Imagination Improves Multimodal Translation

Desmond Elliott[⋆◇] and **Ákos Kádár**[†]
[⋆]ILLC, University of Amsterdam
[◇]School of Informatics, University of Edinburgh
[†]TiCC, Tilburg University
d.elliott@ed.ac.uk, a.kadar@uvt.nl

Abstract

We decompose multimodal translation into two sub-tasks: learning to translate and learning visually grounded representations. In a multitask learning framework, translations are learned in an attention-based encoder-decoder, and grounded representations are learned through image representation prediction. Our approach improves translation performance compared to the state of the art on the Multi30K dataset. Furthermore, it is equally effective if we train the image prediction task on the external MS COCO dataset, and we find improvements if we train the translation model on the external News Commentary parallel text.

1 Introduction

Multimodal machine translation is the task of translating sentences in context, such as images paired with a parallel text (Specia et al., 2016). This is an emerging task in the area of multilingual multimodal natural language processing. Progress on this task may prove useful for translating the captions of the images illustrating online news articles, and for multilingual closed captioning in international television and cinema.

Initial efforts have not convincingly demonstrated that visual context can improve translation quality. In the results of the First Multimodal Translation Shared Task, only three systems outperformed an off-the-shelf text-only phrase-based machine translation model, and the best performing system was equally effective with or without the visual features (Specia et al., 2016). There remains an open question about how translation models should take advantage of visual context.

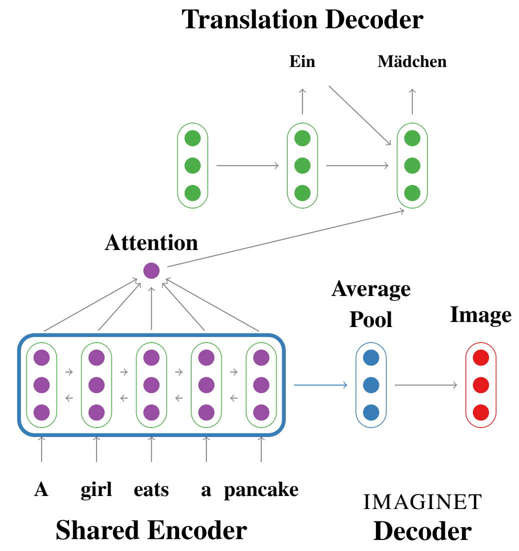


Figure 1: The Imagination model learns visually-grounded representations by sharing the encoder network between the Translation Decoder with image prediction in the IMAGINET Decoder.

We present a multitask learning model that decomposes multimodal translation into learning a translation model and learning visually grounded representations. This decomposition means that our model can be trained over external datasets of parallel text or described images, making it possible to take advantage of existing resources. Figure 1 presents an overview of our model, Imagination, in which source language representations are shared between tasks through the Shared Encoder. The translation decoder is an attention-based neural machine translation model (Bahdanau et al., 2015), and the image prediction decoder is trained to predict a global feature vector of an image that is associated with a sentence (Chrupała et al., 2015, IMAGINET). This decomposition encourages grounded learning in the shared encoder because the IMAGINET decoder is trained to imagine

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 130–141,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

the image associated with a sentence. It has been shown that grounded representations are qualitatively different from their text-only counterparts (Kádár et al., 2016) and correlate better with human similarity judgements (Chrupała et al., 2015). We assess the success of the grounded learning by evaluating the image prediction model on an image–sentence ranking task to determine if the shared representations are useful for image retrieval (Hodosh et al., 2013). In contrast with most previous work, our model does not take images as input at translation time, rather it learns grounded representations in the shared encoder.

We evaluate Imagination on the Multi30K dataset (Elliott et al., 2016) using a combination of in-domain and out-of-domain data. In the in-domain experiments, we find that multitasking translation with image prediction is competitive with the state of the art. Our model achieves 55.8 Meteor as a single model trained on multimodal in-domain data, and 57.6 Meteor as an ensemble.

In the experiments with out-of-domain resources, we find that the improvement in translation quality holds when training the IMAGINET decoder on the MS COCO dataset of described images (Chen et al., 2015). Furthermore, if we significantly improve our text-only baseline using out-of-domain parallel text from the News Commentary corpus (Tiedemann, 2012), we still find improvements in translation quality from the auxiliary image prediction task. Finally, we report a state-of-the-art result of 59.3 Meteor on the Multi30K corpus when ensembling models trained on in- and out-of-domain resources.

The main contributions of this paper are:

- We show how to apply multitask learning to multimodal translation. This makes it possible to train models for this task using external resources alongside the expensive triple-aligned source-target-image data.

- We decompose multimodal translation into two tasks: learning to translate and learning grounded representations. We show that each task can be trained on large-scale external resources, e.g. parallel news text or images described in a single language.

- We present a model that achieves state of the art results without using images as an input. Instead, our model learns visually grounded source language representations using an auxiliary image prediction objective. Our model does not need any additional parameters to translate unseen sentences.

2 Problem Formulation

Multimodal translation is the task of producing target language translation y, given the source language sentence x and additional context, such as an image v (Specia et al., 2016). Let x be a source language sentence consisting of N tokens: x_1, x_2, . . ., x_n and let y be a target language sentence consisting of M tokens: y_1, y_2, . . ., y_m. The training data consists of tuples $\mathcal{D} \in (x, y, v)$, where x is a description of image v, and y is a translation of x.

Multimodal translation has previously been framed as minimising the negative log-likelihood of a translation model that is additionally conditioned on the image, i.e. $J(\theta) = -\sum_j \log \mathrm{p}(y_j|y_{<j}, x, v)$. Here, we decompose the problem into learning to translate and learning visually grounded representations. The decomposition is based on sharing parameters θ between these two tasks, and learning task-specific parameters ϕ. We learn the parameters in a multitask model with shared parameters in the source language encoder. The translation model has task-specific parameters ϕ^t in the attention-based decoder, which are optimized through the translation loss $J_T(\theta, \phi^t)$. Grounded representations are learned through an image prediction model with task-specific parameters ϕ^g in the image-prediction decoder by minimizing $J_G(\theta, \phi^g)$. The joint objective is given by mixing the translation and image prediction tasks with the parameter w:

$$J(\theta, \phi) = wJ_T(\theta, \phi^t) + (1 - w)J_G(\theta, \phi^g) \quad (1)$$

Our decomposition of the problem makes it straightforward to optimise this objective without paired tuples, e.g. where we have an external dataset of described images $\mathcal{D}_{image} \in (x, v)$ or an external parallel corpus $\mathcal{D}_{text} \in (x, y)$.

We train our multitask model following the approach of Luong et al. (2016). We define a primary task and an auxiliary task, and a set of parameters θ to be shared between the tasks. A minibatch of updates is performed for the primary task with probability w, and for the auxiliary task with $1-w$. The primary task is trained until convergence and weight w determines the frequency of parameter updates for the auxiliary task.

3 Imagination Model

3.1 Shared Encoder

The encoder network of our model learns a representation of a sequence of N tokens $x_{1 \dots n}$ in the source language with a bidirectional recurrent neural network (Schuster and Paliwal, 1997). This representation is shared between the different tasks. Each token is represented by a one-hot vector $\mathbf{x_i}$, which is mapped into an embedding $\mathbf{e_i}$ through a learned matrix $\mathbf{E}$:

$$\mathbf{e_i} = \mathbf{x_i} \cdot \mathbf{E} \tag{2}$$

A sentence is processed by a pair of recurrent neural networks, where one captures the sequence left-to-right (forward), and the other captures the sequence right-to-left (backward). The initial state of the encoder $\mathbf{h_{-1}}$ is a learned parameter:

$$\overrightarrow{\mathbf{h_i}} = \overrightarrow{\text{RNN}}(\overrightarrow{\mathbf{h_{i-1}}}, \mathbf{e_i}) \tag{3}$$

$$\overleftarrow{\mathbf{h_i}} = \overleftarrow{\text{RNN}}(\overleftarrow{\mathbf{h_{i-1}}}, \mathbf{e_i}) \tag{4}$$

Each token in the source language input sequence is represented by a concatenation of the forward and backward hidden state vectors:

$$\mathbf{h_i} = [\overrightarrow{\mathbf{h_i}}; \overleftarrow{\mathbf{h_i}}] \tag{5}$$

3.2 Neural Machine Translation Decoder

The translation model decoder is an attention-based recurrent neural network (Bahdanau et al., 2015). Tokens in the decoder are represented by a one-hot vector $\mathbf{y_j}$, which is mapped into an embedding $\mathbf{e_j}$ through a learned matrix $\mathbf{E_y}$:

$$\mathbf{e_j} = \mathbf{y_j} \cdot \mathbf{E_y} \tag{6}$$

The inputs to the decoder are the previously predicted token $\mathbf{y_{j-1}}$, the previous decoder state $\mathbf{d_{j-1}}$, and a timestep-dependent context vector $\mathbf{c_j}$ calculated over the encoder hidden states:

$$\mathbf{d_j} = \text{RNN}(\mathbf{d_{j-1}}, \mathbf{y_{j-1}}, \mathbf{e_j}) \tag{7}$$

The initial state of the decoder $\mathbf{d}_{\text{-1}}$ is a nonlinear transform of the mean of the encoder states, where $\mathbf{W}_{init}$ is a learned parameter:

$$\mathbf{d}_{\text{-1}} = \tanh(\mathbf{W}_{init} \cdot \frac{1}{N} \sum_{i}^{N} \mathbf{h_i}) \tag{8}$$

The context vector c_j is a weighted sum over the encoder hidden states, where N denotes the length of the source sentence:

$$\mathbf{c_j} = \sum_{i=1}^{N} \alpha_{ji} \mathbf{h_i} \tag{9}$$

The α_{ji} values are the proportion of which the encoder hidden state vectors $\mathbf{h_{1 \dots n}}$ contribute to the decoder hidden state when producing the jth token in the translation. They are computed by a feed-forward neural network, where $\mathbf{v_a}$, $\mathbf{W_a}$ and $\mathbf{U_a}$ are learned parameters:

$$\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_{l=1}^{N} \exp(e_{li})} \tag{10}$$

$$e_{ji} = \mathbf{v_a} \cdot \tanh(\mathbf{W_a} \cdot \mathbf{d_{j-1}} + \mathbf{U_a} \cdot \mathbf{h_i}) \tag{11}$$

From the hidden state $\mathbf{d_j}$ the network predicts the conditional distribution of the next token y_j, given a target language embedding $\mathbf{e_{j-1}}$ of the previous token, the current hidden state $\mathbf{d_j}$, and the calculated context vector $\mathbf{c_j}$. Note that at training time, y_{j-1} is the true observed token; whereas for unseen data we use the inferred token $\hat{y}_{j-1}$ sampled from the output of the softmax:

$$p(y_j | y_{<j}, c) = \text{softmax}(\tanh(\mathbf{e_{j-1}} + \mathbf{d_j} + \mathbf{c_j})) \tag{12}$$

The translation model is trained to minimise the negative log likelihood of predicting the target language output:

$$J_{NLL}(\theta, \phi^t) = -\sum_j \log \text{p}(y_j | y_{<j}, x) \tag{13}$$

3.3 Imaginet Decoder

The image prediction decoder is trained to predict the visual feature vector of the image associated with a sentence (Chrupała et al., 2015). It encourages the shared encoder to learn grounded representations for the source language.

A source language sentence is encoded using the Shared Encoder, as described in Section 3.1. Then we transform the shared encoder representation into a single vector by taking the mean pool over the hidden state annotations, the same way we initialise the hidden state of the translation decoder (Eqn. 8). This sentence representation is the input to a feed-forward neural network that predicts the visual feature vector $\hat{\mathbf{v}}$ associated with a

	Size	Tokens	Types	Images
Multi30K: parallel text with images				
En	31K	377K	10K	31K
De		368K	16K	
MS COCO: external described images				
En	414K	4.3M	24K	83K
News Commentary: external parallel text				
En	240K	8.31M	17K	–
De		8.95M		–

Table 1: The datasets used in our experiments.

sentence with parameters $\mathbf{W_{vis}}$:

$$\hat{\mathbf{v}} = \tanh(\mathbf{W_{vis}} \cdot \frac{1}{N} \sum_{i}^{N} \mathbf{h_i}) \tag{14}$$

This decoder is trained to predict the true image vector $\mathbf{v}$ with a margin-based objective, parameterised by the minimum margin α, and the cosine distance $d(\cdot, \cdot)$. A margin-based objective has previously been used in grounded representation learning (Vendrov et al., 2016; Chrupała et al., 2017). The contrastive examples $\mathbf{v}'$ are drawn from the other instances in a minibatch:

$$\begin{aligned} J_{MAR}(\theta, \phi^t) = \sum_{\mathbf{v}' \neq \mathbf{v}} \max\{0, \alpha - d(\hat{\mathbf{v}}, \mathbf{v}) \\ + d(\hat{\mathbf{v}}, \mathbf{v}')\} \end{aligned} \tag{15}$$

4 Data

We evaluate our model using the benchmark Multi30K dataset (Elliott et al., 2016), which is the largest collection of images paired with sentences in multiple languages. This dataset contains 31,014 images paired with an English language sentence and a German language translation: 29,000 instances are reserved for training, 1,014 for development, and 1,000 for evaluation.[1]

The English and German sentences are preprocessed by normalising the punctuation, lowercasing and tokenizing the text using the Moses toolkit. We additionally decompound the German text using Zmorge (Sennrich and Kunz, 2014). This results in vocabulary sizes of 10,214 types for English and 16,022 for German.

We also use two external datasets to evaluate our model: the MS COCO dataset of English described images (Chen et al., 2015), and the English-German News Commentary parallel corpus (Tiedemann, 2012). When we perform experiments with the News Commentary corpus, we first calculate a 17,597 sub-word vocabulary using SentencePiece (Schuster and Nakajima, 2012) over the concatentation of the Multi30K and News Commentary datasets. This gives us a shared vocabulary for the external data that reduces the number of out-of-vocabulary tokens.

Images are represented by 2048D vectors extracted from the 'pool5/7x7_s1' layer of the GoogLeNet v3 CNN (Szegedy et al., 2015).

5 Experiments

We evaluate our multitasking approach with in- and out-of-domain resources. We start by reporting results of models trained using only the Multi30K dataset. We also report the results of training the IMAGINET decoder with the COCO dataset. Finally, we report results on incorporating the external News Commentary parallel text into our model. Throughout, we report performance of the En→De translation using Meteor (Denkowski and Lavie, 2014) and BLEU (Papineni et al., 2002) against lowercased tokenized references.

5.1 Hyperparameters

The encoder is a 1000D Gated Recurrent Unit bidirectional recurrent neural network (Cho et al., 2014, GRU) with 620D embeddings. We share all of the encoder parameters between the primary and auxiliary task. The translation decoder is a 1000D GRU recurrent neural network, with a 2000D context vector over the encoder states, and 620D word embeddings (Sennrich et al., 2017). The Imaginet decoder is a single-layer feed-forward network, where we learn the parameters $\mathbf{W_{vis}} \in \mathbb{R}^{2048\text{x}2000}$ to predict the true image vector with $\alpha = 0.1$ for the Imaginet objective (Equation 15). The models are trained using the Adam optimiser with the default hyperparameters (Kingma and Ba, 2015) in minibatches of 80 instances. The translation task is defined as the primary task and convergence is reached when BLEU has not increased for five epochs on the validation data. Gradients are clipped when their norm ex-

[1] The Multi30K dataset also contains 155K independently collected descriptions in German and English. In order to make our experiments more comparable with previous work, we do not make use of this data.

	Meteor	BLEU
NMT	54.0 ± 0.6	35.5 ± 0.8
Calixto et al. (2017)	55.0	36.5
Calixto and Liu (2017)	55.1	37.3
Imagination	55.8 ± 0.4	36.8 ± 0.8
Toyama et al. (2016)	56.0	36.5
Hitschler et al. (2016)	56.1	34.3
Moses	56.9	36.9

Table 2: En→De translation results on the Multi30K dataset. Our Imagination model is competitive with the state of the art when it is trained on in-domain data. We report the mean and standard deviation of three random initialisations.

	Meteor	BLEU
Imagination	55.8 ± 0.4	36.8 ± 0.8
Imagination (COCO)	55.6 ± 0.5	36.4 ± 1.2

Table 3: Translation results when using out-of-domain described images. Our approach is still effective when the image prediction model is trained over the COCO dataset.

	Meteor	BLEU
NMT	52.8 ± 0.6	33.4 ± 0.6
+ NC	56.7 ± 0.3	37.2 ± 0.7
+ Imagination	56.7 ± 0.1	37.4 ± 0.3
+ Imagination (COCO)	57.1 ± 0.2	37.8 ± 0.7
Calixto et al. (2017)	56.8	39.0

Table 4: Translation results with out-of-domain parallel text and described images. We find further improvements when we multitask with the News Commentary (NC) and COCO datasets.

ceeds 1.0. Dropout is set to 0.2 for the embeddings and the recurrent connections in both tasks (Gal and Ghahramani, 2016). Translations are decoded using beam search with 12 hypotheses.

5.2 In-domain experiments

We start by presenting the results of our multitask model trained using only the Multi30K dataset. We compare against state-of-the-art approaches and text-only baselines. Moses is the phrase-based machine translation model (Koehn et al., 2007) reported in (Specia et al., 2016). NMT is a text-only neural machine translation model. Calixto et al. (2017) is a double-attention model over the source language and the image. Calixto and Liu (2017) is a multimodal translation model that conditions the decoder on semantic image vector extracted from the VGG-19 CNN. Hitschler et al. (2016) uses visual features in a target-side retrieval model for translation. Toyama et al. (2016) is most comparable to our approach: it is a multimodal variational NMT model that infers latent variables to represent the source language semantics from the image and linguistic data.

Table 2 shows the results of this experiment. We can see that the combination of the attention-based translation model and the image prediction model is a 1.8 Meteor point improvement over the NMT baseline, but it is 1.1 Meteor points worse than the strong Moses baseline. Our approach is competitive with previous approaches that use visual features as inputs to the decoder and the target-side reranking model. It also competitive with Toyama et al. (2016), which also only uses images for training. These results confirm that our multitasking approach uses the image prediction task to improve the encoder of the translation model.

5.3 External described image data

Recall from Section 2 that we are interested in scenarios where x, y, and v are drawn from different sources. We now experiment with separating the translation data from the described image data using $\mathcal{D}_{image}$: MS COCO dataset of 83K described images[2] and $\mathcal{D}_{text}$: Multi30K parallel text.

Table 3 shows the results of this experiment. We find that there is no significant difference between training the IMAGINET decoder on in-domain (Multi30K) or out-of-domain data (COCO). This result confirms that we can separate the parallel text from the described images.

5.4 External parallel text data

We now experiment with training our model on a combination of the Multi30K and the News Commentary English-German data. In these experiments, we concatenate the Multi30K and News

[2] Due to differences in the vocabularies of the respective datasets, we do not train on examples where more than 10% of the tokens are out-of-vocabulary in the Multi30K dataset.

	Parallel text		Described images			
	Multi30K	News Commentary	Multi30K	COCO	Meteor	BLEU
Zmorge	✓				56.2	37.8
	✓		✓		57.6	39.0
Sub-word	✓				54.4	35.0
	✓	✓			58.6	39.4
	✓	✓	✓		59.0	39.5
	✓	✓		✓	**59.3**	**40.2**

Table 5: Ensemble decoding results. Zmorge denotes models trained with decompounded German words; Sub-word denotes joint SentencePiece word splitting (see Section 4 for more details).

Commentary datasets into a single $\mathcal{D}_{text}$ training dataset, similar to Freitag and Al-Onaizan (2016). We compare our model against Calixto et al. (2017), who pre-train their model on the WMT'15 English-German parallel text and back-translate (Sennrich et al., 2016) additional sentences from the bilingual independent descriptions in the Multi30K dataset (Footnote 2).

Table 4 presents the results. The text-only NMT model using sub-words is 1.2 Meteor points lower than decompounding the German text. Nevertheless, the model trained over a concatenation of the parallel texts is a 2.7 Meteor point improvement over this baseline (+ NC) and matches the performance of our Multitasking model that uses only in-domain data (Section 5.2). We do not see an additive improvement for the multitasking model with the concatenated parallel text and the in-domain data (+ Imagination) using a training objective interpolation of $w = 0.89$ (the ratio of the training dataset sizes). This may be because we are essentially learning a translation model and the updates from the IMAGINET decoder are forgotten. Therefore, we experiment with multitasking the concatenated parallel text and the COCO dataset ($w = 0.5$). We find that balancing the datasets improves over the concatenated text model by 0.4 Meteor (+ Imagination (COCO)). Our multitasking approach improves upon Calixto et al. by 0.3 Meteor points. Our model can be trained in 48 hours using 240K parallel sentences and 414K described images from out-of-domain datasets. Furthermore, recall that our model does not use images as an input for translating unseen data, which results in 6.2% fewer parameters compared to using the 2048D Inception-V3 visual features to initialise the hidden state of the decoder.

5.5 Ensemble results

Table 5 presents the results of ensembling different randomly initialised models. We achieve a start-of-the-art result of 57.6 Meteor for a model trained on only in-domain data. The improvements are more pronounced for the models trained using sub-words and out-of-domain data. An ensemble of baselines trained on sub-words is initially worse than an ensemble trained on Zmorge decompounded words. However, we always see an improvement from ensembling models trained on in- and out-of-domain data. Our best ensemble is trained on Multi30K parallel text, the News Commentary parallel text, and the COCO descriptions to set a new state-of-the-art result of 59.3 Meteor.

5.6 Multi30K 2017 results

We also evaluate our approach against 16 submissions to the WMT Shared Task on Multimodal Translation and Multilingual Image Description (Elliott et al., 2017). This shared task features a new evaluation dataset: Multi30K Test 2017 (Elliott et al., 2017), which contains 1,000 new evaluation images. The shared task submissions are evaluated with Meteor and human direct assessment (Graham et al., 2017). We submitted two systems, based on whether they used only the Multi30K dataset (constrained) or used additional external resources (unconstrained). Our constrained submission is an ensemble of three Imagination models trained over only the Multi30K training data. This achieves a Meteor score of 51.2, and a joint 3rd place ranking according to human assessment. Our unconstrained submission is an ensemble of three Imagination models trained with the Multi30K, News Commentary, and MS COCO datasets. It achieves a Meteor score of

	Source:	two children on their stomachs lay on the ground under a pipe
	NMT:	zwei kinder auf ihren gesichtern liegen unter dem boden auf dem boden
	Ours:	zwei kinder liegen bäuchlings auf dem boden unter einer schaukel
	Source:	small dog in costume stands on hind legs to reach dangling flowers
	NMT:	ein kleiner hund steht auf dem hinterbeinen und läuft , nach links von blumen zu sehen
	Ours:	ein kleiner hund in einem kostüm steht auf den hinterbeinen , um die blumen zu erreichen
	Source:	a bird flies across the water
	NMT:	ein vogel fliegt über das wasser
	Ours:	ein vogel fliegt durch das wasser

Table 6: Examples where our model improves or worsens the translation compared to the NMT baseline. Top: NMT translates the wrong body part; both models skip "pipe". Middle: NMT incorrectly translates the verb and misses several nouns. Bottom: Our model incorrectly translates the preposition.

53.5, and 2nd place in the human assessment.

5.7 Qualitative examples

Table 6 shows examples of where the multitasking model improves or worsens translation performance compared to the baseline model[3]. The first example shows that the baseline model makes a significant error in translating the pose of the children, translating "on their stomachs" as "on their faces"). The middle example demonstrates that the baseline model translates the dog as walking ("läuft") and then makes grammatical and sense errors after the clause marker. Both models neglect to translate the word "dangling", which is a low-frequency word in the training data. There are instances where the baseline produces better translations than the multitask model: In the bottom example, our model translates a bird flying through the water ("durch") instead of "over" the water.

6 Discussion

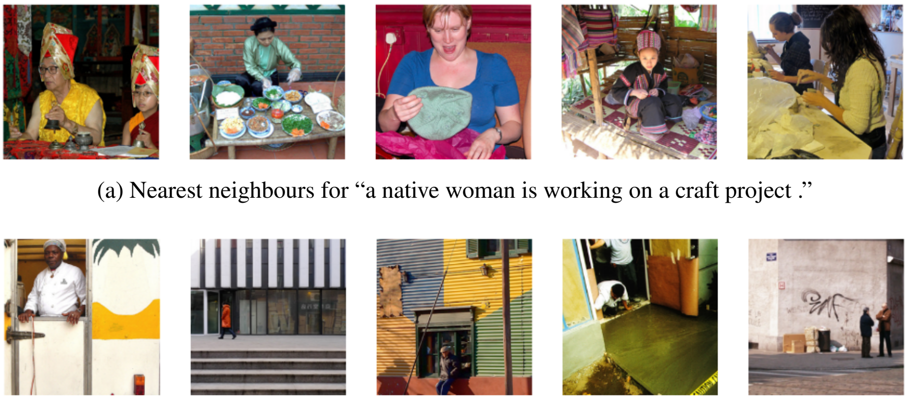
6.1 Does the model learn grounded representations?

A natural question to ask if whether the multitask model is actually learning representations that are relevant for the images. We answer this question by evaluating the Imaginet decoder in an image–sentence ranking task. Here the input is a source language sentence, from which we predict its image vector $\hat{\mathbf{v}}$. The predicted vector $\hat{\mathbf{v}}$ can be compared against the true image vectors $\mathbf{v}$ in the evaluation data using the cosine distance to produce a ranked order of the images. Our model returns a median rank of 11.0 for the true image compared to the predicted image vector. Figure 2 shows examples of the nearest neighbours of the images predicted by our multitask model. We can see that the combination of the multitask source language representations and IMAGINET decoder leads to the prediction of relevant images. This confirms that the shared encoder is indeed learning visually grounded representations.

6.2 The effect of visual feature vectors

We now study the effect of varying the Convolutional Neural Network used to extract the visual features used in the Imaginet decoder. It has previously been shown that the choice of visual features can affect the performance of vision and language models (Jabri et al., 2016; Kiela et al., 2016). We compare the effect of training the IMAGINET decoder to predict different types of image features, namely: 4096D features extracted from the 'fc7'' layer of the VGG-19 model (Simonyan and Zisserman, 2015), 2048D features extracted from the 'pool5/7x7_s1' layer of InceptionNet V3 (Szegedy et al., 2015), and 2048D features extracted from 'avg_pool' layer of ResNet-50 (He et al., 2016). Table 7 shows the results of this experiment. There is a clear difference between predicting the 2048D

[3] We used MT-ComparEval (Klejch et al., 2015)



(a) Nearest neighbours for "a native woman is working on a craft project ."

(b) Nearest neighbours for "there is a cafe on the street corner with an oval painting on the side of the building ."

Figure 2: We can interpret the IMAGINET Decoder by visualising the predictions made by our model.

	Meteor	Median Rank
Inception-V3	56.0 ± 0.1	11.0 ± 0.0
Resnet-50	54.7 ± 0.4	11.7 ± 0.5
VGG-19	53.6 ± 1.8	13.0 ± 0.0

Table 7: The type of visual features predicted by the IMAGINET Decoder has a strong impact on the Multitask model performance.

vectors (Inception-V3 and ResNet-50) compared to the 4096D vector from VGG-19). This difference is reflected in both the translation Meteor score and the Median rank of the images in the validation dataset. This is likely because it is easier to learn the parameters of the image prediction model that has fewer parameters (8.192 million for VGG-19 vs. 4.096 million for Inception-V3 and ResNet-50). However, it is not clear why there is such a pronounced difference between the Inception-V3 and ResNet-50 models[4].

7 Related work

Initial work on multimodal translation used semantic or spatially-preserving image features as inputs to a translation model. Semantic image features are typically extracted from the final layer of a pre-trained object recognition CNN, e.g. 'pool5/7x7_s1' in GoogLeNet (Szegedy et al., 2015). This type of vector has been used as input to the encoder (Elliott et al., 2015; Huang et al., 2016), the decoder (Libovický et al., 2016), or as features in a phrase-based translation model (Shah et al., 2016; Hitschler et al., 2016). Spatially-preserving image features are extracted from deeper inside a CNN, where the position of a feature is related to its position in the image. These features have been used in "double-attention models", which calculate independent context vectors for the source language and a convolutional image features (Calixto et al., 2016; Caglayan et al., 2016; Calixto et al., 2017). We use an attention-based translation model but our multitask model does not use images for translation.

More related to our work is an extension of Variational Neural Machine Translation to infer latent variables to *explicitly* model the semantics of source sentences from visual and linguistic information (Toyama et al., 2016). They report improvements on the Multi30K data set but their model needs additional parameters in the "neural inferrer" modules. In our model, the grounded semantics are represented *implicitly* in the shared encoder. They assume Source-Target-Image training data, whereas our approach achieves equally good results if we train on separate Source-Image and Source-Target datasets. Saha et al. (2016) study cross-lingual image description where the task is to generate a sentence in language L_1 given the image, using only Image-L_2 and L_1-L_2 training corpora. They propose a Correlational Encoder-Decoder to model the Image-L_2 and L_1-L_2 data, which learns correlated representations for paired Image-L_2 data and decodes L_1 from the joint representation. Similar to our work, the encoder is trained by minimizing two loss functions: the Image-L_2 correlation loss, and the L_1 decoding

[4] We used pre-trained CNNs (https://github.com/fchollet/deep-learning-models), which claim equal ILSVRC object recognition performance for both models: 7.8% top-5 error with a single-model and single-crop.

cross-entropy loss. Nakayama and Nishida (2017) consider a zero-resource problem, where the task is to translate from L_1 to L_2 with only Image-L_1 and Image-L_2 corpora. Their model embeds the image, L_1, and L_2 in a joint multimodal space learned by minimizing a multi-task ranking loss between both pairs of examples. In this paper, we focus on *enriching* source language representations with visual information instead of zero-resource learning.

Multitask Learning improves the generalisability of a model by requiring it to be useful for more than one task (Caruana, 1997). This approach has recently been used to improve the performance of sentence compression using eye gaze as an auxiliary task (Klerke et al., 2016), and to improve shallow parsing accuracy through the auxiliary task of predicting keystrokes in an out-of-domain corpus (Plank, 2016). More recently, Bingel and Søgaard (2017) analysed the beneficial relationships between primary and auxiliary sequential prediction tasks. In the translation literature, multitask learning has been used to learn a one-to-many languages translation model (Dong et al., 2015), a multi-lingual translation model with a single attention mechanism shared across multiple languages (Firat et al., 2016), and in multitask sequence-to-sequence learning without an attention-based decoder (Luong et al., 2016). We explore the benefits of grounded learning in the specific case of multimodal translation. We combine sequence prediction with continuous (image) vector prediction, compared to previous work which multitasks different sequence prediction tasks.

Visual representation prediction has been studied using static images or videos. Lin and Parikh (2015) use a conditional random field to imagine the composition of a clip-art scene for visual paraphrasing and fill-in-the-blank tasks. Chrupała et al. (2015) predict the image vector associated with a sentence using an L2 loss; they found this improves multi-modal word similarity compared to text-only baselines. Gelderloos and Chrupała (2016) predict the image vector associated with a sequence of phonemes using a max-margin loss, similar to our image prediction objective. Collell et al. (2017) learn to predict the visual feature vector associated with a word for word similarity and relatedness tasks. As a video reconstruction problem, Srivastava et al. (2015) propose an LSTM Autoencoder to predict video frames as a reconstruction task or as a future prediction task. Pasunuru and Bansal (2017) propose a multitask model for video description that combines unsupervised video reconstruction, lexical entailment, and video description. They find improvements from using out-of-domain resources for entailment and video prediction, similar to the improvements we find from using out-of-domain parallel text and described images.

8 Conclusion

We decompose multimodal translation into two sub-problems: learning to translate and learning visually grounded representations. In a multitask learning framework, we show how these sub-problems can be addressed by sharing an encoder between a translation model and an image prediction model[5]. Our approach achieves state-of-the-art results on the Multi30K dataset without using images for translation. We show that training on separate parallel text and described image datasets does not hurt performance, encouraging future research on multitasking with diverse sources of data. Furthermore, we still find improvements from image prediction when we improve our text-only baseline with the out-of-domain parallel text. Future work includes adapting our decomposition to other NLP tasks that may benefit from out-of-domain resources, such as semantic role labelling, dependency parsing, and question-answering; exploring methods for inputting the (predicted) image into the translation model; experimenting with different image prediction architectures; multitasking different translation languages into a single shared encoder; and multitasking in both the encoder and decoder(s).

Acknowledgments

We are grateful to the anonymous reviewers for their feedback. We thank Joost Bastings for sharing his multitasking Nematus model, Wilker Aziz for discussions about formulating the problem, Stella Frank for finding and explaining the qualitative examples to us, and Afra Alishahi, Grzegorz Chrupała, and Philip Schulz for feedback on earlier drafts of the paper. DE acknowledges the support of an Amazon Research Award, NWO Vici grant nr. 277-89-002 awarded to K. Sima'an, and a hardware grant from the NVIDIA Corporation.

[5] Code: http://github.com/elliottd/imagination

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

J. Bingel and A. Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 164–169.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. DCU-UvA Multimodal MT System Report. In *Proceedings of the First Conference on Machine Translation*, pages 634–638.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1014.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. pages 1724–1734.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622.

Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 112–118.

Guillem Collell, Teddy Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 4378–4384.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

D. Dong, H. Wu, W. He, D. Yu, and H. Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1723–1732.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR*, abs/1510.04709.

Desmond Elliott, Stella Frank, Khalil. Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*.

O. Firat, K. Cho, and Y. Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29*, pages 1019–1027.

Lieke Gelderloos and Grzegorz Chrupała. 2016. From phonemes to images: levels of representation in a recurrent neural model of visually-grounded language learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 1309–1319.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2399–2409.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 639–645.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739.

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2016. Representation of linguistic form and function in recurrent neural networks. *arXiv preprint arXiv:1602.08952*.

Douwe Kiela, Anita L. Verő, and Stephen Clark. 2016. Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, pages 447–456.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Ondřej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. Mt-compareval: Graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, 104(1):63–74.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual meeting of Association for Computational Linguistics*, pages 177–180.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. Cuni system for wmt16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation*, pages 646–654.

Xiao Lin and Devi Parikh. 2015. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2984–2993.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *ICLR*.

Hideki Nakayama and Noriki Nishida. 2017. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

R. Pasunuru and M. Bansal. 2017. Multi-Task Video Captioning with Video and Entailment Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1273–1283.

Barbara Plank. 2016. Keystroke dynamics as signal for shallow syntactic parsing. In *26th International Conference on Computational Linguistics*, pages 609–619.

Amrita Saha, Mitesh M. Khapra, Sarath Chandar, Janarthanan Rajendran, and Kyunghyun Cho. 2016. A correlational encoder decoder architecture for pivot based sequence generation. In *26th International Conference on Computational Linguistics: Technical Papers*, pages 109–118.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. Valerio Miceli Barone, J. Mokry, and M. Nădejde. 2017. Nematus: a Toolkit for Neural Machine Translation. pages 65–68.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Rico Sennrich and Beat Kunz. 2014. Zmorge: A german morphological lexicon extracted from wiktionary. In *Language Resources and Evaluation Conference*, pages 1063–1067.

Kashif Shah, Josiah Wang, and Lucia Specia. 2016. Shef-multimodal: Grounding machine translation on images. In *Proceedings of the First Conference on Machine Translation*, pages 660–665.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pages 543–553.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning*, pages 843–852.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Joji Toyama, Masanori Misono, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. Neural machine translation with latent semantic of image and text. *CoRR*, abs/1611.08459.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. *ICLR*.