# Multimodal Neural Machine Translation of Fashion E-Commerce Descriptions

Katrien Laenen(✉) and Marie-Francine Moens

KU Leuven, Louvain, Belgium
{katrien.laenen, sien.moens}@kuleuven.be

**Abstract.** Neural networks become extremely popular in artificial intelligence. In this paper we show how they aid in automatically translating fashion item descriptions and how they use fashion images to generate the translations. More specifically, we propose a multimodal neural machine translation model in which the decoder that generates the translation attends to visually grounded representations that capture both the semantics of the fashion words in the source language and regions in the fashion image. We introduce this novel neural architecture in the context of fashion e-commerce, where product descriptions need to be available in multiple languages. We report state-of-the-art multimodal translation results on a real-world fashion e-commerce dataset.

**Keywords:** Multimodal neural machine translation · Multimodal multilingual space · Alignment model · Stacked cross-attention · Fashion e-commerce

## 1 Introduction

Internationalisation is considered as a big trend in e-commerce. There is an increasing interest by e-commerce businesses to expand to other countries. Language is here an important barrier. E-retailers struggle to efficiently translate their product descriptions and websites in a variety of languages. Currently, this is still done manually. However, consumers prefer to read product descriptions in their native language to get an optimal understanding of the product specifications and to be able to compare products.

Neural machine translation (NMT) is an approach to machine translation which uses an artificial neural network to predict a sequence of words in the target language given a sequence of words in the source language. In multimodal neural machine translation (MNMT), the source sequence is paired with an image and the target sequence is generated aided by the information in the image. The fashion e-commerce domain, where product descriptions reference to fine-grained product attributes somewhere in the image (e.g., V-neck, floral print), is a challenging but interesting domain for MNMT which requires to efficiently integrate the visual and textual information. State-of-the-art NMT systems are sequence-to-sequence networks with an attention-based encoder-decoder architecture. The encoder encodes each source word with a vector representation which captures the word's semantics. At each timestep, the decoder outputs the most likely target word by looking at the source word representations and the target words generated in previous timesteps. In this work, we propose a

N. Kalbaska et al. (Eds.): FACTUM 2019, *Fashion Communication in the Digital Age*, pp. 46–57, 2019.
https://doi.org/10.1007/978-3-030-15436-3_4

MNMT model which jointly learns to align semantically related source words, target words and image regions and to translate. Hence, it infers a multimodal, multilingual space where a source word, target word and image region that refer to the same fashion attribute have vector representations which are close together. This way the source word representations become visually contextualised or *visually grounded*, which informs the decoder about the visual context in an efficient way.

The main contributions of our paper are:

- We infer a multimodal, multilingual space in which we embed an image region, source word and target word that refer to the same fashion attribute close together. In this space, they are aligned through an attention-based alignment model which uses cosine similarity to measure semantic relatedness. Next, the decoder attends to the inferred visually grounded representations of our source words.
- We propose a new, natural setting for multimodal translation, that is fashion e-commerce, which is challenging because of its references to fine-grained fashion attributes and the limited amount of training data.
- We show state-of-the art multimodal translation results on a real-world fashion e-commerce dataset.

The remainder of this paper is structured as follows. In Sect. 2 we review other work related to the subject of this paper. Next, we elaborate our model architecture in Sect. 3. In Sect. 4 we describe our experimental setup. The results of the conducted experiments can be found in Sect. 5. Finally, we present our conclusions and provide directions for future work in Sect. 6.

## 2 Related Work

Unimodal machine translation models are trained with pairs of sentences, where the target language sentence is the translation of the source language sentence. Currently, neural machine translation is the most popular and successful technique. The neural networks are in the form of sequence-to-sequence networks with an attention-based encoder-decoder architecture. [2] were the first to introduce an attention mechanism in the decoder. The intuition behind it is to compute the expected alignment of every source word with the next target word and to jointly translate. The pure text-based model of [2] will serve as our unimodal neural machine translation (UNMT) baseline.

There is a current interest in MNMT and more specifically in using additional visual information to aid the translation [3–5, 7, 9, 19]. Although these works achieve promising results, they indicate that further exploration to what is the best way to benefit from the visual context is needed. One approach in MNMT is to use a double attention mechanism, one over the source words and another over different regions of the image [3, 5]. However, this approach neglects to exploit the semantic relatedness between the image regions, source words and target words which is an important indicator for the relevance of the visual information. Our approach makes use of an additional alignment model to align the image regions, source words and target words to infer visually grounded source word representations. This is different from [9] who project the visual features to the space of source word embeddings and append these visual words to the head/tail of the source sentence. The encoder then encodes both these visual words and the source words. In contrast to our work, they do not use an alignment model to infer their multimodal space and do not attempt to include the target language in this space. Most closely related to our work is the work of [19] where the visual context is grounded into the encoder through the joint learning of a multimodal space and of a translation model. More precisely, they embed images close to their attended source sentence representations in a multimodal shared space. Additionally, they initialise the decoder hidden state in such a way that the source words closest related to the visual context have more influence during decoding. In contrast, we do not embed full images and sentences in our shared space, but instead work at a finer level to find the latent alignment of image regions and words, which proves to be valuable especially for fashion data. Moreover, we also include the target language to obtain a space which is both multimodal and multilingual [19] report the state-of-the-art results for MNMT and therefore we use their model as our MNMT baseline.

In order to find the semantic correspondences between the image regions, source words and target words we make use of an alignment model. Alignment models have already proven to be useful for other tasks that require to jointly reason over vision and language, such as image captioning [10], visual question answering [1, 17], multimodal search [11] and image-text matching [12, 18].

Neural networks and deep learning models have become an essential item in the toolbox of fashion-related businesses (e.g., in apparel recognition, fashion search, product recommendation and outfit combination). Closer to this work is the work of [13] who generate persuasive textual descriptions of fashion items given a number of key terms that describe the item in order to encourage an online buyer towards a successful purchase. However, their neural architecture ignores the image when generating the persuasive descriptions. The neural architecture proposed in this paper could expand the work of [13] in multimodal and multilingual settings.

## 3 Methodology

First, we describe the baseline models for UNMT and MNMT in respectively Sects. 3.1 and 3.2. Next, we elaborate our proposed MNMT architecture which aligns the image regions, source words and target words with stacked cross-attention in Sect. 3.2. In all formulas, matrices are written with capital letters and vectors are bolded. We use letters $W$ and $\boldsymbol{b}$ to refer to respectively the weights and bias in linear and non-linear transformations.

During the training phase, all models learn from a training set of examples of paired descriptions in source and target language. The MNMT models also have access to a corresponding image. During the testing phase, the models only have access to the source sentence and image.

### 3.1 UNMT Baseline

In UNMT the goal is to translate a source sentence $X = (x_1, x_2, \ldots, x_M)$ consisting of $M$ words into the correct target sentence $Y = (y_1, y_2, \ldots, y_N)$ consisting of $N$ words. Our UNMT baseline is the attention-based encoder-decoder architecture of [2]. For more details, the reader is referred to [2].

### 3.2 MNMT Baseline

In MNMT, a source sentence $X = (x_1, x_2, \ldots, x_M)$ is translated into a target sentence $Y = (y_1, y_2, \ldots, y_N)$ aided by the visual information in image $I$ paired with source sentence $X$. Our MNMT baseline is the model of [19]. The model obtains visually grounded source word representations by sharing the encoder between the translation task and a multimodal space inference task.

**Encoder.** The encoder is a bidirectional recurrent neural network (BRNN) [15] with gated recurrent units (GRUs) [6]. It produces a source word representation $\boldsymbol{s}_j \in \mathbb{R}^{2d_x}$ for each word $\boldsymbol{x}_j$ of source sentence $X$ by concatenating the forward and backward hidden states.

**Shared Space Inference Task.** The objective is to infer a shared space for images and source sentences which captures the semantic meaning across the two modalities. Each image is represented with vector $\boldsymbol{v} \in \mathbb{R}^{2048}$ obtained from the *pool5* layer of the convolutional neural network ResNet50 [8] pre-trained on ImageNet [14]. The representation of the source sentence $\boldsymbol{s}_{att}$ is obtained by applying attention to each source word representation $\boldsymbol{s}_j$ with image representation $\boldsymbol{v}$. This produces attention scores $z_j$ which measure how well the source word at position $j$ corresponds with the image. Next, the attention scores $z_j$ are normalized with the softmax function and used to weight the source words $\boldsymbol{s}_j$. This way, the words which are more related to the image content get a higher weight in the generated source sentence representation:

$$z_j = \tanh\left(W_s \boldsymbol{s}_j\right) . \tanh(W_v \boldsymbol{v}) \tag{1}$$

$$\boldsymbol{s}_{att} = \sum_{j=1}^{M} \beta_j \boldsymbol{s}_j, \text{ with } \beta_j = \operatorname{softmax}([z_1, z_2, \ldots, z_M])_j \tag{2}$$

Next, image $\boldsymbol{v}$ and source sentence $\boldsymbol{s}_{att}$ are projected to their representations $\widehat{\boldsymbol{v}}$ and $\widehat{\boldsymbol{s}}$ in the multimodal space:

$$\widehat{\boldsymbol{v}} = \tanh(W_{v\,emb} \boldsymbol{v} + \boldsymbol{b}_{v\,emb}) \tag{3}$$

$$\widehat{\boldsymbol{s}} = \tanh(W_{s\,emb} \boldsymbol{s}_{att} + \boldsymbol{b}_{s\,emb}) \tag{4}$$

with $\widehat{\boldsymbol{v}}, \widehat{\boldsymbol{s}} \in \mathbb{R}^d$. The projection to the multimodal space is learned by minimizing a triplet loss which enforces that a corresponding image-sentence pair should be closer than a non-corresponding pair:

$$\begin{aligned}\mathcal{L}_{triplet1} = &\sum_{e}^{E}\sum_{e' \neq e}^{E} \max(0, m - f(\widehat{\boldsymbol{v}}_e, \widehat{\boldsymbol{s}}_e) + f(\widehat{\boldsymbol{v}}_e, \widehat{\boldsymbol{s}}_{e'})) \\ &+ \sum_{e}^{E}\sum_{e' \neq e}^{E} \max(0, m - f(\widehat{\boldsymbol{v}}_e, \widehat{\boldsymbol{s}}_e) + f(\widehat{\boldsymbol{v}}_{e'}, \widehat{\boldsymbol{s}}_e))\end{aligned} \tag{5}$$

where index $e$ ranges over the number of training examples and $m$ is the margin. In the multimodal space, cosine similarity $f(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T.\boldsymbol{y}}{||\boldsymbol{x}||.||\boldsymbol{y}||}$ measures semantic relatedness.

**Translation Task.** The visually grounded source word representations $\boldsymbol{s}_j$ are used by the decoder, which is a conditional GRU [16] consisting of two stacked GRUs. At each timestep $t$, the decoder produces the next target word $y_t$ starting from the previously emitted word $y_{t-1}$, the previous decoder hidden state $\boldsymbol{h}_{t-1}$ and the source context vector $\boldsymbol{c}_t^{att}$:

$$\boldsymbol{o}_t = \tanh\left(E_y y_{t-1} + W_h \boldsymbol{h}_t + W_c \boldsymbol{c}_t^{att}\right) \tag{6}$$

$$P\left(y_t | y_{t-1}, \boldsymbol{h}_t, \boldsymbol{c}_t^{att}\right) = \text{softmax}(W_{out} \boldsymbol{o}_t) \tag{7}$$

where $E_y y_{t-1} \in \mathbb{R}^{d_y}$ is the vector representation of the previously emitted word and context vector $\boldsymbol{c}_t^{att}$ is acquired by applying Bahdanau's attention [2] on the source word representations $\boldsymbol{s}_j$ based on the decoder hidden state proposal $\boldsymbol{h}_t^{'}$ from the first GRU. At timestep $t = 0$ the decoder hidden state $\boldsymbol{h}_0$ is initialized such that the source words most closely related to the image have a bigger influence during translation decoding. More precisely, $\boldsymbol{h}_0$ is computed as a weighed sum of the attended source sentence representation $\boldsymbol{s}_{att}$ and the mean of the source word representations $\boldsymbol{s}_j$:

$$\boldsymbol{h}_0 = \tanh\left(W_{init}\left(\lambda \boldsymbol{s}_{att} + (1 - \lambda)\frac{1}{M}\sum_{j=1}^{M} \boldsymbol{s}_j\right)\right) \tag{8}$$

with weight $\lambda$ a hyperparameter. During training, we quantify the quality of the translation with the cross entropy loss:

$$\mathcal{L}_{cross-entropy} = -\sum_{e}^{E}\sum_{t}^{T} \boldsymbol{y}_{et}.\log(\overline{\boldsymbol{y}}_{et}) \tag{9}$$

where indices $e$ and $t$ range over respectively the number of training examples and number of timesteps, $\boldsymbol{y}_{et}$ is the one-hot encoded ground truth vector for training example $e$ at timestep $t$, and $\overline{\boldsymbol{y}}_{et}$ is the vector of predicted probabilities as outputted by the softmax layer for training example $e$ at timestep $t$. Therefore, the complete loss function for the MNMT baseline is:

$$\mathcal{L} = \alpha \mathcal{L}_{cross-entropy} + (1-\alpha)\mathcal{L}_{triplet1} \tag{10}$$

where $\alpha$ determines the contribution of the translation loss versus the visual grounding loss.

### 3.3 MNMT with Alignment Model Based on Stacked Cross-Attention

Similar to the MNMT baseline, our model learns a shared space jointly with the translation task to obtain visually grounded source word representations. In this shared space, we align source words, target words and image regions which refer to the same fashion attribute. Hence in contrast with the MNMT baseline, our space is both multimodal and multilingual and our alignment is finer, resulting in a space which captures fine-grained semantics across the visual and textual modalities. Note that the alignment at the region and word level is latent: we know which sentence corresponds with which image, but which words and image regions correspond is unknown. Therefore, we use an alignment model to learn these correspondences from frequent combinations of words and visual patterns in our training set. The alignment model is based on stacked cross-attention [12]. We will further refer to our model as the MNMT SCA model.

**Encoder.** The encoder is identical to the one of the MNMT baseline in Sect. 3.2.

**Shared Space Inference Task.** We obtain image regions by representing the image with the *res4f*-features $\boldsymbol{v}_k \in \mathbb{R}^{1024}(k = 1..196)$ of ResNet50 [8] pre-trained on ImageNet [14]. The image regions $\boldsymbol{v}_k$, source words $\boldsymbol{s}_j$ and target words $E_y y_t$ are projected to $\widehat{\boldsymbol{v}}_k$, $\widehat{\boldsymbol{s}}_j$ and $\widehat{\boldsymbol{y}}_t$ in the multimodal, multilingual space:

$$\widehat{\boldsymbol{v}}_k = W_{vk_{emb}}\boldsymbol{v}_k + \boldsymbol{b}_{vk_{emb}} \tag{11}$$

$$\widehat{\boldsymbol{s}}_j = W_{sj_{emb}}\boldsymbol{s}_j + \boldsymbol{b}_{sj_{emb}} \tag{12}$$

$$\widehat{\boldsymbol{y}}_t = W_{yt_{emb}}E_y y_t + \boldsymbol{b}_{yt_{emb}} \tag{13}$$

with $\widehat{\boldsymbol{v}}_k, \widehat{\boldsymbol{s}}_j, \widehat{\boldsymbol{y}}_t \in \mathbb{R}^d$. The projections to the multimodal, multilingual space are learned by minimizing a triplet loss which enforces that corresponding image regions, source words and target words should be closer than non-corresponding ones:

$$\mathcal{L}_{triplet2} = \frac{\ell\left(\widehat{V},\widehat{S}\right) + \ell\left(\widehat{V},\widehat{T}\right) + \ell\left(\widehat{S},\widehat{T}\right)}{3} \tag{14}$$

$$\begin{gathered}\text{with } \widehat{V} = \{\widehat{\boldsymbol{v}}_1,\ldots,\widehat{\boldsymbol{v}}_{196}\}, \widehat{S} = \{\widehat{\boldsymbol{s}}_1,\ldots,\widehat{\boldsymbol{s}}_M\}, \widehat{T} = \{\widehat{\boldsymbol{y}}_1,\ldots,\widehat{\boldsymbol{y}}_T\} \\ \ell(Q,K) = \max(0, m - SCA(Q,K) + SCA(Q,K_{hard}))\end{gathered} \tag{15}$$

$$+ \max(0, m - SCA(Q,K) + SCA(Q_{hard},K)) \tag{16}$$

where $m$ is the margin and $SCA(Q,K)$ is the similarity score of two sets of features $Q$ and $K$. Note that we use hard negative sampling here, i.e., $Q_{hard}$ and $K_{hard}$ are the hardest negatives for the corresponding feature sets $(Q,K)$ and are given by $Q_{hard} = argmax_{Q' \neq Q} SCA(Q',K)$ and $K_{hard} = argmax_{K' \neq K} SCA(Q,K')$. Similarity score $SCA(Q,K)$ of feature set $Q = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{Q_{tot}}\}, \boldsymbol{q}_i \in \mathbb{R}^d$ and feature set $K = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_{K_{tot}}\}, \boldsymbol{k}_i \in \mathbb{R}^d$ is computed with stacked cross-attention. Stacked cross-attention works in two stages of attention. In the first stage, we compute the cosine similarities $f(\boldsymbol{q}_i, \boldsymbol{k}_j)$ of all pairs of $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$. These cosine similarities are thresholded at zero and normalized to get attention scores $\boldsymbol{c}_{ij}$ for each $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$:

$$c_{ij} = \frac{\max\left(0, f\left(\boldsymbol{q}_i, \boldsymbol{k}_j\right)\right)}{\sqrt{\sum_{i=1}^{Q_{tot}} \max\left(0, f\left(\boldsymbol{q}_i, \boldsymbol{k}_j\right)\right)^2}} \tag{17}$$

Next a context vector $\boldsymbol{c}_i^{att}$ is computed for each $\boldsymbol{q}_i$ as a weighted combination of the $\boldsymbol{k}_j$:

$$\boldsymbol{c}_i^{att} = \sum_{j=1}^{K_{tot}} \gamma_{ij} \boldsymbol{k}_j, \text{with } \gamma_{ij} = \text{softmax}([\eta c_{i1}, \eta c_{i2}, \ldots, \eta c_{iK_{tot}}])_j \tag{18}$$

with $\eta$ a hyperparameter. If $\boldsymbol{q}_i$ corresponds with some $\boldsymbol{k}_j$, then $\boldsymbol{c}_i^{att}$ will be highly correlated with this $\boldsymbol{k}_j$. Otherwise, $\boldsymbol{c}_i^{att}$ will not be correlated with any of the $\boldsymbol{k}_j$. In the second stage, the similarity score of the two feature sets is calculated as the average cosine similarity $f$ between feature $\boldsymbol{q}_i$ and its context vector $\boldsymbol{c}_i^{att}$:

$$SCA(Q,K) = \frac{\sum_{i=1}^{Q_{tot}} f\left(\boldsymbol{q}_i, \boldsymbol{c}_i^{att}\right)}{Q_{tot}} \tag{19}$$

**Translation Task.** Aligning the image regions, source words and target words in the multimodal, multilingual space makes that the source word representations $\widehat{\boldsymbol{s}}_j$ become visually grounded. Therefore, we feed these $\widehat{\boldsymbol{s}}_j$ to the decoder (instead of the $\boldsymbol{s}_j$) to let the decoder benefit from the visual context. The decoder hidden state is initialized with Eq. 8 but with $\boldsymbol{s}_{att}$ computed as:

$$z_j = \max(0, \max_{\mathrm{k}}\left(f\left(\widehat{\boldsymbol{s}}_j, \widehat{\boldsymbol{v}}_k\right)\right) \tag{20}$$

$$\boldsymbol{s}_{att} = \sum_{j=1}^{M} \beta_j \widehat{\boldsymbol{s}}_j, \text{with } \beta_j = \text{softmax}([z_1, z_2, \ldots, z_M])_j \tag{21}$$

with $f$ the cosine similarity. The complete loss function for our MNMT SCA model is the same as in Eq. 10, but with the triplet loss $\mathcal{L}_{triplet2}$ of Eq. 14 instead.

# 4 Experimental Setup

## 4.1 Dataset

For this task we acquired a new, real-world e-commerce dataset from the company e5 mode, with product descriptions in English, French and Dutch and images of fashion products. The product descriptions describe the main features of a product, but do not provide an exhaustive description. Moreover, not all described product features are visible in the image, e.g., they might apply to the back of the product. The English and Dutch descriptions are sentence-aligned, i.e., they are exact parallel translations. The English and French descriptions have comparable content, i.e., they have similar content but are not translations of each other. The product descriptions are associated with one image that displays the fashion product on a clear, white background. A fashion product can either be a clothing item such as a dress, blouse, pants or underwear, or a clothing accessory like a necklace, belt, scarf or tie. The dataset consists of 3082 product images with associated descriptions in the three languages. The amount of products in this dataset is a realistic size for most e-retailers. Of the total amount of products, 2460 ($\sim$80%) are used for training, 314 ($\sim$10%) for testing and 308 ($\sim$10%) for validation. The validation set is used for hyperparameter tuning during training.

## 4.2 Experiments and Evaluation

We train the UNMT baseline, MNMT baseline and our MNMT SCA model on the e5 fashion dataset for English→Dutch and English→French. We evaluate the translation quality of the resulting models with the BLEU score. The BLEU score has a high correlation with human judgements of translation quality and is one of the most popular metrics to evaluate translation systems. It computes the number of matching $N$-grams (with $N = 1..4$) between the generated translation and the ground truth reference translation. We use beam search with a beam size of 12 for translation decoding.

## 4.3 Training Details

All hyperparameters are set based on our validation set. For models trained with both the cross entropy loss and triplet loss, a factor $\alpha$ of 0.99 and a margin $m$ of 0.1 were found to work well. The dimensions $d_x$ and $d_y$ of the source and target word representations are set to 256. The dimension $d$ of the shared spaces is set to 512. The hidden state of the decoder is 512-dimensional. The decoder initialization weight $\lambda$ is set to 0.5 and the inversed temperature of the softmax function $\eta$ to 4. We stop the training phase if there is no improvement in BLEU score on the validation set for 10 consecutive evaluation steps.

# 5 Results

Table 1 shows the BLEU scores obtained by all models on the e5 fashion dataset. These results indicate that our MNMT SCA model outperforms the MNMT baseline on both language pairs. Hence, a multimodal, multilingual space which aligns images and sentences at the level of regions and words is best for visually contextualizing the source word representations. Figure 1 compares some of the translations generated by our MNMT SCA model with those of the MNMT baseline. In the first example, the MNMT baseline incorrectly interprets *loosely* as referring to the shape of the pullover, while it refers to the knit. Both the MNMT baseline and our MNMT SCA model generate a wrong translation for *flattering*, but while the word *rounded (afgeronde)* generated by our MNMT SCA model also applies to the neckline, the word *yellow (geel)* generated by the MNMT baseline does not.

**Table 1.** Translation results for English→Dutch and English→French in terms of BLEU score on the e5 fashion test set.

| Method | English→Dutch | English→French |
|---|---|---|
| MNMT baseline [19] | 69.47 | 44.57 |
| MNMT SCA model | **74.06** | **45.62** |

**Table 2.** Percentage of test examples where the model outperforms the UNMT baseline for English→Dutch and English→French in terms of BLEU score.

| Method | English→Dutch | English→French |
|---|---|---|
| MNMT baseline [19] | 20.38 | 31.53 |
| MNMT SCA model | **34.08** | **34.71** |

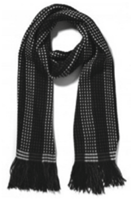
**Source:** A light pink, loosely knitted pullover. The flattering boatneck and short bat sleeves give a very soft and feminine profile. The fabric contains 20% wool.
**MNMT baseline:** Een lichtroze, losse pull. De gele boothals en korte vleermuismouwen geven de pull een heel zacht en vrouwelijk profiel. De stof bestaat uit 20% wol.
**MNMT SCA:** Een lichtroze, losjes gebreide pull. De afgeronde boothals en korte vleermuismouwen geven een heel zacht en vrouwelijk profiel. Het stofje bestaat uit 20% wol.

**Source:** Elegant printed short skirt. Ultra-soft and lightweight A-line model. Closes with a concealed zipper at the back. Fitted with a petticoat.
**MNMT baseline:** Elegant kort mouwloos kleedje. Elegant zacht en makkelijk te combineren. Sluit met een lichtbruin - rits achteraan. Voorzien van een onderrok.
**MNMT SCA:** Elegante bedrukte korte rok. Sober en licht A-lijn model. Sluit met een blinde rits aan de achterzijde. Voorzien van een onderrok.

**Fig. 1.** Comparison of translations generated by the MNMT baseline and our MNMT SCA model for English→Dutch (best viewed in color).

In the second example, the MNMT baseline generates a description of a *sleeveless dress (mouwloos kleedje)* instead of a *printed skirt*. Moreover it misidentifies the shape as being *easy to combine (makkelijk te combineren)* and the zipper as being *light brown (lichtbruin).* These mistakes are not made by our MNMT SCA model.

As also confirmed in previous works [3, 4], the MNMT models are still surpassed by the pure text-based UNMT baseline, which achieves a BLEU score of 74.38 for English→Dutch and of 48.05 for English→French. This is because the signal coming from the text in a MNMT model is stronger than the one coming from the vision side, and distilling the relevant fine-grained details from an image is a difficult task. However even if the UNMT baseline performs better overall, we can also compare the BLEU scores of the individual test examples. Table 2 reports the percentage of test examples where the associated image helps generate a better translation. These results show that in a third of the test examples, supplying an image with the source sentence results in an improved translation when using our MNMT SCA model. One of the test examples for English→Dutch for which this is the case is shown in Fig. 2.



**Source:** A navy scarf with white-blue squares. With fringes. 30 cm on 160 cm.
**UNMT baseline:** Een navy sjaal met zwart-blauwe blokjes. Met franjes. 30 cm op 160 cm.
**MNMT SCA:** Een navy sjaal met wit-blauwe blokjes. Met franjes. 30 cm op 160 cm.

**Fig. 2.** Example for which our MNMT SCA model outperforms the UNMT baseline for English→Dutch (best viewed in color).

While the BLEU score is a good metric to determine translation quality, it has some disadvantages. For instance, a translation which is significantly different from the reference translation will get a low BLEU score, even if it is still valid and acceptable to the human reader. Moreover, a translation which does not sound that smooth or contains a rather unexpected word may not get penalised as much by BLEU if it still closely resembles the reference translation. For a human though it will be clear that such a translation was generated by a machine. However, e-retailers might prefer having consumers find and buy desired products through machine-generated translations instead of not at all, or through human translations which are much more expensive to obtain.

## 6 Conclusion

In this paper, we have proposed a novel neural architecture for MNMT, which learns a multimodal, multilingual space jointly with a translation model to obtain visually grounded source word representations. By attending to the visually grounded source word representations we can jointly reason over vision and language in a way that is effective to produce the translation in the target language. We introduced this model in the context of fashion e-commerce, where the product descriptions describe fine-grained product attributes somewhere in the associated image. Moreover, we have improved state-of-the-art multimodal translation results on a real-word fashion e-commerce dataset.

As future work and to further improve the results, we would like to expand our model by integrating multiple languages and to investigate neural architectures that still better recognise fine-grained fashion attributes in images. We would also like to further explore the possibility to train on comparable data as this forms a realistic setting when dealing with product descriptions in different languages. Finally, the model proposed in this paper offers opportunities to automatically generate different types of fashion item descriptions (in one or multiple languages) that are adapted to its users, to the targeted country or culture, or to marketing strategies, which will take into account images of the fashion item.

## References

1. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. In: Proceedings of NAACL-HLT 2016, pp. 1545–1554. ACL (2016)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473 (2014)
3. Caglayan, O., Aransa, W., Wang, Y., Masana, M., Garcìa-Martìnez, M., Bougares, F., Barrault, L., van de Weijer, J.: Does multimodality help human and machine for translation and image captioning? In: Proceedings of WMT 2016, pp. 627–633. ACL (2016)
4. Caglayan, O., Aransa, W., Bardet, A., Garcìa-Martìnez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., van de Weijer, J.: LIUM-CVC submissions for WMT17 multimodal translation task. In: Proceedings of WMT 2017, Volume 2: Shared Task Papers, pp. 432–439 (2017)
5. Calixto, I., Liu, Q.: Incorporating global visual features into attention-based neural machine translation. In: Proceedings of EMNLP 2017, pp. 992–1003 (2017)
6. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555 (2014)
7. Elliott, D., Kàdàr, À.: Imagination improves multimodal translation. In: Proceedings of IJCNLP 2017, pp. 130–141 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)
9. Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., Dyer, C.: Attention-based multimodal neural machine translation. In: Proceedings of WMT 2016, Volume 2: Shared Task Papers, pp. 639–645 (2016)
10. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of CVPR 2015, pp. 3128–3137 (2015)
11. Laenen, K., Zoghbi, S., Moens, M.-F.: Web search of fashion items with multimodal querying. In: Proceedings of WSDM 2018 (2018)
12. Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross-attention for image-text matching. In: Proceedings of ECCV 2018, pp. 212–228 (2018)
13. Munigala, V., Mishra, A., Tamilselvam, S.G., Khare, S., Dasgupta, R., Sankaran, A.: Persuaide! An adaptive persuasive text generation system for fashion domain. In: Companion Proceedings of the Web Conference 2018, pp. 335–342. ACM (2018)

14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV **115**, 211–252 (2015)
15. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**, 2673–2681 (1997)
16. Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A.V., Mokry, J., Nadejde, M.: Nematus: a toolkit for neural machine translation. In: Proceedings of the Software Demonstrations of EACL 2017, pp. 65–68. ACL (2017)
17. Xu, H., Saenko, K.: Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: Proceedings of ECCV 2016, pp. 451–466 (2016)
18. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Proceedings of ECCV 2018, pp. 707–723 (2018)
19. Zhou, M., Cheng, R., Lee, Y.J., Yu, Z.: A visual attention grounding neural model for multimodal machine translation. In: Proceedings of EMNLP 2018, pp. 3643–3653. ACL (2018)