Multimodal Pivots for Image Caption Translation

Julian Hitschler and **Shigehiko Schamoni**
Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{hitschler,schamoni}@cl.uni-heidelberg.de

Stefan Riezler
Computational Linguistics & IWR
Heidelberg University
69120 Heidelberg, Germany
riezler@cl.uni-heidelberg.de

Abstract

We present an approach to improve statistical machine translation of image descriptions by multimodal pivots defined in visual space. The key idea is to perform image retrieval over a database of images that are captioned in the target language, and use the captions of the most similar images for crosslingual reranking of translation outputs. Our approach does not depend on the availability of large amounts of in-domain parallel data, but only relies on available large datasets of monolingually captioned images, and on state-of-the-art convolutional neural networks to compute image similarities. Our experimental evaluation shows improvements of 1 BLEU point over strong baselines.

1 Introduction

Multimodal data consisting of images and natural language descriptions (henceforth called *captions*) are an abundant source of information that has led to a recent surge in research integrating language and vision. Recently, the aspect of multilinguality has been added to multimodal language processing in a shared task at the WMT16 conference.[1] There is clearly also a practical demand for multilingual image captions, e.g., automatic translation of descriptions of art works would allow access to digitized art catalogues across language barriers and is thus of social and cultural interest; multilingual product descriptions are of high commercial interest since they would allow to widen e-commerce transactions automatically to international markets. However, while datasets of images and monolingual captions already include millions of tuples (Ferraro et al., 2015), the largest multilingual datasets of images and captions known to the authors contain 20,000 (Grubinger et al., 2006) or 30,000[2] triples of images with German and English descriptions.

In this paper, we want to address the problem of multilingual captioning from the perspective of statistical machine translation (SMT). In contrast to prior work on generating captions directly from images (Kulkarni et al. (2011), Karpathy and Fei-Fei (2015), Vinyals et al. (2015), *inter alia*), our goal is to integrate visual information into an SMT pipeline. Visual context provides orthogonal information that is free of the ambiguities of natural language, therefore it serves to disambiguate and to guide the translation process by grounding the translation of a source caption in the accompanying image. Since datasets consisting of source language captions, images, and target language captions are not available in large quantities, we would instead like to utilize large datasets of images and target-side monolingual captions to improve SMT models trained on modest amounts of parallel captions.

Let the task of *caption translation* be defined as follows: For production of a target caption e_i of an image i, a system may use as input an image caption for image i in the source language f_i, as well as the image i itself. The system may safely assume that f_i is relevant to i, i.e., the identification of relevant captions for i (Hodosh et al., 2013) is not itself part of the task of caption translation. In contrast to the inference problem of finding $\hat{e} = \operatorname{argmax}_e p(e|f)$ in text-based SMT, multimodal caption translation allows to take into consideration i as well as f_i in finding $\hat{e}_i$:

$$\hat{e}_i = \underset{e_i}{\operatorname{argmax}}\, p(e_i|f_i, i)$$

[1] http://www.statmt.org/wmt16/multimodal-task.html

[2] The dataset used at the WMT16 shared task is based on translations of Flickr30K captions (Rashtchian et al., 2010).

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2399–2409,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

In this paper, we approach caption translation by a general crosslingual reranking framework where for a given pair of source caption and image, monolingual captions in the target language are used to rerank the output of the SMT system. We present two approaches to retrieve target language captions for reranking by pivoting on images that are similar to the input image. One approach calculates image similarity based deep convolutional neural network (CNN) representations. Another approach calculates similarity in visual space by comparing manually annotated object categories. We compare the multimodal pivot approaches to reranking approaches that are based on text only, and to standard SMT baselines trained on parallel data. Compared to a strong baseline trained on 29,000 parallel caption data, we find improvements of over 1 BLEU point for reranking based on visual pivots. Notably, our reranking approach does not rely on large amounts of in-domain parallel data which are not available in practical scenarios such as e-commerce localization. However, in such scenarios, monolingual product descriptions are naturally given in large amounts, thus our work is a promising pilot study towards real-world caption translation.

2 Related Work

Caption generation from images alone has only recently come into the scope of realistically solvable problems in image processing (Kulkarni et al. (2011), Karpathy and Fei-Fei (2015), Vinyals et al. (2015), *inter alia*). Recent approaches also employ reranking of image captions by measuring similarity between image and text using deep representations (Fang et al., 2015). The tool of choice in these works are neural networks whose deep representations have greatly increased the quality of feature representations of images, enabling robust and semantically salient analysis of image content. We rely on the CNN framework (Socher et al., 2014; Simonyan and Zisserman, 2015) to solve semantic classification and disambiguation tasks in NLP with the help of supervision signals from visual feedback. However, we consider image captioning as a different task than caption translation since it is not given the information of the source language string. Therefore we do not compare our work to caption generation models.

In the area of SMT, Wäschle and Riezler (2015) presented a framework for integrating a large, in-domain, target-side monolingual corpus into machine translation by making use of techniques from crosslingual information retrieval. The intuition behind their approach is to generate one or several translation hypotheses using an SMT system, which act as queries to find matching, semantically similar sentences in the target side corpus. These can in turn be used as templates for refinement of the translation hypotheses, with the overall effect of improving translation quality. Our work can be seen as an extension of this method, with visual similarity feedback as additional constraint on the crosslingual retrieval model. Calixto et al. (2012) suggest using images as supplementary context information for statistical machine translation. They cite examples from the news domain where visual context could potentially be helpful in the disambiguation aspect of SMT and discuss possible features and distance metrics for context images, but do not report experiments involving a full SMT pipeline using visual context. In parallel to our work, Elliott et al. (2015) addressed the problem of caption translation from the perspective of neural machine translation.[3] Their approach uses a model which is considerably more involved than ours and relies exclusively on the availability of parallel captions as training data. Both approaches crucially rely on neural networks, where they use a visually enriched neural encoder-decoder SMT approach, while we follow a retrieval paradigm for caption translation, using CNNs to compute similarity in visual space.

Integration of multimodal information into NLP problems has been another active area of recent research. For example, Silberer and Lapata (2014) show that distributional word embeddings grounded in visual representations outperform competitive baselines on term similarity scoring and word categorization tasks. The orthogonality of visual feedback has previously been exploited in a multilingual setting by Kiela et al. (2015) (relying on previous work by Bergsma and Van Durme (2011)), who induce a bilingual lexicon using term-specific multimodal representations obtained by querying the Google image

[3] We replicated the results of Elliott et al. (2015) on the IAPR TC-12 data. However, we decided to not include their model as baseline in this paper since we found our hierarchical phrase-based baselines to yield considerably better results on IAPR TC-12 as well as on MS COCO.

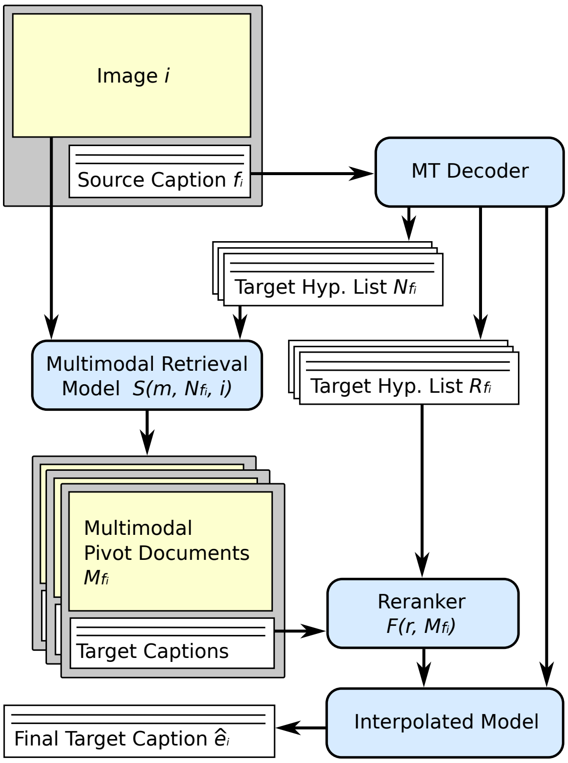


Image i
Source Caption f_i
MT Decoder
Target Hyp. List N_{f_i}
Multimodal Retrieval Model $S(m, N_{f_i}, i)$
Target Hyp. List R_{f_i}
Multimodal Pivot Documents M_{f_i}
Target Captions
Reranker $F(r, M_{f_i})$
Interpolated Model
Final Target Caption $\hat{e}_i$

Figure 1: Overview of model architecture.

search engine.[4] Funaki and Nakayama (2015) use visual similarity for crosslingual document retrieval in a multimodal and bilingual vector space obtained by generalized canonical correlation analysis, greatly reducing the need for parallel training data. The common element is that CNN-based visual similarity information is used as a "hub" (Funaki and Nakayama, 2015) or pivot connecting corpora in two natural languages which lack direct parallelism, a strategy which we apply to the problem of caption translation.

3 Models

3.1 Overview

Following the basic approach set out by Wäschle and Riezler (2015), we use a crosslingual retrieval model to find sentences in a target language document collection C, and use these to rerank target language translations e of a source caption f.

The systems described in our work differ from that of Wäschle and Riezler (2015) in a number of aspects. Instead of a two-step architecture of coarse-grained and fine-grained retrieval, our system uses relevance scoring functions for retrieval of matches in the document collection C, and for reranking of translation candidates that are based on inverse document frequency of terms (Spärck Jones, 1972) and represent variants of the popular TF-IDF relevance measure.

A schematic overview of our approach is given in Figure 1. It consists of the following components:

Input: Source caption f_i, image i, target-side collection C of image-captions pairs

Translation: Generate unique list N_{f_i} of k_n-best translations, generate unique list R_{f_i} of k_r-best list of translations[5] using MT decoder

Multimodal retrieval: For list of translations N_{f_i}, find set M_{f_i} of k_m-most relevant pairs of images and captions in a target-side collection C, using a heuristic relevance scoring function $S(m, N_{f_i}, i), m \in C$

Crosslingual reranking: Use list M_{f_i} of image-caption pairs to rerank list of translations R_{f_i}, applying relevance scoring function $F(r, M_{f_i})$ to all $r \in R_{f_i}$

Output: Determine best translation hypothesis $\hat{e}_i$ by interpolating decoder score d_r for a hypothesis $r \in R_{f_i}$ with its relevance score $F(r, M_{f_i})$ with weight λ s.t.

$$\hat{e}_i = \underset{r \in R_{f_i}}{\text{argmax}}\, d_r + \lambda \cdot F(r, M_{f_i})$$

The central concept is the scoring function $S(m, N_{f_i}, i)$ which defines three variants of target-side retrieval (TSR), all of which make use of the procedure outlined above. In the baseline text-based reranking model (TSR-TXT), we use relevance scoring function S_{TXT}. This function is purely text-based and does not make use of multimodal context information (as such, it comes closest to the models used for target-side retrieval in Wäschle and Riezler (2015)). In the retrieval model enhanced by visual information from a deep convolutional neural network (TSR-CNN), the scoring function S_{CNN} incorporates a textual relevance score with visual similarity information extracted from the neural network. Finally, we evaluate these models against a relevance score based on human object-category annotations (TSR-HCA), using the scoring function

[4] `https://images.google.com/`

[5] In practice, the first hypothesis list may be reused. We distinguish between the two hypothesis lists N_{f_i} and R_{f_i} for notational clarity since in general, the two hypothesis lists need not be of equal length.

S_{HCA}. This function makes use of the object annotations available for the MS COCO corpus (Lin et al., 2014) to give an indication of the effectiveness of our automatically extracted visual similarity metric. The three models are discussed in detail below.

3.2 Target Side Retrieval Models

Text-Based Target Side Retrieval. In the TSR-TXT retrieval scenario, a match candidate $m \in C$ is scored in the following way:

$$S_{TXT}(m, N_{f_i}) = \\ Z_m \sum_{n \in N_{f_i}} \sum_{w_n \in tok(n)} \sum_{w_m \in typ(m)} \delta(w_m, w_n) idf(w_m),$$

where δ is the Kronecker δ-function, N_{f_i} is the set of the k_n-best translation hypotheses for a source caption f_i of image i by decoder score, $typ(a)$ is a function yielding the set of types (unique tokens) contained in a caption a,[6] $tok(a)$ is a function yielding the tokens of caption a, $idf(w)$ is the inverse document frequency (Spärck Jones, 1972) of term w, and $Z_m = \frac{1}{|typ(m)|}$ is a normalization term introduced in order to avoid biasing the system towards long match candidates containing many low-frequency terms. Term frequencies were computed on monolingual data from Europarl (Koehn, 2005) and the News Commentary and News Discussions English datasets provided for the WMT15 workshop.[7] Note that in this model, information from the image i is not used.

Multimodal Target Side Retrieval using CNNs. In the TSR-CNN scenario, we supplement the textual target-side TSR model with visual similarity information from a deep convolutional neural network. We formalize this by introduction of the positive-semidefinite distance function $v(i_x, i_y) \rightarrow [0, \infty)$ for images i_x, i_y (smaller values indicating more similar images). The relevance scoring function S_{CNN} used in this model takes the following form:

$$S_{CNN}(m, N_{f_i}, i) \\ = \begin{cases} S_{TXT}(m, N_{f_i}) e^{-bv(i_m, i)}, & v(i_m, i) < d \\ 0 & otherwise, \end{cases}$$

where i_m is the image to which the caption m refers and d is a cutoff maximum distance, above which match candidates are considered irrelevant, and b is a weight term which controls the impact of the visual distance score $v(i_m, i)$ on the overall score.[8]

Our visual distance measure v was computed using the VGG16 deep convolutional model of Simonyan and Zisserman (2015), which was pre-trained on ImageNet (Russakovsky et al., 2014). We extracted feature values for all input and reference images from the penultimate fully-connected layer (`fc7`) of the model and computed the Euclidean distance between feature vectors of images. If no neighboring images fell within distance d, the text-based retrieval procedure S_{TXT} was used as a fallback strategy, which occurred 47 out of 500 times on our test data.

Target Side Retrieval by Human Category Annotations. For contrastive purposes, we evaluated a TSR-HCA retrieval model which makes use of the human object category annotations for MS COCO. Each image in the MS COCO corpus is annotated with object polygons classified into 91 categories of common objects. In this scenario, a match candidate m is scored in the following way:

$$S_{HCA}(m, N_{f_i}, i) \\ = \delta(cat(i_m), cat(i)) S_{TXT}(m, N_{f_i}),$$

where $cat(i)$ returns the set of object categories with which image i is annotated. The amounts to enforcing a strict match between the category annotations of i and the reference image i_m, thus pre-filtering the S_{TXT} scoring to captions for images with strict category match.[9] In cases where i was annotated with a unique set of object categories and thus no match candidates with nonzero scores were returned by S_{HCA}, S_{TXT} was used as a fallback strategy, which occurred 77 out of 500 times on our test data.

[6] The choice for per-type scoring of reference captions was primarily driven by performance considerations. Since captions rarely contain repetitions of low-frequency terms, this has very little effect in practice, other than to mitigate the influence of stopwords.

[7] `http://www.statmt.org/wmt15/translation-task.html`

[8] The value of $b = 0.01$ was found on development data and kept constant throughout the experiments.

[9] Attempts to relax this strict matching criterion led to strong performance degradation on the development test set.

3.3 Translation Candidate Re-scoring

The relevance score $F(r, M_{f_i})$ used in the reranking model was computed in the following way for all three models:

$$F(r, M_{f_i}) = Z_{M_{f_i}} \sum_{m \in M_{f_i}} \sum_{w_m \in typ(m)} \sum_{w_r \in tok(r)} \delta(w_m, w_r) idf(w_m)$$

with normalization term

$$Z_{M_{f_i}} = (\sum_{m \in M_{f_i}} |tok(m)|)^{-1},$$

where r is a translation candidate and M_{f_i} is a list of k_m-top target side retrieval matches. Because the model should return a score that is reflective of the relevance of r with respect to M_{f_i}, irrespective of the length of M_{f_i}, normalization with respect to the token count of M_{f_i} is necessary. The term $Z_{M_{f_i}}$ serves this purpose.

4 Experiments

4.1 Bilingual Image-Caption Data

We constructed a German-English parallel dataset based on the MS COCO image corpus (Lin et al., 2014). 1,000 images were selected at random from the 2014 training section[10] and, in a second step, one of their five English captions was chosen randomly. This caption was then translated into German by a native German speaker. Note that our experiments were performed with German as the source and English as the target language, therefore, our reference data was not produced by a single speaker but reflects the heterogeneity of the MS COCO dataset at large. The data was split into a development set of 250 captions, a development test set of 250 captions for testing work in progress, and a test set of 500 captions. For our retrieval experiments, we used only the images and captions that were not included in the development, development test or test data, a total of 81,822 images with 5 English captions per image. All data was tokenized and converted to lower case using the `cdec`[11] utilities `tokenized-anything.pl` and `lowercase.pl`. For the German data, we performed compound-splitting using the method described by Dyer (2009), as implemented by the `cdec` utility `compound-split.pl`. Table 1 gives an overview of the dataset. Our parallel development, development test and test data is publicly available.[12]

Section	Images	Captions	Languages
DEV	250	250	DE-EN
DEVTEST	250	250	DE-EN
TEST	500	500	DE-EN
RETRIEVAL (C)	81,822	409,110	EN

Table 1: Number of images and sentences in MS COCO image and caption data used in experiments.

4.2 Translation Baselines

We compare our approach to two baseline machine translation systems, one trained on out-of-domain data exclusively and one domain-adapted system. Table 2 gives an overview of the training data for the machine translation systems.

Out-of-Domain Baseline. Our baseline SMT framework is hierarchical phrase-based translation using synchronous context free grammars (Chiang, 2007), as implemented by the `cdec` decoder (Dyer et al., 2010). Data from the Europarl (Koehn, 2005), News Commentary and Common Crawl corpora (Smith et al., 2013) as provided for the WMT15 workshop was used to train the translation model, with German as source and English as target language.

Like the retrieval dataset, training, development and test data was tokenized and converted to lower case, using the same `cdec` tools. Sentences with lengths over 80 words in either the source or the target language were discarded before training. Source text compound splitting was performed using `compound-split.pl`. Alignments were extracted bidirectionally using the `fast-align` utility of `cdec` and symmetrized with the `atools` utility (also part of `cdec`) using the `grow-diag-final-and` symmetrization heuristic. The alignments were then used by the `cdec` grammar extractor to extract a synchronous context free grammar from the parallel data.

[10] We constructed our parallel dataset using only the training rather than the validation section of MS COCO so as to keep the latter pristine for future work based on this research.

[11] `https://github.com/redpony/cdec`

[12] `www.cl.uni-heidelberg.de/decoco/`

Corpus	Sentences	Languages	System
Europarl	1,920,209	DE-EN	O/I
News Commentary	216,190	DE-EN	O/I
Common Crawl	2,399,123	DE-EN	O/I
Flickr30k WMT16	29,000	DE-EN	I
Europarl	2,218,201	EN	O/I
News Crawl	28,127,448	EN	O/I
News Discussions	57,803,684	EN	O/I
Flickr30k WMT16	29,000	EN	I

Table 2: Parallel and monolingual data used for training machine translation systems. Sentence counts are given for raw data without preprocessing. O/I: both out-of-domain and in-domain system, I: in-domain system only.

The target language model was trained on monolingual data from Europarl, as well as the News Crawl and News Discussions English datasets provided for the WMT15 workshop (the same data as was used for estimating term frequencies for the retrieval models) with the KenLM toolkit (Heafield et al., 2013; Heafield, 2011).[13]

We optimized the parameters of the translation system for translation quality as measured by IBM BLEU (Papineni et al., 2002) using the Margin Infused Relaxed Algorithm (MIRA) (Crammer and Singer, 2003). For tuning the translation models used for extraction of the hypothesis lists for final evaluation, MIRA was run for 20 iterations on the development set, and the best run was chosen for final testing.

In-Domain Baseline. We also compared our models to a domain-adapted machine translation system. The domain-adapted system was identical to the out-of-domain system, except that it was supplied with additional parallel training data from the image caption domain. For this purpose, we used 29,000 parallel German-English image captions as provided for the WMT16 shared task on multimodal machine translation. The English captions in this dataset belong to the Flickr30k corpus (Rashtchian et al., 2010) and are very similar to those of the MS COCO corpus. The German captions are expert translations. The English captions were also used as additional training data for the target-side language model. We generated k_n- and k_r-best lists of translation candidates using this in-domain baseline system.

Model	k_n	k_m	k_r	λ
TSR-TXT	300	500	5	$5 \cdot 10^4$
TSR-CNN	300	300	5	$70 \cdot 10^4$
TSR-HCA	300	500	5	$10 \cdot 10^4$

Table 3: Optimized hyperparameter values used in final evaluation.

4.3 Optimization of TSR Hyperparameters

For each of our retrieval models, we performed a step-wise exhaustive search of the hyperparameter space over the four system hyperparameters for IBM BLEU on the development set: The length of the k_n-best list the entries of which are used as queries for retrieval; the number of k_m-best-matching captions retrieved; the length of the final k_r-best list used in reranking; the interpolation weight λ of the relevance score F relative to the translation hypothesis log probability returned by the decoder. The parameter ranges to be explored were determined manually, by examining system output for prototypical examples. Table 3 gives an overview over the hyperparameter values obtained.

For TSR-CNN, we initially set the cutoff distance d to 90.0, after manually inspecting sets of nearest neighbors returned for various maximum distance values. After optimization of retrieval parameters, we performed an exhaustive search from $d = 80.0$ to $d = 100.0$, with step size 1.0 on the development set, while keeping all other hyperparameters fixed, which confirmed out initial choice of $d = 90.0$ as the optimal value.

Explored parameter spaces were identical for all models and each model was evaluated on the test set using its own optimal configuration of hyperparameters.

4.4 Significance Testing

Significance tests on the differences in translation quality were performed using the approximate randomization technique for measuring performance differences of machine translation systems described in Riezler and Maxwell (2005) and implemented by Clark et al. (2011) as part of the Multeval toolkit.[14]

[13] https://kheafield.com/code/kenlm/

[14] https://github.com/jhclark/multeval

System	BLEU ↑	p_c	p_t	p_d	p_o
cdec out-dom.	25.5				
cdec in-dom.	29.6				0.00
TSR-TXT	29.7			0.45	0.00
TSR-CNN	**30.6**		0.04	0.02	0.00
TSR-HCA	**30.3**	0.42	0.01	0.00	0.00
System	METEOR ↑	p_c	p_t	p_d	p_o
cdec out-dom.	31.7				
cdec in-dom.	34.0				0.00
TSR-TXT	34.1			0.41	0.00
TSR-CNN	**34.7**		0.00	0.00	0.00
TSR-HCA	**34.4**	0.09	0.00	0.00	0.00
System	TER ↓	p_c	p_t	p_d	p_o
cdec out-dom.	49.3				
cdec in-dom.	46.1				0.00
TSR-TXT	45.8			0.12	0.00
TSR-CNN	**45.1**		0.03	0.00	0.00
TSR-HCA	**45.3**	0.34	0.02	0.00	0.00

Table 4: Metric scores for all systems and their significance levels as reported by Multeval. p_o-values are relative to the cdec out-of-domain baseline, p_d-values are relative to the cdec in-domain baseline, p_t-values are relative to TSR-TXT and p_c-values are relative to TSR-CNN. Best results are reported in **bold** face.[15]

4.5 Experimental Results

Table 4 summarizes the results for all models on an unseen test set of 500 captions. Domain adaptation led to a considerable improvement of +4.1 BLEU and large improvements in terms of METEOR and Translation Edit Rate (TER). We found that the target-side retrieval model enhanced with multimodal pivots from a deep convolutional neural network, TSR-CNN and TSR-HCA, consistently outperformed both the domain-adapted cdec baseline, as well as the text-based target side retrieval model TSR-TXT. These models therefore achieve a performance gain which goes beyond the effect of generic domain-adaptation. The gain in performance for TSR-CNN and TSR-HCA was significant at $p < 0.05$ for BLEU, METEOR, and TER. For all evaluation metrics, the difference between TSR-CNN and TSR-HCA was not significant, demonstrating that retrieval using our CNN-derived distance metric could match retrieval based the human object category annotations.

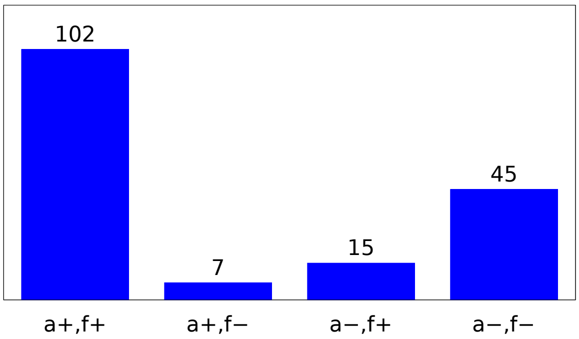


Figure 2: Results of the human pairwise preference ranking experiment, given as the joint distribution of both rankings: $a+$ denotes preference for TSR-CNN in terms of accuracy, $f+$ in terms of fluency; $a-$ denotes preference for the in-domain baseline in terms of accuracy, $f-$ in terms of fluency.

The text-based retrieval baseline TSR-TXT never significantly outperformed the in-domain cdec baseline, but there were slight nominal improvements in terms of BLEU, METEOR and TER. This finding is actually consistent with Wäschle and Riezler (2015) who report performance gains for text-based, target side retrieval models only on highly technical, narrow-domain corpora and even report performance degradation on medium-diversity corpora such as Europarl. Our experiments show that it is the addition of visual similarity information by incorporation of multimodal pivots into the image-enhanced models TSR-CNN and TSR-HCA which makes such techniques effective on MS COCO, thus upholding our hypothesis that visual information can be exploited for improvement of caption translation.

4.6 Human Evaluation

The in-domain baseline and TSR-CNN differed in their output in 169 out of 500 cases on the test set. These 169 cases were presented to a human judge alongside the German source captions in a double-blinded pairwise preference ranking experiment. The order of presentation was randomized for the two systems. The judge was asked to rank fluency and accuracy of the translations independently. The results are given in Figure 2. Overall, there was a clear preference for the output of TSR-CNN.

[15] A baseline for which a random hypothesis was chosen from the top-5 candidates of the in-domain system lies between the other two baseline systems: 27.5 / 33.3 / 47.7 (BLEU / METEOR / TER).

4.7 Examples

Table 5 shows example translations produced by both cdec baselines, TSR-TXT, TSR-CNN, and TSR-HCA, together with source caption, image, and reference translation. The visual information induced by target side captions of pivot images allows a disambiguation of translation alternatives such as “skirt” versus “rock (music)” for the German “Rock”, “pole” versus “mast” for the German “Masten”, and is able to repair mistranslations such as “foot” instead of “mouth” for the German “Maul”.

5 Conclusions and Further Work

We demonstrated that the incorporation of multimodal pivots into a target-side retrieval model improved SMT performance compared to a strong in-domain baseline in terms of BLEU, METEOR and TER on our parallel dataset derived from MS COCO. The gain in performance was comparable between a distance metric based on a deep convolutional network and one based on human object category annotations, demonstrating the effectiveness of the CNN-derived distance measure. Using our approach, SMT can, in certain cases, profit from multimodal context information. Crucially, this is possible without using large amounts of in-domain parallel text data, but instead using large amounts of monolingual image captions that are more readily available.

Learning semantically informative distance metrics using deep learning techniques is an area under active investigation (Wu et al., 2013; Wang et al., 2014; Wang et al., 2015). Despite the fact that our simple distance metric performed comparably to human object annotations, using such high-level semantic distance metrics for caption translation by multimodal pivots is a promising avenue for further research.

The results were achieved on one language pair (German-English) and one corpus (MS COCO) only. As with all retrieval-based methods, generalized statements about the relative performance on corpora of various domains, sizes and qualities are difficult to substantiate. This problem is aggravated in the multimodal case, since the relevance of captions with respect to images varies greatly between different corpora (Hodosh et al., 2013). In future work, we plan to evaluate our approach in more naturalistic settings, such machine translation for captions in online multimedia repositories

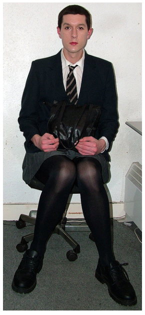
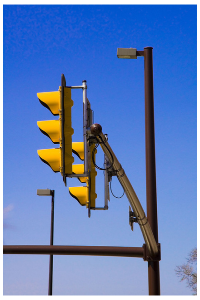
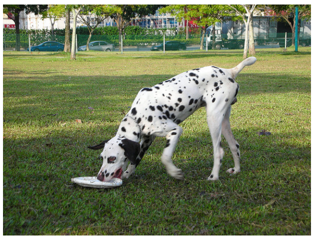
Image:	
Source:	Eine Person in einem Anzug und Krawatte und einem Rock.
cdec out-dom:	a person in a suit and tie and a rock .
cdec in-dom:	a person in a suit and tie and a rock .
TSR-TXT:	a person in a suit and tie and a rock .
TSR-CNN:	a person in a suit and tie and a skirt .
TSR-HCA:	a person in a suit and tie and a rock .
Reference:	a person wearing a suit and tie and a skirt
Image:	
Source:	Ein Masten mit zwei Ampeln für Autofahrer.
cdec out-dom:	a mast with two lights for drivers .
cdec in-dom:	a mast with two lights for drivers .
TSR-TXT:	a mast with two lights for drivers .
TSR-CNN:	a pole with two lights for drivers .
TSR-HCA:	a pole with two lights for drivers .
Reference:	a pole has two street lights on it for drivers .
Image:	
Source:	Ein Hund auf einer Wiese mit einem Frisbee im Maul.
cdec out-dom:	a dog on a lawn with a frisbee in the foot .
cdec in-dom:	a dog with a frisbee in a grassy field .
TSR-TXT:	a dog with a frisbee in a grassy field .
TSR-CNN:	a dog in a grassy field with a frisbee in its mouth .
TSR-HCA:	a dog with a frisbee in a grassy field .
Reference:	a dog in a field with a frisbee in its mouth

Table 5: Examples for improved caption translation by multimodal feedback.

such as Wikimedia Commons[16] and digitized art catalogues, as well as e-commerce localization.

A further avenue of future research is improving models such as that presented in Elliott et al. (2015) by crucial components of neural MT such as "attention mechanisms". For example, the attention mechanism of Bahdanau et al. (2015) serves as a soft alignment that helps to guide the translation process by influencing the sequence in which source tokens are translated. A similar mechanism is used in Xu et al. (2015) to decide which part of the image should influence which part of the generated caption. Combining these two types of attention mechanisms in a neural caption translation model is a natural next step in caption translation. While this is beyond the scope of this work, our models should provide an informative baseline against which to evaluate such methods.

Acknowledgments

This research was supported in part by DFG grant RI-2221/2-1 "Grounding Statistical Machine Translation in Perception and Action", and by an Amazon Academic Research Award (AARA) "Multimodal Pivots for Low Resource Machine Translation in E-Commerce Localization".

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California, USA.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Barcelona, Spain.

Iacer Calixto, Teófilo de Compos, and Lucia Specia. 2012. Images as context in statistical machine translation. In *Proceedings of the Workshop on Vision and Language (VL)*, Sheffield, England, United Kingdom.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the Association for Computational Lingustics (ACL)*, Portland, Oregon, USA.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Chris Dyer. 2009. Using a maximum entropy model to build segmentation lattices for mt. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, Boulder, Colorado, USA.

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

Hao Fang, Li Deng, Margaret Mitchell, Saurabh Gupta, Piotr Dollar, John C. Platt, Forrest Iandola, Jianfeng Gao, C. Lawrence Zitnick, Rupesh K. Srivastava, Xiaodeng He, and Geoffrey Zweit. 2015. From captions to visual concepts and back. In *In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao (Kenneth) Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: A new evaluatioin resource for visual information systems. In *In Proceedings of LREC*, Genova, Italy.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT)*, Edinburgh, Scotland, United Kingdom.

[16] `https://commons.wikimedia.org/wiki/Main_Page`

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.

Andrey Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachusetts, USA.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit*, Phuket, Thailand.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, Colorado, USA.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *Computing Research Repository*, abs/1405.0312.

Kishore Papineni, Salim Roukos, Todd Ard, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania, USA.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, USA.

Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Methods for MT and Summarization (MTSE) at the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, USA.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2014. Imagenet large scale visual recognition challenge. *Computing Research Repository*, abs/1409.0575.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland, USA.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA.

Jason Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2(1):207–218.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston,Massachusetts, USA.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio, USA.

Zhaowen Wang, Jianchao Yang, Zhe Lin, Jonathan Brandt, Shiyu Chang, and Thomas Huang. 2015. Scalable similarity learning using large margin neighborhood embedding. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Washington, DC, USA.

Katharina Wäschle and Stefan Riezler. 2015. Integrating a large, monolingual corpus as translation memory into statistical machine translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*, Antalya, Turkey.

Pengcheng Wu, Steven C.H. Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. 2013. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia*, Barcelona, Spain.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, Lille, France.