Visual Grounding in Video for Unsupervised Word Translation

Gunnar A. Sigurdsson[3*] Jean-Baptiste Alayrac[1] Aida Nematzadeh[1] Lucas Smaira[1]
Mateusz Malinowski[1] João Carreira[1] Phil Blunsom[1,2] Andrew Zisserman[1,2]

[1]DeepMind
[2]Department of Engineering Science, University of Oxford
[3]Carnegie Mellon University
github.com/gsig/visual-grounding

Abstract

There are thousands of actively spoken languages on Earth, but a single visual world. Grounding in this visual world has the potential to bridge the gap between all these languages. Our goal is to use visual grounding to improve unsupervised word mapping between languages. The key idea is to establish a common visual representation between two languages by learning embeddings from unpaired *instructional videos narrated in the native language. Given this shared embedding we demonstrate that (i) we can map words between the languages, particularly the 'visual' words; (ii) that the shared embedding provides a good initialization for existing unsupervised text-based word translation techniques, forming the basis for our proposed hybrid visual-text mapping algorithm, MUVE; and (iii) our approach achieves superior performance by addressing the shortcomings of text-based methods – it is more robust, handles datasets with less commonality, and is applicable to low-resource languages. We apply these methods to translate words from English to French, Korean, and Japanese – all without any parallel corpora and simply by watching many videos of people speaking while doing things.*

1. Introduction

Children can learn multiple languages by merely observing their environment and interacting with others, without any explicit supervision or instruction; multilingual children do not hear a sentence and its translation simultaneously, and they do not hear a sentence in multiple languages while observing the same situation [20]. Instead, they can leverage visual similarity across situations: what they observe while hearing "the dog is eating" on Monday is similar to what they see as they hear "le chien mange" on Friday.

Figure 1: Across the world, there are many different ways to refer to 🐶. But in the visual domain, a 🐶 is simply a 🐶 everywhere on Earth. In this work, we leverage this observation to learn to translate words in different languages without *any* paired bilingual data.

We take a first step towards building an unsupervised multimodal translation system by relating the machine translation task to the way children learn multiple languages: we expose the system to videos of people from different countries performing a task while explaining what they are doing in their native languages. There are many such videos in YouTube: for example, we can learn how to squeeze orange juice by watching Korean or English videos. Instructional videos tend to look visually similar and the underlying concepts being spoken are often the same. We obtained a large number of such videos and the corresponding

[*] Work done while Gunnar was an intern at DeepMind.

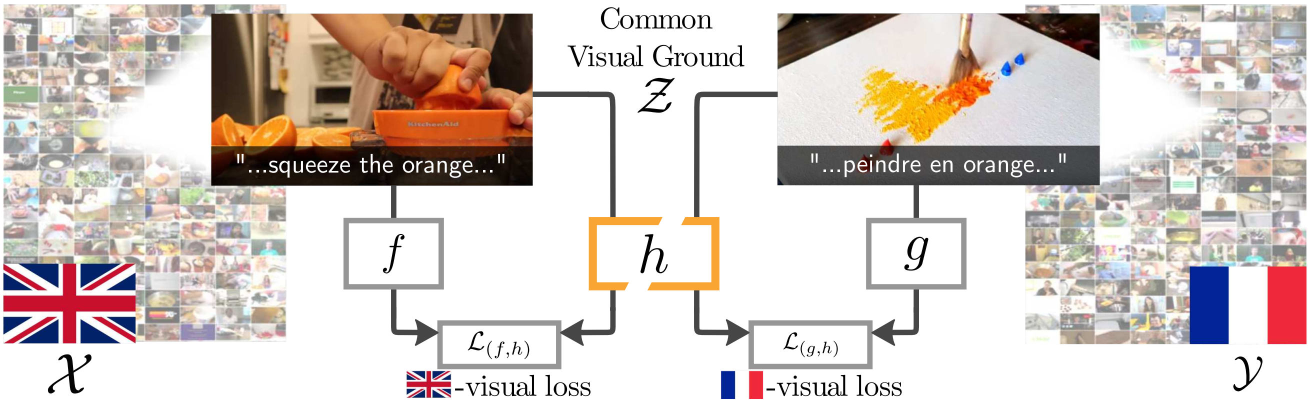


Figure 2: We build on recent advances in video modeling and train an unsupervised system that learns to translate words in multiple languages by grounding the language in video and without any paired data. ("peindre en orange"="painting in orange".)

subtitles using automatic speech recognition, extending the recent procedure of [35] to multiple languages.

Working with this data introduces various challenges. First of all, despite significant recent progress, visual understanding in videos is far from solved – even with the state-of-the-art models, clustering similar activities is not easy. Additionally, and in contrast to manually-captioned datasets where words tend to describe the scene, in instructional videos the words correspond to what the instructors are saying. While performing a task, the instructors often talk about random topics (such as subscriber counts and audience interaction) that do not have any visual relevance.

This paper demonstrates that, despite these challenges, a shared visual representation can facilitate the mapping of different languages at the word level. As illustrated in Fig. 2, we propose a model that maps two languages through the visual domain (videos). For English and French, the model correctly translates 28.0% and 45.3% of common words and visual words, all by only watching videos. For comparison, a retrieval-based baseline (without sharing the visual representation) achieves 12.5% and 18.6% for common words and visual words.

Moreover, we show that our model is more robust than the state-of-the-art unsupervised *text-based* word mapping models which exploit co-occurrence statistics [4, 10], in terms of sensitivity to (a) the degree to which the two languages differ (*e.g.*, English is more similar to French than Korean), (b) the dissimilarity of the training corpora of the two languages (*e.g.*, English and French Wikipedia are highly similar), and (c) the amount of training data. Finally, we show that the combination approach (with text-based approaches) is reliable for a large variety of tasks. For example, when the training corpora in French and English are dissimilar (instructional videos in French and Wikipedia in English), our method achieves a 32.6% recall while that of the text-based ones is less than 0.5%.

Contributions. The contributions are threefold. **(i)** We propose a method to map languages through the visual domain using *only* unpaired instructional videos, **(ii)** we demonstrate that our method is effective at connecting words in different languages through vision in an unsupervised manner, and finally **(iii)** we show that our method can serve as a good initialization for existing word mapping techniques addressing many shortcomings of *text-based* methods.

2. Prior Work

Bilingual child language acquisition. An open question in the field of bilingual language acquisition is to what extent the systems and representations learned for each language are shared. This sharing can happen for different aspects of language such as grammar, morphology, or the conceptual representations [11, 19]. For example, bilingual children eventually learn that both "chien" and "dog" refer to the actual animal dog, but whether and when this representation is shared is a matter of debate. We explore whether sharing the conceptual (visual) representation improve the quality of word translation for different languages.

Unsupervised text-based word alignment. Words often occur in the same context in different languages – in both English and French, "dog", "catch", and "ball" co-occur together. Previous work has used this insight to align the embedding space of different languages and use the aligned space to translate words from one language to another language [31, 36]. Earlier work used various degrees of supervision through ground-truth dictionaries or heuristics [4, 26, 42]; recently, fully unsupervised approaches achieved a similar performance on word alignment for different language pairs without any supervision [6, 10]. However, because these methods take advantage of the similarity between both the language pairs and their training corpora, they are not robust when the languages (or their training corpora) are very different [5, 43].

Vision and language. There is a growing interest in combining methods developed in computer vision and natural language processing to solve more challenging problems at the intersection of these fields [2, 13, 25, 27, 30, 33, 41, 45]. Grounding language is at the core of the interest of these two communities. It also has a long tradition in symbolic artificial intelligence, where "meaningless symbols cannot be grounded in anything but other meaningless symbols" [23]. The same problem of assigning meaning to symbols has been a fruitful research direction in computer vision. Early work explored weak supervision and the correspondence problem between text annotations and image regions [7, 14], with more modern approaches exploring joint image-text word embeddings [17], or building a language conditioned attention map over the images in caption generation, visual question answering and text-based retrieval [3, 12, 24, 32, 38, 39, 45, 48, 50]. Of particular interest, recent work has focused on multimodal and multilingual settings such as producing captions in many languages, visual-guided translations [8, 16, 44, 46], or bilingual visual question answering [18]. However, these use a paired corpora, *i.e.* same video or images are associated with captions in multiple languages [46]. Obtaining paired corpora in several languages is expensive, and does not scale.

Instructional videos. In this work, we rely on instructional videos [1, 40, 49] since they can be obtained at scale *without any manual annotation* [35]: they consist of YouTube videos and their associated narrations which is generated using automatic speech recognition (ASR). We propose to use instructional videos in different languages to show that we can translate words by only watching and listening to people performing various tasks.

3. Unsupervised Multilingual Learning

We describe our approach for unsupervised multilingual word alignment through grounding in the visual domain $\mathcal{Z}$. Our method is *unsupervised* in that it learns the correspondences between two languages $\mathcal{X}$ and $\mathcal{Y}$ (*e.g.* English and French) without *any* parallel (paired) corpora. Instead, we are given two distinct collections of instructional videos, *i.e.* n videos narrated with language $\mathcal{X}$ and another m *different* videos with language $\mathcal{Y}$. Equipped with this, our goal is to learn to map languages $\mathcal{X}$ and $\mathcal{Y}$ by leveraging the shared visual modality $\mathcal{Z}$ – the videos. We evaluate this ability in terms of the accuracy of word translation, *i.e.* how well the vocabulary in one language can be mapped to the other one.

Mapping languages through instructional videos is challenging: first, learning video-text embeddings from instructional videos is difficult as the speech in these videos is only *loosely* related to the scene.* Second, in multilingual setting, such errors compound since both languages have this low video-text relevance; moreover, visually similar videos may not be semantically similar.

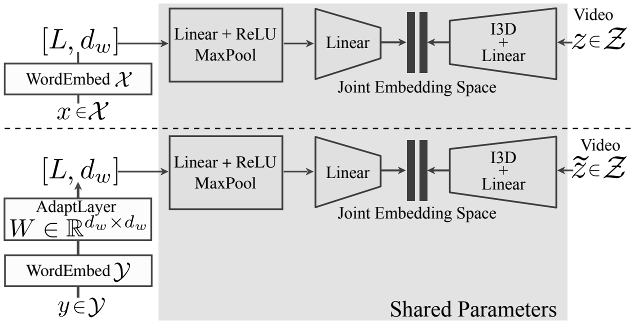


Figure 3: Details of the three encoders: one for language $\mathcal{X}$, one for language $\mathcal{Y}$, and one for videos $\mathcal{Z}$. Coupling of the two languages is obtained by sharing parts of the model (shaded region).

This challenge *cannot* be addressed by using the similarity of videos to construct a parallel text corpora (see Fig. 4). Instead, following Miech *et al.* [35], we learn a joint (monolingual) video-text embedding space from instructional videos. We extend the training strategy to the multilingual case by defining the following objective:

$$\min_{f,g,h} \underbrace{\mathcal{L}_{(f,h)}(\mathcal{X} \times \mathcal{Z})}_{\text{Language } \mathcal{X} \text{ and vision}} + \underbrace{\mathcal{L}_{(g,h)}(\mathcal{Y} \times \mathcal{Z})}_{\text{Language } \mathcal{Y} \text{ and vision}}, \tag{1}$$

where $\mathcal{L}$ is a metric-learning loss between text and video embeddings [34]. The parameters f, g, and h define the embedding functions of the language $\mathcal{X}$, language $\mathcal{Y}$, and the video domain $\mathcal{Z}$, respectively. The idea is that *sharing* the visual encoder h across the two languages is crucial to align the two languages $\mathcal{X}$ and $\mathcal{Y}$.

Next, we describe the proposed approach (Eq. (1)) in detail. Sec. 3.1 explains our choice of embedding models f, g and h. Sec. 3.2 defines the loss function $\mathcal{L}$. Finally, in Sec. 3.3, we explain how our initial model can be used to improve text-based word mapping techniques.

3.1. Multilingual Visual Embedding: Architecture

An illustration of our architecture is given in Fig. 3.

Input to the model. We represent sentences as a fixed-length sequence of integers, *i.e.* $\mathcal{X}$ and $\mathcal{Y}$ are of the form $\{1, \dots, K\}^L$ where K and L are the vocabulary size and sentence length, respectively. On average, sentences consist of 10 words. Videos are in pixel space: $\mathcal{Z} = \mathbb{R}^{T \times H \times W \times 3}$, where T is the number of frames in the video clips (here 32 frames at 10 FPS); H and W are the height and width of the video respectively, with 3 RGB channels.

Text encoders. The text encoder f in language $\mathcal{X}$, following [35], consists of: (i) a word embedding layer that takes as input a sequence composed of L tokens and outputs L vectors of dimension d_w, (ii) a position-wise fully

*e.g. only 50% of captions and videos in HowTo100M are related [35].

connected feed-forward layer followed by max pooling over the words to generate a single d_i-dimensional vector for the whole sequence, and finally (iii) a linear layer to map the intermediate representation to the joint embedding space $\mathbb{R}^d$.

For the text encoder g in language $\mathcal{Y}$, we share model weights across languages [22, 28]. Specifically, we share the weights of the feed forward layers and the last linear layer between f and g. To input different languages to the shared layers, we add a linear layer, referred to as the *AdaptLayer*, after the word embedding layer in language $\mathcal{Y}$.

Intuitively, the role of the *AdaptLayer* is to transform the word embedding space of language $\mathcal{Y}$ such that word embeddings in language $\mathcal{Y}$ become as similar as possible to the word embeddings in language $\mathcal{X}$. Then, the rest of the network can be shared, and yet preserve the monolingual properties of the word embeddings if needed. Our architecture is *symmetric*, yet AdaptLayer appears asymmetric. However, the orthogonality constraint used in AdaptLayer enforces *symmetry* of the overall model. Indeed, a symmetric case with each language equipped with AdaptLayer is equivalent to our case; this can be shown by multiplying the AdaptLayer for X and Y by the inverse of the AdaptLayer for X, ending up with a single AdaptLayer for Y.

Video encoder. For the video encoder, we use the standard I3D [9] model followed by a linear layer that maps the output into the joint embedding space.

3.2. The Base Model: Training and Inference

Training data. We are given a set of n videos narrated in language $\mathcal{X}$: $\{(x_i, z_i)\}_{i=1}^n$ and a set of m *different* videos narrated in language $\mathcal{Y}$: $\{(y_j, \tilde{z}_j)\}_{j=1}^m$. Note that there is *no* overlap in videos in the first and second set, *i.e.* we do not have access to *paired* bilingual data.

Training objective. The first term $\mathcal{L}_{(f,h)}$ in our objective function Eq. (1) is defined as follows:

$$\mathcal{L}_{(f,h)}\left(\{(x_i, z_i)\}_{i=1}^n\right) = \sum_i -\log \text{NCE}\left(f(x_i), h(z_i)\right), \tag{2}$$

where NCE corresponds to the noise contrastive estimation [21, 29] discriminative operator:

$$\text{NCE}(x, z) = \frac{e^{f(x)^\top h(z)}}{e^{f(x)^\top h(z)} + \sum\limits_{(x',z')\sim\mathcal{N}} e^{f(x')^\top h(z')}}, \tag{3}$$

where $\mathcal{N}$ is a set of negative pairs used to enforce that video and narration that co-occur in the data are close in the space and those that do not are far. In this work, the negatives are x and z paired with other x' and z' chosen uniformly at random from the training set $\mathcal{X}$, following [34]. In practice, each training batch includes clips from either language, and the negatives for each element in the NCE loss are the other elements from the batch in the same language. $\mathcal{L}_{(g,h)}$ in Eq. (1) has the same form, except with g and $\{(y_j, \tilde{z}_j)\}_{j=1}^m$.

Inference. Because we use the same visual encoder h for the two languages, we can assume that the outputs of the language encoders f and g are in the same space. After training our model with the joint loss in Eq. (1), we can directly map the first language to the second one; for a given $x \in \mathcal{X}$, we find $y \in \mathcal{Y}$ for which the embedding $g(y)$ has the smallest cosine distance to $f(x)$.

3.3. MUVE: Improving Unsupervised Translation

In this section, we explain how the Base Model can be used to improve a state-of-the-art *text-based* word translation technique.

Text-based word translation. It has been shown that distributed representations of words (*e.g.* Word2Vec [37]) share similarities across languages. In particular, Mikolov *et al.* [36] show that a word embedding matrix in a target language can be approximated by simply applying a *linear* mapping on a word embedding matrix in a different source language. To recover that linear mapping, Mikolov *et al.* [36] employ a supervised method where, given a subset of 5,000 pairs of words in the two languages, the mapping is learned by minimizing a L_2 distance between the word embeddings of the source language and the linearly mapped word embeddings of the target language. Xing et al. [47] show that the results can be improved by adding an orthogonality constraint. This can be done in closed form with the Procrustes algorithm (see [10] for details).

The unsupervised *MUSE* method. Conneau et al. [10] propose the *MUSE* approach that, in contrast to the method of Mikolov *et al.* [36], does not require any supervised pairs of words. *MUSE* has three main steps **(i)** finding an initial linear mapping via an adversarial approach, then **(ii)** refining the mapping with the Procrustes algorithm, and finally **(iii)** normalizing the distances using the local neighborhood.

MUVE: aligning words through vision. As explained in Section 3.1, the intuition behind the linear *AdaptLayer* (see Fig. 3) is to map word embeddings from language $\mathcal{Y}$ to a similar vector space as word embeddings from language $\mathcal{X}$ before being fed to the shared layers. Given this, we propose to *replace* the step **(i)** (adversarial initialization) of the *MUSE* algorithm by the *AdaptLayer* of our Base Model, after training it on videos. We call that method MUVE for *Multilingual Unsupervised Visual Embeddings*. To further improve the performance, we follow the observation of [47] by adding to the objective (1) an orthogonal penalty $\|WW^\top - I\|_F^2$ on the weights $W \in \mathbb{R}^{d_w \times d_w}$ of the *AdaptLayer*, where I is the d_w-dimensional identity matrix. In Sec. 5.3, we demonstrate that MUVE is more robust than its text-based counterparts in multiple aspects.

4. Multimodal and Multilingual Datasets

This section explains the training and evaluation datasets used in Sec. 5. All datasets are available at github.com/gsig/visual-grounding.

4.1. The HowToWorld Dataset

Existing instructional video datasets curated from YouTube (*e.g.*, the HowTo100M dataset) are in English. We follow the approach of [35] to obtain data in three new languages: French (Fr), Japanese (Ja) and Korean (Ko). We use their list of 23,000 tasks (*e.g.*, making a latte) and translate them to Fr, Ja and Ko. We obtain 31M, 30M and 34M unique clips with narration from automatic speech recognition for the Fr, Ja and Ko datasets, respectively. We use HowTo100M [35] as the English (En) dataset. To ensure that our datasets are stricly unpaired we removed any videos present in more than one of the datasets. More details are provided in the Appendix.

4.2. Text Corpora for Training Embeddings

To compare MUVE to the state-of-the-art unsupervised text-based word alignment methods, we use three text corpora: **(i) Wiki-En/Fr**: the publicly available release of Wikipedia in English and French. We filter the structured output to extract the sentences before processing as described in Sec. 5.1, **(ii) HowToW-Text-{En,Fr,Ko,Ja}**: we use the narration extracted from the videos of HowToWorld in multiple languages and **(iii) WMT Fr-En corpus**: we use the publicly available WMT French-English corpus, that consists of En-Fr translations for various news articles.

4.3. Evaluation benchmarks

Our goal translating words from one language to another (*e.g.*, En-Fr, En-Ko, En-Ja). We describe the datasets used to analyze the translations, also found in the Appendix.

The *Dictionary* En-{Fr,Ko,Ja}. We use the test split of the ground-truth bilingual dictionaries used in the MUSE paper [10] to compare our method to text-based word mapping methods. Each dictionary provides the translation of 1500 English words in another language (*e.g.*, Fr) and list multiple translations for each English word. There are 2943 En-Fr, 1922 En-Ko, and 1799 En-Ja pairs. As we focus on vision and to understand how different methods compare on visual versus non-visual words, we also manually annotate the bilingual dictionary for en-fr to select words that can be visually observed (*Dictionary (Visual)*). This results in 637 English words and 1430 En-Fr pairs. Among example words in the *Dictionary* dataset are: {*torpedo, giovanni, chat, catholics, herald, chuck, ...*} whereas the *Dictionary (Visual)* contains {*torpedo, chuck, pit, garrison, sprint, ...*}.

***Simple Words* En-{Fr,Ko,Ja}.** To examine the role of word frequency, we create a list of the 1000 most common English words from the Simple English Wikipedia. We translate this list to Fr, Ko, and Ja using the Google Translate interface. We manually filter these words to create a list of visual words (*Simple Words (Visual)*). Example words in the *Simple Words* dataset include {*correct, touch, hit, either, regard, carry, with, three, ...*} and *Simple Words (Visual)* contains {do, fall, police, carry, make, station, afternoon, money, club...}

Human Queries En-{Fr,Ko,Ja}. In order to also qualitatively assess the performance of our proposed model in Sec. 5.5, we create a text dataset (*Human Queries*) containing expressions similar to narrations contained in instructional videos. We manually defined a set of 444 visual queries along with their translations in En, Fr, Ko, and Ja. Examples include {*oil painting, make snowman, glue wood, cut tomato, play violin, open car door, paint shirt, tennis service, brew coffee, dribbling basketball, ...*}.

5. Experiments

In this section, we first provide our implementation details (Sec. 5.1); in Sec. 5.2, we demonstrate the effectiveness of our Base Model in word translation benchmarks. In Sec. 5.3, we show that the representations learned by our model can be used to improve the quality of text-based word translation methods. We also show that our method (MUVE) is more robust than the text-based methods (Sec. 5.4). Finally, in Sec. 5.5, we showcase various qualitative results that give further insight into our method.

5.1. Implementation Details

We tokenize the transcripts of the videos and lowercase. We create a vocabulary of the 65,536 most common words for each language, and map the rest to the UNK symbol. After preprocessing, we train monolingual word embeddings using Word2Vec [37] (Skip-Gram, 300 dim, 5 words, 5 negatives). We use these pretrained embeddings in MUVE, MUSE, and VecMap models.

At training, we sample a video clip (32 frames at 10 FPS) with its corresponding narration from the given datasets (*e.g.*, HowToW-En or the relevant HowToW-{Fr-Ko-Ja}). Each training batch includes clips from either language, and the negatives for each element in the NCE loss are the other elements from the batch in the same language. For the video encoder, we finetune an I3D model [9] pretrained on the Kinetics-400 dataset [9]. For the language models (Sec. 3.1), the word embedding layers are pretrained on the corresponding HowToW-Text datasets to incorporate distributional semantics. We use the Adam optimizer with an initial learning rate of 10^{-3} with batch size of 128 and train the model for 200k iterations on 2 Cloud TPUs.

English-French		Dictionary		Simple Words	
		All	Visual	All	Visual
1)	Random Chance	0.1	0.2	0.1	0.2
2)	Video Retrieval	6.3	7.6	12.5	18.6
3)	Base Model	9.1	15.2	28.0	45.3
4)	MUVE	**28.9**	**39.5**	**58.3**	**67.5**

Table 1: The performance of our models and the baselines as Recall@1 on En-Fr *Dictionary* and *Simple Words*.

Evaluation metrics. We report Recall@n in our experiments: given a query (*e.g*. 'Dog'), we retrieve n results (*e.g*. 'Chien', 'Chienne', 'Chiot', ...), and the retrieval is a success if *any* of the n results are listed as a correct translation in the ground-truth dictionary. If not specified otherwise, we report Recall@1 in the paper. We observed the same trend with Recall@10 and report it in the Appendix.

5.2. The Base Model Evaluation

We investigate whether sharing the visual encoder across languages improves the quality of word translations; to do so, we compare the results of our Base Model with two baselines which we explain below.

Baselines. Our first baseline method (*Random Chance*) retrieves a random hypothesis translation without using videos. The second baseline – *Video Retrieval* – uses videos to create a parallel corpus between the two languages. We first extract I3D features pretrained on Kinetics [9] for all video clips in HowToW-En and HowToW-Fr. We then, for each of the English video clips (100M), find the three closest French video clips (in terms of the L2 distance). Finally, we take the narrations associated with these video pairs to create a parallel text corpus. Given the parallel corpus, we can find alignments between English and French words based on their co-occurrence. More specifically, we calculate the joint probability between the English and French word pairs. For each English word, we can then rank the French words using this joint probability.

Results. We report the results of our models and the baselines on the Dictionary and Simple Words benchmarks in Table 1. We observe that our Base Model outperforms the baseline by a significant margin in both benchmarks. Moreover, not surprisingly, the performance of all methods is better on the *Visual* portion of these benchmarks. In Fig. 4, we provide two examples of the two types of failures of the *Video Retrieval* model: In the first row, the retrieved video is correct (visually related to the query) but the narrations in English and French do not convey the same meaning. In the second row, the frame from the retrieved video is somewhat visually similar to the query (both contain food) but does not depict the same concept. This example shows how visual understanding poses a challenge for this task.

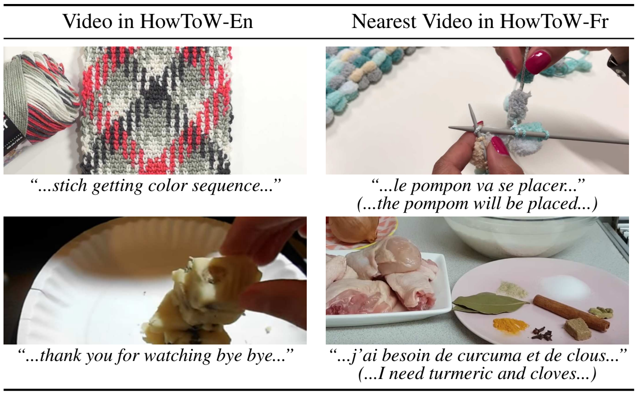

Video in HowToW-En
Nearest Video in HowToW-Fr
"...stich getting color sequence..."
"...le pompon va se placer..."
(...the pompom will be placed...)
"...thank you for watching bye bye..."
"...j'ai besoin de curcuma et de clous..."
(...I need turmeric and cloves...)

Figure 4: Examples of two types of failures of the *Video Retrieval* baseline. In the first row, the videos are visually related (knitting), but no words match, making learning translation challenging. In the second, the videos are related (food), but the left caption is irrelevant to the visual content.

5.3. MUVE: Improving Text-Based Alignment

We evaluate the proposed MUVE approach, that is how much the representations learned by our Base Model can improve the text-based word translation methods. We first describe text-based methods that use large scale corpora for word translation. Then, we show how using representations from our model (Sec. 3.1) improves over text-based approaches: three unsupervised, and one supervised methods described below. All methods use word embeddings trained on *HowToW-Text* for their respective languages.

Iterative Procrustes iteratively maps word embeddings of two languages using a distance-based heuristic; then it finds the orthogonal matrix that best maps the chosen pairs. We choose the best solution from 25 different initializations (either the identity matrix or random matrices).

MUSE [10] uses adversarial training to map the word embeddings to a space where they are indistinguishable, which provides better starting point for the *Iterative Procrustes* method. The results obtained from *MUSE* [10] have been found to be sensitive to the initialization [4].

VecMap [4] is more robust to initialization and differences across languages when compared to *MUSE*; it obtains better linear transformation by careful normalization, whitening, and dimensionality reduction.

Supervised provides an upper bound on the unsupervised methods: it uses 5,000 words and their translations to find an optimal orthonormal matrix that aligns the embeddings.

Results. In Table 2, we present the word translation results between English and French, Korean, and Japanese. Our method, MUVE, outperforms all the text-based methods. We observe a bigger improvement over the text-based methods for English-Korean and English-Japanese pairs. These results confirm previous findings that suggest text-

Dictionary		En-Fr		En-Ko	En-Ja
		All	Visual	All	All
1)	Iterative Procrustes	0.2	0.3	0.3	0.3
2)	MUSE [10]	26.3	36.2	11.8	11.6
3)	VecMap [4]	28.4	**40.8**	13.0	13.7
4)	MUVE	**28.9**	39.5	**17.7**	**15.1**
5)	Supervised	57.9	60.3	41.8	41.1

Table 2: Performance of our and text-based methods across different language pairs. We report Recall@1 on the *Dictionary* dataset. All method use word embeddings trained on *HowToW-Text* for their respective languages.

based methods are more suited for similar languages (*e.g.*, English and French) [4, 43] and shows that grounding in visual domain for word translation is especially effective in that regime. Finally, we also observe in Table 1 a significant improvement of MUVE (row 4) over our Base model alone (row 3) (+19.8% and +30.3% absolute improvement on the Dictionary and Simple Words benchmarks, respectively). Overall, this experiment validates our intuition that the information contained in the visual domain is *complementary* to the word co-occurence statistics used by the text-based methods for the task of unsupervised word translation.

Importance of the orthogonal constraint. As explained in Sec. 3.3, we add an orthogonal constraint to the *AdaptLayer* when applying MUVE. We observe that this penalty was a *key* component for MUVE. Precisely, there is a 43.0% relative drop of performance for Recall@1 on the Dictionary En-Fr (going from 28.9 in Table 2 to 16.6) benchmark when removing the orthogonal constraint. This further corroborates the findings described in [47].

5.4. Robustness of Unsupervised Word Translation

Sec. 5.3 shows that MUVE is more robust to the difference between language pairs when compared to the text-based methods (*i.e.* performance degrades less when going from French to Japanese and Korean in Table 2). Here we examine two other axes of robustness: the dissimilarity of the training corpora of the two languages and the amount of training data. All results reported in this section are on English and French languages because text-based models perform better for this pair.

Model selection. We observe that MUSE [10] and VecMap [4] are both sensitive to initialization. To address this, we select the optimal hyperparameters for the text-based method on the *test set*: we perform an extensive search over hyperparameters and random initialisations, *e.g.* 213 runs for the *MUSE* method, and compute the performance of these runs. We then select the *best* performing run on the test set, and hence reporting an upper bound of the true performance of these baselines. Note that when reporting numbers for MUVE we *only* use the monolingual validation loss for model selection, and all numbers for MUVE use the same hyperparameters.

	HowToW-Fr				WMT-Fr				Wiki-Fr			
	∼	[10]	[4]	MUVE	∼	[10]	[4]	MUVE	∼	[10]	[4]	MUVE
HTW-En	.62	45.8	45.4	**47.3**	.67	0.3	0.7	**35.1**	.65	0.3	0.1	**41.2**
WMT-En	.54	0.3	0.2	**26.4**	.40	**88.0**	87.2	85.0	.44	45.9	1.3	**54.9**
Wiki-En	.54	0.3	0.1	**32.6**	.46	**56.7**	52.3	55.9	.39	86.2	**86.7**	82.4

Table 3: Robustness of different methods to the dissimilarity of training corpora. We report Recall@10 on *English-French Dictionary* dataset for *MUSE* [10], *VecMap* [4], and *MUVE*, as well as the dissimilarity (∼) of the training corpora expressed with the Jensen Shannon Distance.

Dissimilarity of the training corpora. We examine how the dissimilarity of the training corpora affects the models. Following [15] we measure the dissimilarity of two corpora by comparing their word co-occurrence statistics. Specifically, we count the co-occurrence of each pair of words in the same sentence, and normalize to get a distribution per word. Then, we align pair of words in English and French using the Google Translate interface, and compute the Jensen Shannon distance between the distributions.

We report the results in Table 3; all methods are evaluated on the Dictionary dataset with the Recall@10 metric. Looking at the diagonal of the table, we observe when the corpora are *similar* (*e.g.*, Wiki-En and Wiki-Fr), all methods perform well. However when the corpora are *less similar* (off-diagonal elements), we observe that MUVE significantly outperforms its text-based counterparts. We note that methods trained on Wiki-En and WMT-Fr perform better compared to Wiki-Fr and WMT-En. This is likely due to the combination of Wiki-Fr and WMT-En being a smaller corpora: Wiki-En is much larger than Wiki-Fr while the WMT corpora in both languages are of the same size. In conclusion, our method by using visual grounding is more robust to the dissimilarity of the corpora in two languages.

Amount of training data. Unsupervised word translation is especially appealing for low-resource languages where there is no large corpora available. We investigate to what extent MUVE and the text-based methods are robust to the varying size of training data. More specifically, we use 100%, 10%, and 1% of the target training corpora (Wiki-Fr or HowToW-Fr) and report Recall@10. For MUVE, when reducing HowToW-Fr, we also reduce the amount of videos processed. Our results are shown in Fig. 5. MUVE is more robust to conditions where the training corpora is small when compared to the text-based methods, revealing another advantage of visual grounding for the task of unsupervised word translation.

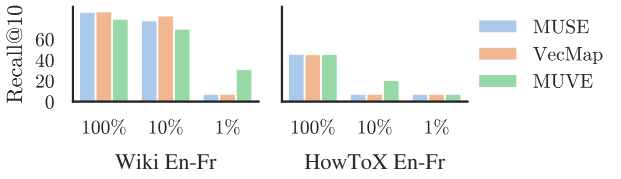


Figure 5: Recall@10 on *En-Fr Dictionary* for *MUSE*, *VecMap*, and *MUVE* varying amount of data.

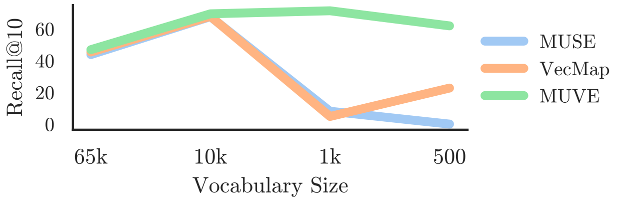


Figure 6: Recall@10 on *English-French Dictionary* for *MUSE*, *VecMap*, and *MUVE* for English and French pretrained word embeddings with various vocabulary sizes in French (65k, 10k, 1k, or 500 most common French words). All methods use HowToW-En and HowToW-Fr.

Vocabulary size. The text-based methods rely on words' context to align the space of two languages; consequently, the size of vocabulary (and the number of words' neighbors) can play a role in their performance. For low-resource languages, we do not have access to a large corpus and as a result words might not have many neighbors. We explore to what extent the vocabulary size influences the performance of different methods. Fig. 6 shows Recall@10 for different methods and vocabulary sizes. We keep the full English vocabulary and vary the size of French vocabulary. We only evaluate on words that are seen in both English and French vocabularies. We observe that MUVE is the only method whose performance does not deteriorate when vocabulary size decreases (even when it is as small as 500).

5.5. Qualitative Results

In Fig. 7, we visualize a 2-stage inference process: **(1)** given an English query (from the *Human Queries* dataset), using our Base Model, we retrieve the video from the training set that is most similar to that query. **(2)** Given that video, we retrieve the closest text from the French Human Queries dataset. The model is able to retrieve relevant videos. However, we also observe that such 2-stage approach can be problematic for translation (*e.g.* the second row of Fig. 7 where both individual steps makes sense but the overall result is incorrect due to model drift).

In Table 4, we visualize the 1-stage inference process described in Sec. 3.2. The model is often accurate, and errors often result in semantically similar words, such as translating *"a man with a dog"* as *"walk dog"* and *"feed dog"*.

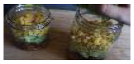
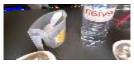
English	Retrieved Video	Top French Hypotheses Given Video	
Beach		Plage (*Beach*)	Courir sur la plage (*Running on the beach*)
Point at the sky		Des Nuages (*Clouds*)	Le coucher du soleil (*Sunset*)
Christmas tree		Sapin de noël (*Christmas Tree*)	Faire un bonhomme de neige (*Make snowman*)
Cut carrot		Couper la carotte (*Cut carrot*)	Carotte (*Carrot*)
Add pickle		Ajoutez des cornichons (*Add pickle*)	Mélanger les legumes verts (*Mix greens*)
Add water		Verser de l'eau (*Pour water*)	Bien mélanger (*Mix thoroughly*)

Figure 7: Left: a frame from the video that the model chose as most related to the english query. Right: top 2 french predictions conditioned on the video. The visual grounding provides a weak but usable signal for translation.

English Text	1st Model Retrieval *(English Meaning)*	2nd Model Retrieval *(English Meaning)*
Boy Playing	Balle qui rebondit par le chat *(Ball Bouncing by the Cat)*	Homme jouant au foot *(Man Playing Football)*
Girl Eats Ice Cream	Chocolat *(Chocolate)*	Sucrer les pancakes *(Top Pancake Sugar)*
Man Driving Red Car	Homme conduit voiture rouge *(Man Driving Red Car)*	Voiture rouge *(Red Car)*
A Man with a Dog	Promener un chien *(Walk Dog)*	Nourrir un chien *(Feed Dog)*
Air Conditioning	Voler dans les airs *(Fly Air)*	Air conditionné *(Air Conditioning)*

Table 4: Top 2 retrieved results in French on the *Human Queries* dataset given an English query.

6. Conclusion

Learning multiple languages is a challenging problem that multilingual children tackle with ease. The shared visual domain can help as it allows children to relate words in different languages through the similarity of their visual experience. Inspired by this, we propose an unsupervised multimodal model for word translation that learns from instructional YouTube videos. This is beneficial over text-based methods, allowing for more robust translation when faced with diverse corpora. Future work needs to explore extensions to the proposed model for translating full sentences.

Acknowledgements. The authors would like to thank Antoine Miech for invaluable advice and their HowTo100M code, as well as Carl Doersch, Ankush Gupta, Relja Arandjelović, Viorica Pătrăucean, Ellen Clancy, and others at DeepMind for helpful discussions, support, and feedback on the project. The authors would finally like to thank Lisa Anne Hendricks and Sebastian Ruder for their feedback on the manuscript.

References

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016. 3

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 3

[3] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 3

[4] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, 2017. 2, 6, 7

[5] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*, 2018. 2

[6] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, 2018. 2

[7] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *JMLR*, 2003. 3

[8] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. 2018. 3

[9] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 4, 5, 6

[10] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv:1710.04087*, 2017. 2, 4, 5, 6, 7

[11] Annick De Houwer. Bilingual language acquisition. *The handbook of child language*, 2017. 2

[12] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017. 3

[13] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 3

[14] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 3

[15] Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kardas, Sylvain Gugger, and Jeremy Howard. Multifit: Efficient multi-lingual language model fine-tuning. *EMNLP*, 2019. 7

[16] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *ACL*, 2016. 3

[17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 3

[18] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *NeurIPS*, 2015. 3

[19] Fred Genesee. Early bilingual development: One language or two? *Journal of child language*, 1989. 2

[20] Fred Genesee, Johanne Paradis, and Martha B. Crago. *Dual language development & disorders: A handbook on bilingualism & second language learning*. Paul H Brookes Publishing, 2004. 1

[21] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 4

[22] Thanh-Le Ha, Jan Niehues, and Alexander Waibel. Toward multilingual neural machine translation with universal encoder and decoder. *IWSLT*, 2016. 4

[23] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 1990. 3

[24] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. *ICCV*, 2017. 3

[25] Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. Learning to follow directions in street view. *AAAI*, 2020. 3

[26] Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation. *EMNLP*, 2018. 2

[27] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. *ICCV*, 2019. 3

[28] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *ACL*, 2017. 4

[29] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv:1602.02410*, 2016. 4

[30] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 3

[31] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*, 2015. 2

[32] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. Learning visual question answering by bootstrapping hard attention. In *ECCV*, 2018. 3

[33] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A deep learning approach to visual question answering. *IJCV*, 2017. 3

[34] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. *arXiv:1912.06430*, 2019. 3, 4

[35] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 2, 3, 5

[36] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv:1309.4168*, 2013. 2, 4

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 4, 5

[38] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *CVPR*, 2017. 3

[39] Candace Ross, Andrei Barbu, Yevgeni Berzak, Battushig Myanganbayar, and Boris Katz. Grounding language acquisition by training semantic parsers using captioned videos. In *EMNLP*, 2018. 3

[40] Ozan Sener, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Unsupervised semantic parsing of video collections. In *ICCV*, 2015. 3

[41] Kevin Shen, Amlan Kar, and Sanja Fidler. Lifelong learning for image captioning by asking natural language questions. *ICCV 2019*, 2019. 3

[42] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *ICLR*, 2017. 2

[43] Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. *ACL*, 2018. 2, 7

[44] Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. Unsupervised multi-modal neural machine translation. In *CVPR*, 2019. 3

[45] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 3

[46] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *ICCV*, 2019. 3

[47] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *ACL*, 2015. 4, 7

[48] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044*, 2015. 3

[49] Shoou-I Yu, Lu Jiang, and Alexander Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *ACM*, 2014. 3

[50] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 3