimate the mapping parameters $\mathbf{W}$ with stochastic gradient descent and per-parameter learning rates tuned with Adagrad (Duchi et al., 2011). The tuning of hyperparameters γ and k is performed on a random 25% subset of the training data.

[5] Faithful to the zero-shot setup, in our experiments there is never any overlap between train and test words; however, to make the task more challenging, we include the train words in the search space, except where expressly indicated.

[6] At training time, we average the 100 vectors associated to a label into a single representation, to reduce training set size while minimizing information loss. At test time, as normally done, we present the model with single image visual vectors.

[7] The notion of negative samples is not unique to margin-based learning; in Mikolov et al. (2013b), the authors used it to efficiently estimate a word probability distribution.

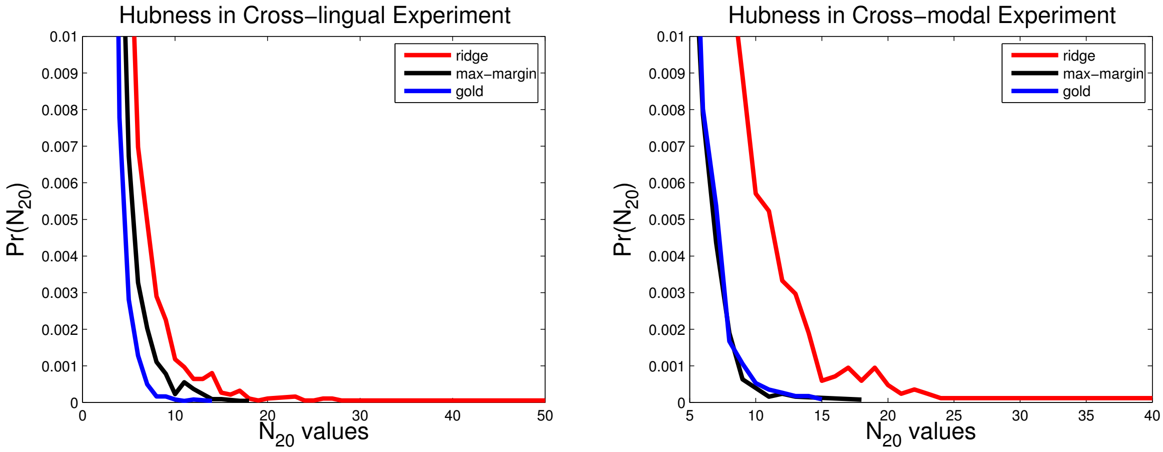

Figure 1: **Hubness distribution in cross-linguistic (left) and cross-modal (right) search spaces.** The hubness score (N_{20}) is computed on the top-20 neighbour lists of the test items, using their original (gold), ridge- or max-margin-mapped vectors as query terms.

3 Hubness

High-dimensional spaces are often affected by *hubness* (Radovanović et al., 2010b; Radovanović et al., 2010a), that is, they contain certain elements – *hubs* – that are near many other points in space without being similar to the latter in any meaningful way. As recently noted by Dinu et al. (2015), the hubness problem is greatly exacerbated when one looks at the nearest neighbours of vectors that have been mapped across spaces with **ridge**.[8] Given a set of query vectors with the corresponding top-k nearest neighbour lists, we can quantify the degree of hubness of an item in the search space (parameterized by k) by the number of lists in which it occurs. $N_k(y)$, the *hubness* at k of an item y, is computed as follows:

$$N_k(y) = |\{x \in T | y \in NN_k(x, S)\}|,$$

where S denotes the search space, T denotes the set of query items and $NN_k(x, S)$ denotes the k nearest neighbors of x in S.

Figure 1 reports N_{20} distributions across the cross-linguistic and cross-modal search spaces, using the respective test items as query vectors. The blue line shows the distributions for the "gold" vectors (that is, the vectors in the target space we would like to approximate). The red line shows the same distributions when neighbours are queried for the ridge-mapped test vectors (ignore black lines for now). In both spaces, when the query vectors are mapped, hubness increases dramatically. The largest hubs for the original test items occur in 15 neighbour lists or less. With the mapped vectors, we find hubs occurring in 40 lists or more. The figure also shows that, in both spaces, we observe more points with smaller but non-negligible N_{20} (e.g., around 10) when mapped vectors are queried. In both spaces, the difference in hubness is very significant according to a cross-tab test ($p<10^{-30}$). Finally, as Table 2 shows, the largest hubs are by no means terms that we might expect to occur as neighbours of many other items on semantic grounds (e.g., very general terms), but rather very specific and rare words whose high hubness cannot possibly be a genuine semantic property.

Cross-linguistic	Cross-modal
blockmonthon (50)	smilodon (40)
hashim (28)	pintle (33)
akayev (27)	knurled (27)
autogiustificazione (27)	handwheel (24)
limassol (26)	circlip (23)
regulars (26)	black-footed (23)
18 (25)	flatbread (22)

Table 2: **Top ridge hubs**, together with N_{20} scores. Note that cross-linguistic hubs are supposed to be *Italian* words.

Causes of hubness Why should the mapping function lead to an increase in hubness? We conjecture that this is due to an intrinsic property of least-squares estimation. Given the training ma-

[8] Dinu et al. (2015) observe, but do not attempt to understand hubness, as we do here. They propose to address it with methods to re-rank neighbour lists, which are less general and should be largely complementary to our effort to improve estimation of the cross-mapping function.

trices $\mathbf{X}$ and $\mathbf{Y}$, and the projection matrix $\mathbf{W}$ obtained by minimizing squared error, each column $\hat{\mathbf{y}}_{*,i}$ of $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}$ is the orthogonal projection of $\mathbf{y}_{*,i}$, the corresponding $\mathbf{Y}$ column onto the column space of $\mathbf{X}$ (Strang, 2003, Ch. 4). Consequently, $\mathbf{y}_{*,i} = \boldsymbol{\epsilon}_i + \hat{\mathbf{y}}_{*,i}$, where the $\boldsymbol{\epsilon}_i$ error vector is orthogonal to $\hat{\mathbf{y}}_{*,i}$. It follows that $||\mathbf{y}_{*,i}||^2 \geq= ||\hat{\mathbf{y}}_{*,i}||^2$. Since $\mathbf{y}_{*,i}$ and $\hat{\mathbf{y}}_{*,i}$ have equal means (because the error terms in $\boldsymbol{\epsilon}_i$ must sum to 0), it immediately follows from the squared length inequality that $\hat{\mathbf{y}}_{*,i}$ has lower or equal variance to $\mathbf{y}_{*,i}$. Since this holds for all columns of $\hat{\mathbf{Y}}$, it follows in turn that the set of mapped vectors in $\hat{\mathbf{Y}}$ has lower or equal variance to the corresponding set of original vectors in $\mathbf{Y}$. Coming back to hubness, a set of lower variance points (such as the mapped vectors) will result in higher hubness since the points will on average be closer to each other. The problem is likely to be further exacerbated by the property of least-squares to ignore relative distances between points (the objective only aims at making predicted and observed vectors look like each other),

Strictly, the theoretical result only holds for the training points. However, to the extent that the training set is representative of what will be encountered in the test set, it should also extend to test data (and if training and testing data are very different, the mapping function will generalize very poorly anyway). Moreover, the result holds for a pure least-squares solution, without the ridge L2 regularization term. Whether it also applies to ridge-based estimates will depend on the relative impact of the least-squares and L2 terms on the final solution (and it is not excluded that the L2 term might also independently reduce variance, of course). Empirically, we find that, indeed, lower variance also characterizes test vectors mapped with a ridge-estimated function.

Interestingly, in the literature on cross-space mapping we find that authors choose a different cost function than ridge, without motivating the choice. Socher et al. (2014) mention in passing that max-margin outperforms a least-squared-error cost for cross-modal mapping.

Max-margin as a solution to hubness Referring back to Figure 1, we see that when ridge estimation is replaced by max-margin (black line), there is a considerable decrease in hubness in both settings. This is directly reflected in a large increase in performance in our cross-linguistic (English-to-Italian) zero-shot task (left two columns of Table 3), with the largest improvement for the all important P@1 measure (equivalent to accuracy).[9] These results are well above the current best cross-language accuracy for cross-modal mapping without added orthographic cues (33%), attained by Mikolov et al. (2013a).[10] The absolute performance figures are low in the challenging cross-modal setting, but here too we observe a considerable improvement in accuracy when max-margin is applied. Indeed, we are already above the cross-modal zero-shot mapping state of the art for a search space of similar size (0.5% accuracy in Frome et al. (2013)). Still, the improvement over ridge (while present) is not as large for the less strict (higher ranks) performance scores.

Table 4 confirms that the improvement brought about by max-margin is indeed (at least partially) due to hubness reduction. A large proportion of vectors retrieved as top-1 predictions (translations/labels) are hubs when mapping is trained with ridge, but the proportion drops dramatically with max-margin. Still, more than 1/5 top predictions for cross-modal mapping with max-margin are hubs (vs. less than 1/10 for the original vectors). Now, the mathematical properties we reviewed above suggest that, for least-squares estimation, hubness is caused by general reduced variance of the space after mapping. Thus, hubs should be vectors that are near the mean of the space. The first row of Table 5 confirms that the hubs found in the neighbourhoods of ridge-mapped query terms are items that tend to be closer to the search space mean vector, and that this effect is radically reduced with max-margin estimation. However, the second row of the table shows another factor at play, that has a major role in the cross-modal setting, and it is only partially addressed by max-margin estimation: Namely, in vision-to-language mapping, there is a strong tendency for hubs (that, recall, have an important effect on performance, as they enter many nearest neighbour lists) to be close to a training data point.

[9] We have no realistic upper-bound estimate, but due to different word senses, synonymy, etc., it is certainly not 100%.

[10] Although the numbers are not fully comparable because of different language pairs and various methodological details, their method is essentially equivalent to our **ridge** approach we are clearly outperforming.

	Cross-linguistic		Cross-modal	
	ridge	**max-margin**	**ridge**	**max-margin**
P@1	29.7	38.4	1.1	1.9
P@5	44.2	54.2	4.8	5.4
P@10	49.1	60.4	7.9	9.0

Table 3: **Ridge vs. max-margin in zero-shot experiments.** Precision @N results cross-linguistically (test items: 1.5K, search space: 200K) and cross-modally (test items: 1K, search space: 5.1K).

Cross-linguistic			Cross-modal		
ridge	**max-margin**	**gold**	**ridge**	**max-margin**	**gold**
19.6	9.8	0.6	55.8	21.6	7.8

Table 4: **Hubs as top predictions**. Percentage of top-1 neighbours of test vectors in zero-shot experiments of Table 3 with $\mathrm{N}_{20} > 5$.

	Cross-linguistic		Cross-modal	
cosine with	**ridge**	**max-margin**	**ridge**	**max-margin**
full-space mean	0.21	0.06	0.13	-0.01
training point	0.15	0.12	0.34	0.24

Table 5: **Properties of hubs.** Spearman ρ of N_{20} scores with cosines to mean vector of full search space (top) and nearest training item (bottom), across all search space elements. All correlations significant (p<0.001) except cross-modal **max-margin** hubness/full-space mean.

4 Pollution

The quantitative results and post-hoc analysis of hubs in Section 3 suggest that cross-modal mapping is facing a serious generalization problem. To get a better grasp of the phenomenon, we define a binary measure of *(training data) pollution* for a queried item x and parameterized by k, such that pollution is 1 if x has a (target) training item y among its k nearest neighbours, 0 otherwise. Formally:

$$\mathrm{N}^{\mathrm{pol}}_{\mathrm{k,S}}(\mathrm{x}) = [\![\exists \mathrm{y} \in \mathbf{Y}^{Tr} : y \in \mathrm{NN}_{\mathrm{k,S}}(\mathrm{x})]\!],$$

where $\mathbf{Y}^{Tr}$ is the matrix of target vectors used in training, $\mathrm{NN}_{\mathrm{k,S}}(\mathrm{y})$ denotes the top k neighbors of y in search space S, and $[\![\mathrm{z}]\!]$ is an indicator function.[11]

The average pollution $\mathrm{N}^{\mathrm{pol}}_{1,\mathrm{S}}$ of all test items in the cross-modal experiment, when $|\mathrm{S}|{=}200K$ is 18%, which indicates that in 1/5 of cases the returned label is that of a training point. The equivalent statistic in the cross-linguistic experiment drops to 8.7% (words tend to be more varied than the set of concrete, imageable concepts used for image annotation tasks, and so the cross-linguistic training set is probably less uniform than the one used in the vision-to-language setting).

The real extent of the generalization problem in the cross-modal setup becomes more obvious if we restrict the search space to labels effectively associated to an image in our data set $(|\mathrm{S}|{=}5.1K)$. In this case, the average pollution $\mathrm{N}^{\mathrm{pol}}_{1,\mathrm{S}}$ across all test items jumps to 88%, that is, the vast majority of test images are annotated with a label coming from the training data. Clearly, there is a serious problem of overfitting to the training subspace. While we came to this observation by inspecting the properties of hubs, other work in zero-shot for image labeling has indirectly noted the same. Frome et al. (2013) empirically showed that the performance of the system is higher when removing training labels from the search space, while Norouzi et al. (2014) proposed a zero-shot method that avoids explicit cross-modal mapping.

Adapting to the full search space by data augmentation High training-data pollution indicates that cross-modal mapping does not generalize well beyond the kind of data points it encountered in learning. This is a special case of the *dataset bias* problem (Torralba and Efros, 2011) and, given that the latter has been addressed as a domain adaptation problem (Gong et al., 2012; Donahue et al., 2013), we adopt here a similar view. Self-training has been successfully used for domain adaptation in NLP, e.g., in syntactic parsing. Given the limited amount of syntactically annotated data coming from monotonous sources (e.g., the Wall Street Journal), parsers show a big drop in performance when applied to different domains (e.g., reviews), since training and test domains differ dramatically, thus affecting their generalization performance. In a nutshell, the idea behind self-training (McClosky et al., 2006; Reichart and Rappoport, 2007) is to use manually annotated data $(\mathrm{x}^{\mathrm{A}}_{\mathrm{i}}, .., \mathrm{x}^{\mathrm{A}}_{\mathrm{N}}, \mathrm{y}^{\mathrm{A}}_{\mathrm{i}}, .., \mathrm{y}^{\mathrm{A}}_{\mathrm{N}})$ from domain A to train a parser, feed the trained parser with data $\mathrm{x}^{\mathrm{B}}_{\mathrm{i}}, .., \mathrm{x}^{\mathrm{B}}_{\mathrm{K}}$ from domain B in order to obtain their automated annotations $\hat{\mathrm{y}}^{\mathrm{B}}_{\mathrm{i}}, .., \hat{\mathrm{y}}^{\mathrm{B}}_{\mathrm{K}}$ and then retrain the parser

[11] Pollution is of course an effect of overfitting, but we use this more specific term to refer to the tendency of training vectors to "pollute" nearest neighbour lists of mapped vectors.

dolphin	tarantula	highland
whale	anteater	whisky
orca	arachnid	lowland
porpoise	spider	bagpipe
cetacean	opossum	glen
shark	scorpion	distillery

Table 6: **Visual chimeras** for *dolphin*, *tarantula* and *highland*.

	none	**chimera-5**	**chimera-10**
P@1	1.9	3.7	3.2
P@5	5.4	10.9	10.5
P@10	9.0	15.8	15.9

Table 7: **Cross-modal zero-shot experiment with data augmentation.** Labeling precision @N with no data augmentation (**none**) and when using top 5 (**chimera-5**) and top 10 (**chimera-10**) nearest neighbors from training set of each item in the search space to build the corresponding chimeras (1K test items, 5.1K search space).

with a combination of "clean" data from domain A and "noisy" data from domain B.

In our setup, self-training would be applied by labeling a larger set of images with a cross-modal mapping function estimated on the initial training data, and then using both sources of labeled data to retrain the function. Although the idea of self-training for inducing cross-modal mapping functions is appealing, especially given the vast amount of unlabeled data available out there, the very low performance of current cross-modal mapping functions makes the effort questionable. We would like to exploit unannotated data representative of the search space, *without* relying on the output of cross-modal mapping for their annotation. One way to achieve this is to use *data augmentation* techniques that are representative of the search space. Data augmentation is popular in computer vision, where it is performed (among others) by data jittering, visual sampling or image perturbations. It has proven beneficial for both "deep" (Krizhevsky et al., 2012; Zeiler and Fergus, 2014) and "shallow" (Chatfield et al., 2014) systems, and it was recently introduced to NLP tasks (Zhang and LeCun, 2015).

Specifically, in order to train the mapping function using both annotated data and points that are representative of the full search space, we rely on a form of data augmentation that we call *visual chimera* creation. For every item $\mathrm{y_i} \notin \mathbf{Y}^{Tr}$ in the search space S, we use linguistic similarity as a proxy of visual similarity, and create its *visual vector* $\hat{\mathrm{x}}_\mathrm{i}$ by averaging the visual vectors corresponding to the nearest words in language space that do occur as labels in the training set. Table 6 presents some examples of visual chimeras. For $\mathrm{y_i}$=*dolphin*, the visual vectors of other cetacean mammals are averaged to create the chimera $\hat{\mathrm{x}}_\mathrm{i}$. Since linguistic similarity is not always determined by visual factors, the method also produces noisy data points. For $\mathrm{y_i}$=*tarantula*, *opossums* enter the picture, while for $\mathrm{y_i}$=*highland* images of "topically" similar concepts are used (e.g., *bagpipe*).

Table 7 reports cross-modal zero-shot labeling when training with **max-margin** and data augmentation. We experiment with visual chimeras constructed using 5 vs. 10 nearest neighbours. While the examples above suggest that the process injects some noise in the training data, we also observe a decrease of pollution $\mathrm{N}^{\mathrm{pol}}_{1,\mathrm{S}}$ from 88% when using the "clean" training data, to 71% and 73% when expanding them with chimeras (for **chimera-5** and **chimera-10**, respectively). Reflecting this drop in pollution, we see large improvements in precision at all levels, when chimeras are used (no big differences between 5 or 10 neighbours).

The improvements brought about by the chimera method are robust. First, Table 8 reports performance when the search space excludes the training labels, showing that data augmentation is beneficial beyond mitigating the bias in favor of the latter. In this setup, **chimera-5** is clearly outperforming **chimera-10** (longer neighbour lists will include more noise), and we focus on it from here on.

All experiments up to here follow the standard cross-modal zero-shot protocol, in which the search space is given by the union of the test and training labels, or a subset thereof. Next, we make the task more challenging by increasing it with 1K extra elements acting as distractors. The distractors are either randomly sampled from our usual 200K English word space, or, in the most challenging scenario, picked among those words, in the same space, that are among the top-5 near-

	none	chimera-5	chimera-10
P@1	6.7	9.3	8.3
P@5	21.7	25.2	21.3
P@10	29.9	34.3	29.7

Table 8: **Cross-modal zero-shot experiment with data augmentation, disjoint train/search spaces.** Same setup as Table 8, but search space excludes training elements (1K test items, 1K search space).

	random		related	
	none	**chimera-5**	**none**	**chimera-5**
P@1	0.8	3.3	1.9	2.8
P@5	5.3	9.0	4.8	8.8
P@10	8.8	13.3	7.9	12.6

Table 9: **Cross-modal zero-shot experiment with data augmentation, enlarged search space.** Labeling precision @N with no data augmentation (**none**) and when using top 5 (**chimera-5**) nearest neighbors from training set of each item in the search space to build the corresponding chimeras. Test items: 1K. Search space: 5.1K+1K extra distractors from a 200K word space, either randomly picked (*random*), or *related* to the training items.

est neighbours of a training element. Again, we create one visual chimera for each label in the search space. Results are presented in Table 9. As expected, performance is negatively affected with both plain and data-augmented models, but the latter is still better in absolute terms. While **chimera-5** undergoes a larger drop when the search contains many elements similar to the training data ("related" column), which is explained by the fact that visual chimeras will often include the distractor items of this setup, it appears to be more resistant against random labels, which in many cases are words that bear no resemblance to the training data (e.g., *naushad*, *yamato*, *13-14*). The picture when using no data augmentation is exactly the opposite, with the model being more harmed, at P@1, by the random labels.

Finally, Table 10 presents results in the cross-linguistic setup, when applying the same data augmentation technique. In this case, we augment the 5K training elements with 11.5K chimeras, for the 1.5K test elements and 10K randomly sampled distractors. For these 11.5K elements, we associate their Italian (target space) label y_i with a

	none	chimera-5
P@1	38.4	31.1
P@5	54.2	46.1
P@10	60.4	51.3

Table 10: **Cross-linguistic zero-shot experiment with data augmentation**. Translation precision @N when learning with **max-margin** and no data augmentation (**none**) or data augmentation using the top 5 (**chimera-5**) nearest neighbors of 11.5K items in the 200K-word search space (1.5K test items).

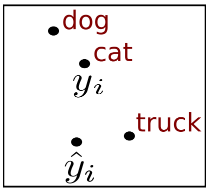

Figure 2: **Looking for intruders.** We pick *truck* rather than *dog* as negative example for *cat*.

"pseudo-translation" vector $\hat{x}_i$ obtained by averaging the vectors of the English (source space) translations of the nearest Italian words to y_i included in the training set. Results, in Table 10, show that in this case our data augmentation method is actually hampering performance. We saw that pollution affects the cross-linguistic setup much less than it affects the cross-modal one, and we conjecture that, consequently, in the translation task, there is not a large-enough generalization gain to make up for the extra noise introduced by augmentation.

5 Picking informative negative examples

An interesting feature of the ranking max-margin objective lies in its active use of negative examples. While previous work in cross-space mapping has paid little attention to the properties that negative samples should possess, this has not gone unnoticed in the NLP literature on structured prediction tasks. Smith and Eisner (2005) propose a contrastive estimation framework in the context of POS-tagging, in which positive evidence derived from gold sentence annotations is extended with negative evidence derived by various *neighbourhood functions* that corrupt the data in particular ways (e.g., by deleting 1 word).

Having shown the effectiveness of max-margin estimation in the previous sections, we now take

	Cross-linguistic		Cross-modal	
	random	**intruder**	**random**	**intruder**
P@1	38.4	40.2	3.7	5.6
P@5	54.2	55.5	10.9	12.4
P@10	60.4	61.8	15.8	17.8

Table 11: **Random vs. intruding negative examples.** Zero-shot precision @N results when cross-space function is estimated using **max-margin** with random or "intruder" negative examples, cross-linguistically (test items: 1.5K, search space: 200K) and cross-modally (test items: 1K, search space: 5.1K).

a first step towards engineering the negative evidence exploited by this method, in the context of inducing cross-space mapping functions. In particular, our idea is that, given a training instance $\mathrm{x_i}$, an informative negative example would be *near* the mapped vector $\mathrm{\hat{y}_i}$, but *far* from the actual gold target space vector $\mathrm{y_i}$. Intuitively, such "intruders" correspond to cases where the mapping function is getting the predictions seriously wrong, and thus they should be very informative in "correcting" the function mapping trajectories. This can seen as a vector-space interpretation of the *max-loss* update protocol (Crammer et al., 2006) that picks negative samples expected to harm performance more. Figure 2 illustrates the idea with a cartoon example. If *cat* is the gold target vector $\mathrm{y_i}$ and $\mathrm{\hat{y}_i}$ the corresponding mapped vector, then we are going to pick *truck* as negative example, since it is an intruder (near the mapped vector, far from the gold one).

More formally, at each step of stochastic gradient descent, given a source space vector $\mathrm{x_i}$, its target gold label/translation $\mathrm{y_i}$ in $\mathbf{Y}^{Tr}$ and the mapped vector $\mathrm{\hat{y}_i}$, we compute $\mathrm{s_j} = \cos(\mathrm{\hat{y}_i}, \mathrm{y_j}) - \cos(\mathrm{y_i}, \mathrm{y_j})$, for all vectors $\mathrm{y_j}$ in $\mathbf{Y}^{Tr}$ s.t. $j \neq i$, and pick as negative example for $\mathrm{x_i}$ the vector with the largest $\mathrm{s_j}$.

Table 11 presents zero-shot mapping results when intruding negative examples are used for max-margin estimation. For cross-modal mapping, we apply data augmentation as described in the previous section. While the absolute performance increase is relatively small (less than 2% in both setups), it is consistent. Furthermore, the proposed protocol results in lower $\mathrm{N^{pol}_{1,S}}$ pollution in the cross-modal setup (from 71% to 63%). Finally, we observe that the learning behaviour of the two protocols (intruders vs. random) is different; the **intruder** approach is already achieving good performance after just few training epochs, since it can rely on more informative negative samples (see Figure 3).

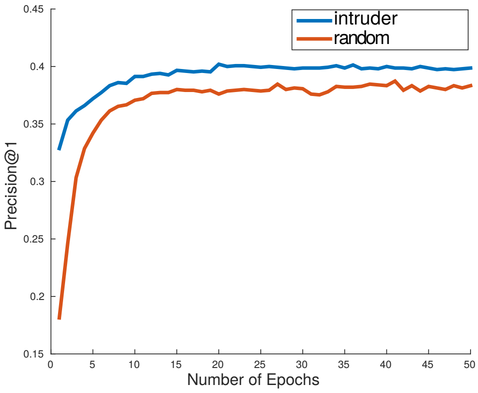

Figure 3: **Learning curve with random or intruding negative samples** in the cross-linguistic experiment.

6 Conclusion

We have considered some general mathematical and empirical properties of linear cross-space mapping functions, suggesting one well-known (max-margin estimation) and two new (chimera augmentation and "intruder" negative sample adjustment) methods to improve their performance. With them, we achieve results well above the state of the art in both the cross-linguistic and the cross-modal setting. Both chimera and the intruder methods are flexible, and we plan to explore them further in future research. In particular, we want to devise more semantically-motivated methods to select chimera components and negative samples.

Acknowledgments

We thank Adam Liska, Yoav Goldberg and the anonymous reviewers for useful comments. We acknowledge ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247, Baltimore, MD.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585.

Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, Miami Beach, FL.

Georgiana Dinu and Marco Baroni. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of ACL*, pages 624–633, Baltimore, MD.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of ICLR Workshop Track*, San Diego, CA. Published online: `http://www.iclr.cc/doku.php?id=iclr2015:main`.

Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. 2013. Semi-supervised domain adaptation with instance constraints. In *In Proceedings of CVPR*, pages 668–675.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A deep visual-semantic embedding model. In *Proceedings of NIPS*, pages 2121–2129, Lake Tahoe, NV.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *In Proceedings of CVPR*, pages 2066–2073.

Kristen Grauman and Bastian Leibe. 2011. *Visual Object Recognition*. Morgan & Claypool, San Francisco.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1097–1105, Lake Tahoe, Nevada.

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

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of HLT-NAACL*, pages 152–159.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL*, pages 746–751, Atlanta, Georgia.

Tom Mitchell, Svetlana Shinkareva, Andrew Carlson, Kai-Min Chang, Vincente Malave, Robert Mason, and Marcel Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191–1195.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. 2014. Zero-shot learning by convex combination of semantic embeddings. In *Proceedings of ICLR*.

Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom Mitchell. 2009. Zero-shot learning with semantic output codes. In *Proceedings of NIPS*, pages 1410–1418, Vancouver, Canada.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531.

Milo˘s Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of SIGIR*, pages 186–193, Geneva, Switzerland.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *In Proceedings of ACL*, pages 616–623.

Noah A Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of ACL*, pages 354–362.

Richard Socher, Milind Ganjoo, Christopher Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Proceedings of NIPS*, pages 935–943, Lake Tahoe, NV.

Richard Socher, Quoc Le, Christopher Manning, and Andrew Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Gilbert Strang. 2003. *Introduction to linear algebra, 3d edition*. Wellesley-Cambridge Press, Wellesley, MA.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC*, pages 2214–2218.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *In Proceedings of CVPR*, pages 1521–1528.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pages 680–690, Edinburgh, UK.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of IJCAI*, pages 2764–2770.

Matthew Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of ECCV (Part 1)*, pages 818–833, Zurich, Switzerland.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scrath. *arXiv preprint arXiv:1502.01710*.