A Filtering Approach to Object Region Detection in Multimodal Machine Translation

Ali Hatami ali.hatami@insight-centre.org
Paul Buitelaar paul.buitelaar@insight-centre.org
Mihael Arcan mihael.arcan@insight-centre.org
Insight SFI Research Centre for Data Analytics,
Data Science Institute, University of Galway, Ireland

Abstract

Recent studies in Multimodal Machine Translation (MMT) have explored the use of visual information in a multimodal setting to analyze its redundancy with textual information. The aim of this work is to develop a more effective approach to incorporating relevant visual information into the translation process and improve the overall performance of MMT models. This paper proposes an object-level filtering approach in Multimodal Machine Translation, where the approach is applied to object regions extracted from an image to filter out irrelevant objects based on the image captions to be translated. Using the filtered image helps the model to consider only relevant objects and their relative locations to each other. Different matching methods, including string matching and word embeddings, are employed to identify relevant objects. Gaussian blurring is used to soften irrelevant objects from the image and to evaluate the effect of object filtering on translation quality. The performance of the filtering approaches was evaluated on the Multi30K dataset in English to German, French, and Czech translations, based on BLEU, ChrF2, and TER metrics.

1 Introduction

In recent years, neural network-based models have been widely used in translation tasks. Neural Machine Translation (NMT) represents remarkable performance in terms of fluency and precision compared with the previous generations of machine translation (Cho et al., 2014a). Recurrent Neural Network (RNN) with an attention mechanism has found broad application in NMT due to its capability to capture long-term dependencies between the most relevant parts of the source sentence (Cho et al., 2014b). The transformer model has demonstrated remarkable improvements in machine translation tasks. The cross-attention mechanism as a crucial component of the transformer-based model enhances the model's ability to capture semantic dependencies by combining self-attention, which allows source words to interact with themselves, with attention mechanisms involving target words (Vaswani et al., 2017).

Most current NMT models have shown incredible improvements in the quality of translations, but they rely solely on parallel text corpora for training. However, recent studies (Yao and Wan, 2020; Zhao et al., 2022; Wang and Xiong, 2021) in NMT have increasingly focused on using visual as well as textual content to enhance the quality of translations. Multimodal Machine Translation (MMT), a subarea of NMT, has been introduced to utilise visual information extracted from other modalities, such as images or videos, to translate an aligned sentence

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 393–405
September 4–8, 2023, Macau SAR, China.

Figure 1: The use of an image helps the translation model disambiguate the word *seal* in the sentence "*Two boys watch a seal.*" and select the correct translation from English to German.

in a source language into the target language. Similar to other multimodal tasks, MMT aims to enhance the model's ability by using visual content as an additional source of information to better understand and translate the source text. The idea behind MMT is to incorporate visual information to assist with word sense disambiguation in the input text.

Despite the fact that text-only NMT models, particularly the hidden states in the attention mechanism, consider contextual information, word sense disambiguation remains an open challenge for NMT (Tang et al., 2018). For example, as shown in Figure 1, the word "*seal*" in the English sentence "*Two boys watch a seal.*" is an ambiguous word and could have at least two different translations in German: (1) "*Zwei Jungs gucken sich einen Seehund an.*", and (2) "*Zwei Jungs gucken sich ein Siegel an.*". The Word *Seehund* in (1) refers to a fish-eating aquatic mammal, and *Siegel* in (2) is a piece of wax with an individual design stamped into it. Given the word "*seal*", the context of the source text does not provide enough information to disambiguate the words in English, and both translated texts in German are correct. However, the aligned image with the source text can provide additional information for disambiguation of the source text. Due to this, visual information can enrich text-only NMT models by leveraging additional information to disambiguate input words and provide correct translations on the target side.

Despite the importance of using visual context, visual resources such as images and videos contain a large amount of information that might not be helpful in the translation step. This additional information does often not help on improving the performance of a translation model and in some cases, it even drops the translation quality. So the recent studies on MMT focus more on finding a suitable approach to reduce the negative effects of rich visual information and enrich the translation model with the related information. To overcome the challenge mentioned above, this work focuses on identifying related visual information in the image encoder before using it in the translation model. Our approach is based on matching identified objects within the images with the captions in the text to detect which identified objects are relevant and useful in the translation process. Therefore, we apply a filtering approach that blurs irrelevant object regions on the image to reduce their negative effects on the translation model.

2 Related Work

There are various approaches proposed to integrate visual information with text-only translation models. These approaches typically utilise a visual attention mechanism in either the decoder or encoder to capture the relationships between words in a sentence and image features. The common method involves extracting visual information by employing Convolutional Neural Networks (CNN) and then integrating this information with textual features (Yao and Wan, 2020).

Regarding visual features, existing studies on MMT employ two types of visual features: global and local visual features. Global features represent the entire image as a single vector

without attention to the spatial layout of the image (Calixto and Liu, 2017). On the other hand, local features describe an image as a sequence of equally sized patches (Calixto et al., 2017). Global and local features represent different information about the image based on texture. Local features are extracted from multiple points in the image and are more robust to clutter than global features (Lisin et al., 2005). CNNs can be used to extract both global and local features from the image (Zheng et al., 2019).

In some studies, global image features are used in the encoder in addition to word sequences, to use both types of features in the decoding stage (Huang et al., 2016). Alternatively, they can be used to initialise the hidden parameters of the encoder and decoder in RNN (Calixto and Liu, 2017). In Caglayan et al. (2017), element-wise multiplication is used to initialise the hidden states of the encoder/decoder in the attention-based model. In Zhou et al. (2018), a visual attention mechanism is used to link visual and corresponding text semantically.

Despite the successful use of multimodal information in MMT, visual features do not always improve the translation quality of the integrated model, especially when textual features are highly informative (Caglayan et al., 2019). Therefore, recent studies on MMT focus more on the quality of the visual modality used as auxiliary information in the translation model (Zhao et al., 2022; Wang and Xiong, 2021), specifically in selecting relevant information and integrating this visual information with textual modality (Caglayan et al., 2016).

Several approaches have been proposed to improve the quality of visual modality in MMT. For instance, Yao and Wan (2020) proposed a multimodal transformer-based self-attention mechanism to encode relevant information in images. To capture various relationships, Yin et al. (2020) proposed a graph-based multimodal fusion encoder. Ive et al. (2019) introduced a translate-and-refine mechanism by using images in a second-stage decoder to refine the text-only NMT model in ambiguous words. Calixto et al. (2019) employed a latent variable model to extract the multimodal relationships between image and text modalities. Recent methods try to reduce the noise of visual information and select visual features related to the text. For example, Wang and Xiong (2021) used object-level visual modelling to mask irrelevant objects and specific words in the source text to analyse visual feature learning. Zhao et al. (2022) employed object detection in the image encoder to extract visual features of object regions from an image and then applied it to a doubly-attentive decoder model.

In our study, we utilised blur filtering on the initial image to conceal irrelevant objects while preserving the relevant ones. Our approach differs from previous works such as Zhao et al. (2022); Wang and Xiong (2021); Yin et al. (2020) in that our blur filter prioritises relevant objects over irrelevant ones. However, it's important to note that the irrelevant objects are only partially blurred, so the MMT model does not completely disregard them. Additionally, by applying the filter to the entire image, the model gains knowledge about the relative positions of all objects in relation to each other.

3 Methodology

In this section, we explain the main steps of our approach: i) detect object regions from images, ii) align object regions with captions and iii) blur irrelevant object regions.

3.1 Object Region Detection

For the image encoder, we use an object detection model to extract object-level features from the input image. As shown in Figure 2, the encoder first uses a bottom-up attention-based object detection model (Anderson et al., 2018) to detect n objects from the image. The bottom-up attention mechanism detects a set of image regions, with each region represented by a pooled convolutional features vector. This mechanism is based on Faster R-CNN (Ren et al., 2015) with ResNet-101 (He et al., 2016) pre-trained on Visual Genome (Krishna et al., 2017) to detect

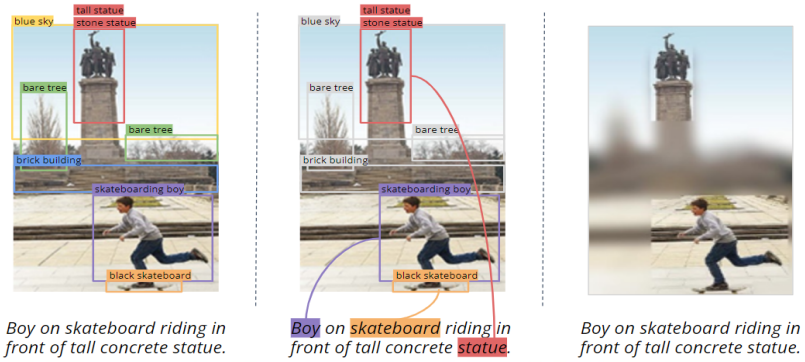

Figure 2: Our proposed image filtering approach involves three steps (left to right): (1) detecting all possible objects from the image, (2) aligning words in the text with identified object classes for irrelevant object detection, (3) applying blur filtering on irrelevant objects.

1,400 objects and 600 object attributes. The bottom-up attention mechanism first generates a fixed-length feature vector for each region proposal in the image. Then, these region proposals are classified using the Faster R-CNN model, and for each identified object, the model returns its object class, object attribute, and bounding box. For example, Figure 2 shows the objects identified from the image including *statue*, *skateboard*, and *boy* as object classes and *tall*, *stone*, and *skateboarding* as object attributes.

3.2 Object Region and Caption Alignment

After obtaining the identified objects, we explore different strategies to align the identified object classes with words in the text captions to be translated. As we discussed, the redundancy of information in the image side is one of the important challenges for MMT. As shown in Figure 2, some of these objects such as *statue*, *skateboard*, and *boy* are important for translating *"Boy on skateboard riding in front of tall concrete statue."*, while other detected objects are not mentioned in the caption to be translated. Thus, finding the relevant visual in regard to the caption plays an important role in MMT tasks. In this work, we used string matching, lemma matching and word embedding similarity approaches to find matching objects that are mentioned in the text caption.

String matching is a technique used to compare two strings and determine whether they match for a specific word or sequence of words within a larger body of text. In this work, we used string matching to align each word in the text caption with the detected object classes in the image. This is an important step in selecting relevant visual information for the translation process. To perform string matching, the words in the text caption are compared with each detected object class to determine whether they match or not.

String matching is a simple approach that compares the overlap of a word or a sequence of words in the caption with the exact string of the object class. However, this can be a limitation, as the words in the captions can be inflected, opposite to the lexicalised object classes that are always in their nominative form. Therefore, we used lemma matching, which is more flexible than string matching. Lemma matching is used for matching the nominative form of words (known as lemmas) in the text caption with the base form of identified object classes. This is particularly useful in cases where there may be variations in the form of words such as plural. For example, using string matching, the word *statues* in the caption was not matched with the nominative form *statue* provided by the object detection tool. Applying lemma matching, we could align *statue* with the object class.

	Training	Validation	Test 2016	Test 2017	Test 2018
Number of Sentences	29,000	1,014	1,000	1,000	1,071
English (words/sent)	13.0	13.1	13.0	11.4	12.9
German (words/sent)	12.4	12.7	12.1	10.8	11.5
French (words/sent)	14.1	14.2	14.0	12.6	13.8
Czech (words/sent)	10.2	10.2	10.5	-	10.2

Table 1: The summary of the Multi30k dataset includes the number of sentences and the average words per sentence for each language.

Furthermore, we leverage word embeddings to align words in the caption with the object detection classes, where each word is represented as a dense vector of real numbers, where each dimension of the vector corresponds to a feature of the word. Using word embeddings, we can find matching words between the text caption and the object classes by computing the similarity between their corresponding vectors. This approach allows us to capture semantic similarities between words, even if they are not exact matches. For instance, the words "*girl*" and "*woman*" can be semantically related using word embeddings, whereas string and lemma matching fail to identify their connection.

In this work, we use two different word embedding methods, GloVe and BERT. GloVe (Pennington et al., 2014) is a word embedding model that aims to capture the semantic and syntactic relationships between words. Unlike GloVe, BERT (Devlin et al., 2019) is a context-based model. BERT is a language model that learns the representation of the contextual relationship between the words in a sentence, known as contextual word embeddings. For example, in GloVe, the word "*bank*" would have the same vector representation in phrases like "*bank account*" and "*bank of the river*" However, in BERT, each word is represented based on the context of the other words in the sentence. To compute the similarity between each word in the caption with all object classes, we use the cosine similarity. This metric measures the cosine of the angle between two vectors and ranges between 0 and 1. A cosine similarity of 1 indicates that two vectors are identical, while a cosine similarity of 0 indicates that they are completely dissimilar.

For each word embedding model, we perform experiments using various cosine similarity thresholds. We select the optimal threshold for each method based on the translation BLEU score. Through empirical observation, we found that thresholds of 0.8 for GloVe and 0.98 for BERT yielded the best translation results. Once the matching between each word in the text caption and the identified object classes is finished, the relevant object classes can be chosen for applying blur filtering on irrelevant objects.

3.3 Irrelevant Object Region Filtering

After selecting the relevant objects for each matching technique, we apply a blur filter to the region boxes of the irrelevant objects in the original image. There are two benefits by using a blur filter for the irrelevant objects in the original image. Firstly, the blur filter helps the model to focus on the relevant objects more than the irrelevant objects. Nevertheless, as we only partially blur the irrelevant objects, the MMT model does not completely ignore them. Secondly, applying the filter to the whole image allows the model to learn the positional information of all relevant objects in the image.

For blur filtering, we use Gaussian blur (also called Gaussian smoothing), a convolution technique widely used in computer vision as a pre-processing step for noise reduction and eliminating details from the image (Ibrahim. et al., 2021). Gaussian blur is a linear low-pass filter that uses a Gaussian function to calculate the pixel value. Equation 1 shows the Gaussian blur filter with a two-dimensional Gaussian function.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-x^2+y^2}{2\sigma^2}} \quad (1)$$

Where *(x,y)* are the coordinates of the pixel, and σ is the standard deviation of the Gaussian distribution. The standard deviation (σ) is a parameter that changes the radius of the Gaussian function and controls the blur intensity. The intensity of blurring refers to the degree or level of blur applied to an image or specific areas within an image. It determines how much the details in the image are smoothed or obscured. By increasing the radius, the Gaussian function considers more neighbouring pixels, leading to an increased degree of intensity. A higher intensity of blurring results in a stronger and more noticeable blur effect, while a lower intensity produces a milder or less pronounced blur. After filtering out irrelevant objects and keeping relevant ones in the original image, we use the ResNet-101 model pre-trained on ImageNet (Deng et al., 2009), to extract visual features from the filtered image. We used the Python Imaging Library (PIL)[1] to apply blur filtering to the image. We perform the experiment with different blur intensities (10, 25, and 75) for the English to German translation task. Based on the BLUE scores for all matching strategies, we determine that a blur intensity of 75 produces the best results.

4 Experimental Setup

This section provides insights into the dataset used in this work, translation evaluation metrics and neural architecture of our model including text/image encoder and decoder.

4.1 Dataset

We used the Multi30K (Elliott et al., 2016) dataset in this work to train and evaluate our models. Multi30K is an extension of the Flickr30K Entities dataset that consists of 29,000 images with paired descriptions expressed in one English sentence and translated sentences in German, French, and Czech (Elliott et al., 2017). The training set of the dataset contains captions aligned with the images. Multi30K also provides three test sets: the 2016 and 2017 test sets, each with 1,000 images, and the 2018 test set with 1,071 images. Table 1 summarises the dataset, including the number of sentences and the average number of words per sentence for each language.

4.2 Object Detection Framework

We use the bottom-up attention (Anderson et al., 2018) mechanism to detect objects in the image encoder to extract all possible objects from an input image. This object detection model is based on the Faster R-CNN model (Ren et al., 2015) and can be used to extract class, attribute and region box for each object. This object detection model is a pre-trained model on Visual Genome (Krishna et al., 2017) to detect 1,400 objects and 600 attributes. For this work, we use the default settings[2] for Faster R-CNN model to extract 36 objects for each image (Anderson et al., 2018). Figure 2 shows an example of the output of the object detection model that extracts object region boxes for an image with the associated object classes and attributes. In this example, object detection model identifies multiple objects from the image and returns a pair of words for each object (attribute class) including: *blue sky, tall statue, stone statue, bare tree, brick building, skateboarding boy, black skateboard.*

4.3 Word Embeddings

We used word embedding methods to align words in the text caption with the detected object classes. Specifically, we utilised GloVe and BERT word embedding models to find relevant object classes for words in the English caption. For this work, we used pre-trained GloVe 50d

[1] `https://github.com/python-pillow/Pillow`

[2] `https://github.com/airsplay/py-bottom-up-attention`

English → German	BLEU ↑	ChrF2 ↑	TER ↓
Text-only NMT	32.5	57.7	53.7
Basline MMT	35.3 ± 1.5	60.9 ± 1.1	50.2 ± 1.5
String matching	**36.9 ± 1.6***	**61.4 ± 1.1***	**49.1 ± 1.7***
Lemma matching	36.3 ± 1.6*	61.2 ± 1.2	**49.1 ± 1.7***
GloVe matching	36.3 ± 1.7*	61.0 ± 1.2	49.5 ± 1.7*
BERT matching	36.2 ± 1.6*	60.9 ± 1.2	49.4 ± 1.6*
English → French	BLEU ↑	ChrF2 ↑	TER ↓
Text-only NMT	53.8	69.7	33.6
Baseline MMT	**56.8 ± 1.7**	**72.6 ± 1.2**	30.8 ± 1.5
String matching	56.6 ± 1.7	**72.6 ± 1.2**	**30.6 ± 1.4**
Lemma matching	56.0 ± 1.8	72.1 ± 1.2	31.6 ± 1.6
GloVe matching	56.7 ± 1.6	72.5 ± 1.2	30.7 ± 1.4
BERT matching	56.5 ± 1.7	72.5 ± 1.2	31.1 ± 1.4
English → Czech	BLEU ↑	ChrF2 ↑	TER ↓
Text-only NMT	26.0	48.7	58.0
Baseline MMT	29.4 ± 1.5	52.1 ± 1.2	53.7 ± 1.6
String matching	29.0 ± 1.5	51.7 ± 1.2	54.5 ± 1.6
Lemma matching	**29.6 ± 1.7**	**52.4 ± 1.2**	**53.0 ± 1.7**
GloVe matching	29.2 ± 1.7	51.8 ± 1.2	53.8 ± 1.7
BERT matching	28.7 ± 1.6	51.6 ± 1.2	54.3 ± 1.6

Table 2: BLEU, ChrF2 and TER scores for baseline and proposed models for English to German, French and Czech on the 2016 test set ($*$ represents a statistically significant result compared to baseline MMT at a significance level of $p < 0.05$).

word embedding to extract word vectors for the words in the text caption and identified object classes. Additionally, we used the pre-trained BERT-base-uncased to extract vectors for each word. This model is trained on lower-cased text, which allows it to generalise better to unseen text with different capitalisation patterns.

4.4 Neural Machine Translation

In this section, we introduce the text-only and multimodal NMT models used in this work.

4.4.1 Text-only NMT

We train a text-only transformer model as a baseline model for our experiment. This model uses only the text captions of the images. OpenNMT (Klein et al., 2018) toolkit is used to train the text-only model on English to German, French and Czech of Multi30k dataset. The architecture of the model includes a 6-layer transformer with an attention mechanism for both the encoder and decoder. We trained the model for 50K steps on the training dataset and set the parameters of the model to the default configuration of OpenNMT. We used Sentencepiece Kudo and Richardson (2018) to split words into sub-word units.

4.4.2 Multimodal NMT

We used the Doubly-Attentive Decoder RNN (Calixto et al., 2017) as the baseline model for our multimodal architecture. The Doubly-Attentive Decoder employs a single decoder RNN that integrates two separate attention mechanisms, one for the source-language words and another for the visual features. The decoder RNN with a Doubly-Attentive mechanism considers the previous hidden state of the decoder and previously generated word, along with two distinct attention mechanisms that handle the source sentence and image separately. For this study, we used the default configuration[3] of the Doubly-Attentive Decoder RNN. The visual features,

[3] https://github.com/iacercalixto/MultimodalNMT

English → German	BLEU ↑	ChrF2 ↑	TER ↓
Text-only NMT	25.0	53.2	63.7
Baseline MMT	28.2 ± 1.6	55.4 ± 1.1	59.6 ± 1.7
String matching	28.8 ± 1.6	55.6 ± 1.2	60.0 ± 1.9
Lemma matching	**29.1 ± 1.7***	55.7 ± 1.1	**59.0 ± 1.8**
GloVe matching	28.6 ± 1.6	**55.8 ± 1.2**	59.9 ± 1.8
BERT matching	28.4 ± 1.6	**55.8 ± 1.2**	60.0 ± 1.9
English → French	BLEU ↑	ChrF2 ↑	TER ↓
Text-only NMT	47.0	65.2	40.2
Baseline MMT	**48.4 ± 1.8**	**66.9 ± 1.2**	38.2 ± 1.7
String matching	47.8 ± 1.9	66.6 ± 1.3	**38.0 ± 1.6**
Lemma matching	47.4 ± 1.7	66.1 ± 1.3	39.2 ± 1.6
GloVe matching	47.6 ± 1.8	66.8 ± 1.3	38.6 ± 1.6
BERT matching	47.7 ± 1.9	66.8 ± 1.2	38.5 ± 1.6

Table 3: BLEU, ChrF2 and TER scores for baseline and proposed models for English to German and French on the 2017 test set ($*$ represents a statistically significant result compared to baseline MMT at a significance level of $p < 0.05$).

with a dimension of 2,048, were obtained by inputting images to a pre-trained ResNet-101 and extracting the activations of the res4f layer. The hidden state dimension of the visual model was set to 500 for both the 2-layer GRU encoder and the 2-layer GRU decoder. The model also set the dimension of the source word embedding to 500, batch size to 400, beam size to 5, text dropout to 0.3, and image region dropout to 0.5. After training the model for 25 epochs using stochastic gradient descent with ADADELTA (Zeiler, 2012) and a learning rate of 0.002, we selected the model of epoch 16 based on comparing the BLEU scores of the final models on the test datasets.

4.5 Evaluation Metrics

We report the translation scores using three metrics: BLEU (Papineni et al., 2002), ChrF2 (Popović, 2015), and TER (Snover et al., 2006). BLEU score is based on the precision of n-grams (contiguous sequences of words) in the candidate translation compared to the reference translations. ChrF2 measures the similarity between the character n-grams in the reference translation and the candidate translation produced by the machine translation system. It is particularly useful for evaluating the quality of machine translations for languages with complex writing systems, where word-based metrics like BLEU may not be as effective. TER measures the number of edits (insertions, deletions, and substitutions) required to transform a machine translation output into a reference translation produced by a human translator.

5 Results

In this section, we present the results of our experiments, where we trained our models on the Multi30k dataset and evaluated the translation quality using the BLEU, ChrF2, and TER metrics. We compare the translation quality of our proposed models, which utilise different matching approaches, i.e., string, lemma, GloVe, and BERT, with MMT baseline models across three different test sets. The text-only NMT model was trained solely on text captions without images. The MMT baseline model was trained on both text captions and original images without applying any filtering of irrelevant objects. We report the results for English-German, English-French, and English-Czech translation pairs. Comparing the text-only NMT and the MMT models, the latter statistically significant ($p < 0.05$) outperform the text-only models.

Table 2 presents the translation results of the 2016 test set from English into German,

English → German	BLEU ↑	ChrF2 ↑	TER ↓
Text-only NMT	24.0	50.8	66.0
Baseline MMT	26.4 ± 1.4	52.9 ± 1.0	63.8 ± 1.6
String matching	26.6 ± 1.4	53.3 ± 1.0	63.8 ± 1.6
Lemma matching	26.9 ± 1.3	53.3 ± 1.0	63.1 ± 1.6
GloVe matching	**27.2 ± 1.3***	**53.8 ± 1.0***	**63.0 ± 1.7**
BERT matching	27.0 ± 1.4	53.4 ± 1.0*	63.3 ± 1.6
English → French	BLEU ↑	ChrF2 ↑	TER ↓
Text-only NMT	31.5	55.3	43.0
Baseline MMT	**34.1 ± 1.4**	**57.6 ± 1.0**	51.0 ± 1.7
String matching	33.7 ± 1.4	57.3 ± 1.0	**50.7 ± 1.4**
Lemma matching	32.8 ± 1.3	56.9 ± 1.0	51.3 ± 1.3
GloVe matching	33.8 ± 1.4	**57.6 ± 1.0**	51.1 ± 1.3
BERT matching	33.7 ± 1.4	57.5 ± 1.0	51.2 ± 1.7
English → Czech	BLEU ↑	ChrF2 ↑	TER ↓
Text-only NMT	20.0	44.4	67.1
Baseline MMT	23.4 ± 1.4	46.7 ± 1.0	64.2 ± 1.6
String matching	23.5 ± 1.3	46.7 ± 1.0	63.8 ± 1.6
Lemma matching	**23.7 ± 1.3**	46.8 ± 1.1	**63.7 ± 1.6**
GloVe matching	23.6 ± 1.3	46.3 ± 1.1	64.1 ± 1.5
BERT matching	23.4 ± 1.3	**46.9 ± 1.0**	64.4 ± 2.0

Table 4: BLEU, ChrF2 and TER scores for baseline and proposed models for English to German, French and Czech on the 2018 test set ($*$ represents a statistically significant result compared to baseline MMT at a significance level of $p < 0.05$).

French, and Czech. String matching resulted in a one-point improvement in the BLEU score compared to the baseline MMT, as verified by the ChrF2 and TER metrics which was statistically significant at a significance level of $p < 0.05$. However, no significant improvements were observed in the proposed approaches for English to French translation. Lemma matching showed a slight improvement in English to Czech translation for three metrics. In Table 3, we can see the translation results for the 2017 test set from English to German and French. The use of lemma matching led to an improvement of 0.9 points in terms of the BLEU score compared to the baseline MMT. However, it was unexpected to find that for the English to French translation direction, the baseline MMT model outperformed the proposed models with blurred irrelevant objects. Table 4 presents the results of translating the 2018 test set from English to German, French and Czech. GloVe matching showed a 0.8-point improvement in terms of the BLEU score compared to the MMT baseline. This improvement is supported by ChrF2 and TER metrics. Other matching approaches demonstrated slight improvements. However, for English to French and Czech translation direction, we only observe minor improvements using the proposed matching approaches.

Figure 3 shows a few examples for the English to German translation direction, where our filtered approach improved over the baseline MMT approach. In the first example, our approach can guide the translation system to translate *riding* into *reitet*, with the meaning of *riding a horse*. The baseline MMT model translated *riding* into *fährt*, with the meaning of *driving a car*. In the second example, the baseline MMT system ignores translating the word *barefoot*, while the filtered MMT model provides the right translation, i.e. *barfüßiges*. Within the last example, the baseline MMT model translates the word *plastic* into *Gewändern*, in the meaning as *garment* or *rob*. The filtered MMT model, on the other hand, provides the right translation as a compound word, i.e., *Plastickstühlen* (en. *plastic chairs*).

Source (En) A cowboy **riding** on the back of a bronco in a competition.
Reference (De) Ein Cowboy **reitet** ein Wildpferd in einem Wettbewerb.
Baseline_MMT (De) Ein Cowboy **fährt** bei einem Wettkampf auf einem Pferd.
Filtered_MMT (De) Ein Cowboy **reitet** auf dem Rücken eines Wettkampfs in einem Rennen.

Source (En) A young **barefoot** girl in a pink dress is jumping outside.
Reference (De) Ein **barfüßiges** junges Mädchen in einem rosa Kleid springt im Freien.
Baseline_MMT (De) Ein kleines Mädchen in einem rosa Kleid springt draußen.
Filtered_MMT (De) Ein junges **barfüßiges** Mädchen in einem rosa Kleid springt im Freien.

Source (En) Two men in white **plastic chairs** sitting in a doorway.
Reference (De) Zwei Männer auf weißen **Plastikstühlen** sitzen in einem Eingang.
Baseline_MMT (De) Zwei Männer in weißen **Gewändern** sitzen in einer Türöffnung.
Filtered_MMT (De) Zwei Männer in weißen **Plastikstühlen** sitzen auf einem Eingang.

Figure 3: Examples for baseline and Filtered-based (string matching) MMT models to translate from English to German. Red and blue words indicate incorrect and correct translations, respectively.

6 Conclusion

Recent studies in Neural Machine Translation have focused on utilising visual information to enhance the quality of translation tasks. However, the success of Multimodal Machine Translation systems is highly dependent on the quality of the visual content used alongside textual datasets. Visual resources like images and videos contain a large amount of visual information, and some of it is irrelevant to the caption translation task. Hence, one of the major challenges in Multimodal Machine Translation is to separate the relevant information from the irrelevant one.

In this study, to improve the translation of the image captions, we propose to use object detection in the image encoder to prioritise relevant objects within the image. For each detected object, we extract its class, attribute, and regional box. Then, we utilise string, lemma matching, and pre-trained word embeddings, such as GloVe and BERT, to align the detected object classes in images with the words in text captions. Our experiments show that blurring irrelevant objects of images statistically significantly improves the performance of the baseline model in English to German translation. However, we observe minor improvements in translations from English to Czech, where the translations from English to French do not show any improvements. For our future work, we plan to leverage visual scene graphs in Multimodal Machine Translation. A visual scene graph is a data structure that represents visual scenes as a graph, where nodes correspond to objects and edges correspond to their relationships. It encodes the objects in the scene and the relationships among them, such as the attributes and locations of the objects and the spatial relationships between them. This representation allows for a rich and structured visual understanding of images.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We would like to thank the anonymous reviewers for their insights on this work.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, Los Alamitos, CA, USA. IEEE Computer Society.

Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., and van de Weijer, J. (2017). LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., and van de Weijer, J. (2016). Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 627–633, Berlin, Germany. Association for Computational Linguistics.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Calixto, I. and Liu, Q. (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 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Calixto, I., Liu, Q., and Campbell, N. (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, Vancouver, Canada. Association for Computational Linguistics.

Calixto, I., Rios, M., and Aziz, W. (2019). Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li, F.-F. (2009). Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Los Alamitos, CA, USA. IEEE Computer Society.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Ibrahim., N. M., ElFarag., A. A., and Kadry., R. (2021). Gaussian blur through parallel computing. In *Proceedings of the International Conference on Image Processing and Vision Engineering - IMPROVE*, pages 175–179. INSTICC, SciTePress.

Ive, J., Madhyastha, P., and Specia, L. (2019). Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. (2018). OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lisin, D., Mattar, M., Blaschko, M., Benfield, M., and Learned-Miller, E. (2005). Combining local and global image features for object class recognition. In *CVPR*, pages 47–47. Max-Planck-Gesellschaft.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Tang, G., Sennrich, R., and Nivre, J. (2018). An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, D. and Xiong, D. (2021). Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2720–2728. AAAI Press.

Yao, S. and Wan, X. (2020). Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Yin, Y., Meng, F., Su, J., Zhou, C., Yang, Z., Zhou, J., and Luo, J. (2020). A novel graph-based multi-modal fusion encoder for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *Computing Research Repository (CoRR)*, abs/1212.5701.

Zhao, Y., Komachi, M., Kajiwara, T., and Chu, C. (2022). Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.

Zheng, Y., Huang, J., Chen, T., Ou, Y., and Zhou, W. (2019). CNN classification based on global and local features. In Kehtarnavaz, N. and Carlsohn, M. F., editors, *Real-Time Image Processing and Deep Learning 2019*, volume 10996, page 109960G. International Society for Optics and Photonics, SPIE.

Zhou, M., Cheng, R., Lee, Y. J., and Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium. Association for Computational Linguistics.