NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES

Masters Thesis

Image Informed Neural Machine Translation with Transformers

Konstantina G. Nikolaidou

ATHENS

June 2020

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Διπλωματική Εργασία

Χρήση Εικόνας για Ενίσχυση Μετάφρασης με Νευρωνικά Δίκτυα

Κωνσταντίνα Γ. Νικολαίδου

ΑΘΗΝΑ

Ιούνιος 2020

Masters Thesis

Image Informed Neural Machine Translation with Transformers

Konstantina G. Nikolaidou
A.M.: DS1180013

SUPERVISOR: Vassilis Katsouros, Research Director, ATHENA Research and Innovation Center

EXAMINATION COMMITTEE:

Vassilis Katsouros, Research Director, ATHENA Research and Innovation Center

Ioannis Emiris, Professor, National and Kapodistrian University of Athens

Marcus Liwicki, Chair Professor, Luleå University of Technology

June 2020

Διπλωματική Εργασία

Χρήση Εικόνας για Ενίσχυση Μετάφρασης με Νευρωνικά Δίκτυα

Κωνσταντίνα Γ. Νικολαίδου
A.M.: DS1180013

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Βασίλης Κατσούρος, Διευθυντής Ερευνών, Ερευνητικό Κέντρο ”Αθηνά”

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Βασίλης Κατσούρος, Διευθυντής Ερευνών, Ερευνητικό Κέντρο ”Αθηνά”

Ιωάννης Εμίρης, Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Marcus Liwicki, Chair Professor, Luleå University of Technology

Ιούνιος 2020

ABSTRACT

Neural Machine Translation is one of the most important tasks in Natural Language Processing. Complex architectures have achieved incredible performances using solely text as input data. The vast amount of data and their accessibility has increased the need in exploiting them in order to achieve better and more human-like results. Multimodal Machine Translation uses additional modalities like images and speech in order to ground and enhance the task of translation assuming that they contain alternative representations of the input data.

We focus our work in investigating several variations of the same system to integrate visual features in a neural machine translation system. To this end, we use the state-of-the-art Transformer architecture. We are performing the task for three different target languages: French, German and Czech, using English as our source language. We use the Multi30K dataset, a large-scale multilingual multimodal dataset publicly available for the task of multimodal machine translation and cross-lingual captioning. We evaluate the results on three different test sets and compute and compare the systems through the BLEU, METEOR and TER scores. We further investigate whether the additional modalities are used by masking color words in the source sets.

The examined systems seem to perform similarly with small differences in their performance, with the multimodal ones giving more natural translations in many cases. Investigation on the predicted texts revealed several inconsistent parts and biases in the dataset. Results of the mask predictions show that it is very likely that systems indeed use the additional modalities when text is not sufficient, while the text model outputs predictions based on training biases.

SUBJECT AREA: Multimodal Machine Translation

KEYWORDS: Multimodality, Neural Networks, Neural Machine Translation

ΠΕΡΙΛΗΨΗ

Η μηχανική μετάφραση μέσω χρήσης νευρωνικών δικτύων αποτελεί ένα από τα πιο σημαντικά προβλήματα της Επεργασίας Φυσικής Γλώσσας. Σύνθετες αρχιτεκτονικές έχουν επιτύχει σημαντικές επιδόσεις χρησιμοποιώντας αποκλειστικά κείμενο ως δεδομένα. Η αφθονία δεδομένων και η εύκολη προσβασιμότητά τους έχουν αυξήσει την ανάγκη εκμετάλλευσής τους προκειμένου να επιτευχθούν καλύτερα και πιο φυσικά αποτελέσματα. Η Πολυτροπική Μηχανική Μετάφραση χρησιμοποιεί επιπρόσθετα δεδομένα όπως εικόνες και ομιλία, για να υποστηρίξει και να ενισχύσει την μετάφρασης, υποθέτοντας ότι περιέχουν εναλλακτικές αναπαραστάσεις των δεδομένων εισόδου.

Εστιάζουμε τη συγκεκριμένη εργασία στη διερεύνηση παραλλαγών του ίδιου συστήματος για την ενσωμάτωση οπτικών χαρακτηριστικών σε ένα νευρωνικό δίκτυο για μετάφραση. Για το σκοπό αυτό, χρησιμοποιούμε την αρχιτεκτονική Transformer η οποία αποτελεί το state-of-the-art. Για την μετάφραση χρησιμοποιούμε τρεις διαφορετικές γλώσσες-στόχους: Γαλλικά, Γερμανικά και Τσέχικα, χρησιμοποιώντας τα Αγγλικά ως την γλώσσα προέλευσης. Χρησιμοποιούμε το σύνολο δεδομένων Multi30K, ένα πολυγλωσσικό σύνολο μεγάλης κλίμακας που διατίθεται στο κοινό για την εφαρμογή σε προβήματα μηχανικής μάθησης. Αξιολογούμε τα αποτελέσματα σε τρία διαφορετικά σύνολα δεδομένων και υπολογίζουμε και συγκρίνουμε τα συστήματα μέσω των μετρικών BLEU, METEOR και TER. Εξετάζουμε περαιτέρω εάν χρησιμοποιούνται τα χαρακτηριστικά των εικόνων αντικαθιστώντας λέξεις που αποτελούν χρώματα με μια μάσκα.

Τα εξεταζόμενα συστήματα φαίνεται να συμπεριφέρονται παρόμοια με μικρές διαφορές στην απόδοσή τους, με τα συστήματα που χρησιμοποιούν τις εικόνες να δίνουν πιο φυσικές μεταφράσεις σε πολλές περιπτώσεις. Η έρευνα των μεταφρασμένων αποτελεσμάτων αποκάλυψε αρκετά ασυνεπή μέρη και προκαταλήψεις στα δεδομένων. Τα αποτελέσματα της εφαρμογής μάσκας δείχνουν ότι είναι πολύ πιθανό τα συστήματα να χρησιμοποιούν πράγματι τα πρόσθετα στοιχεία που προέρχονται από εικόνες ειδικά όταν το κείμενο δεν είναι αρκετό, ενώ το μοντέλο που βασίζεται εξ ολοκλήρου σε κείμενο κάνει προβλέψεις με βάση τις προκαταλήψεις κατά τη διάρκεια της εκπαίδευσης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Πολυτροπική Μηχανική Μετάφραση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Πολυτροπικότητα, Νευρωνικά Δίκτυα, Μηχανική Μετάφραση

ACKNOWLEDGEMENTS

This work was conducted in the Machine Learning group of Embedded Intelligent Systems (EISLAB) division in Luleå University of Technology.

I would like to thank my internal supervisor Vasilis Katsouros for his valuable help and organization that finalized this work, as well as my external supervisor Marcus Liwicki, for his guidance and support provided during this work.

Additionally, I would like to thank the people of EISLAB for the knowledge I have gained throughout this whole time and the pleasant work environment they created. Being welcome and included as a new member of the group meant a lot to me. Finally, I would like to thank my family, friends and boyfriend for the love and support they have provided me throughout this project.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Introduction

Over the last years, Computer Vision (CV) and Natural Language Processing (NLP) have shown great progress in many of their respective goal tasks like image and video processing, generation and understanding for the former and text generation, understanding, translation and more for the latter. Developing machine learning algorithms to comprehend human language has been a challenge for many years. Language has a lot of ambiguities and although humans can easily master it naturally, it remains one of the most complex tasks for a machine to be able to understand it. Neural Machine translation is one of these tasks. With the rise of Deep Learning the field has witnessed innovative solutions by using a vast amount of data and complex architectures. Neural networks though, have long been seen as a black box, where the solutions perform well according to some performance metric, however they are not explainable. A lot of focus has been given on image feature visualization in order to understand and interpret Convolutional Neural Networks (CNNs). Techniques like attention mechanism seem promising in NLP to give some kind of interpretability to the systems. The fields are in the quest of more explainable solutions and although both use similar machine learning techniques and architectures, not enough interaction had been made between them in order to gain valuable information from their combination. Like in human learning, visual information and context can help in learning a language. Humans have the ability to process various forms of data simultaneously in order to improve their understanding and perception. Given an image humans can easily give a language description of it. The tremendous growth of visual and textual data available on web has generated the need of exploiting and combining them for varied reasons and it is mandatory to explore the possibilities, how the integration of both fields can be achieved.Recently, there has been a rise in works that take advantage of both textual and visual information to perform tasks that need both features. Popular tasks that involve both NLP and CV are automatic image description and video captioning [3], visual question answering [4] and multimodal machine translation. Multimodal Machine Translation aims in drawing information from more than one modality to improve the translation task.

In this work, we attempt to exploit information from images and examine whether their use can improve the performance of a natural language processing task such as neural machine translation. The chapters of this work are organized as followed:

- **Chapter 1** makes an introduction and definition of the problem while tackling the delimitations of the work and ethical issues.
- **Chapter 2** focuses on the theory and background of Neural Machine Translation and how the field passes from Recurrent Neural Networks to the Attention mechanism and finally the Transformer.
- **Chapter 3** gives some information about the used dataset, the architecture adaptation to implement the multimodal task and the hyperparameters of the system.
- **Chapter 4** presents the performance of the different systems using evaluation metrics and several translation examples for human evaluation.
- **Chapter 5** outlines the findings and future research directions.

1.2 Related Work

In 2016, a shared task on Multimodal Machine Translation (MMT) and Crosslingual Image Captioning (CIC) was introduced along with the WMT shared tasks [5] to tackle the problem of generating descriptions of images for languages other than English [6]. The main goal of the task was to push existing work on the combination of computer vision and language processing and the multimodal language processing towards multilingual multimodal language processing. Furthermore, the task aimed to investigate the effectiveness of information from images and multiple source language sentences in machine translation. WMT held these multimodal tasks for 3 consecutive years while from 2018 it became a competition organized by the University of Sheffield [7]. Throughout these years, various methodologies have been proposed for the shared task where they compare text-only and multimodal varieties of different MT frameworks.

The majority of submissions in 2016 were based on the RNN architecture with attention of [1]. In [8], a double-attentive model is proposed that attends separately to the source and the image features. A LSTM with multiple sequential global and local visual and textual features as states for attention and parallel LSTM threads is examined in [9]. In other works that use SMT systems instead of neural networks, we see the use of captions of similar images for cross-lingual re-ranking of the translation outputs [10]. In the shared task for WMT17 the best performance is given by a multimodal attentive NMT with separate attention over source text and image features [11]. In the last year of the shared task, the Transformer architecture was used with additional image features extracted from an object detection and localization network in [12] while in [13] with the image representations as additional input, where the visual features were predicted and used as an auxiliary objective. Although these works achieved high performances, authors claim that visual features don't have a big effect on the quality of the translation and that the task is mostly dependent on the language model. The results are also depended on the expansion with synthetic data that most of works used. Other related works use ensemble methods of standard attentive neural machine translation models [14], while others don't exploit visual information [15]. The baseline consisted of a text-only attentive nmt system created with NMTPY [16], with a conditional GRU as a decoder.

1.3 Problem Definition

Multimodal machine translation (MMT) aims in assisting a translation system by adding meaningful representations through image or audio features besides text. Related work on multimodal machine translation claims that the visual modality is either unnecessary or only marginally beneficial. The task is mostly relied on the use of Recurrent Neural Networks with Attention mechanism. The success of the Transformer architecture has triggered our interest in investigating the network architecture. A limitation of the current methods is that they are mostly based on synthetic data to achieve better results. This draws the attention on the size of the data set. The objective of this work is to investigate the use of the Transformer architecture for Neural Machine Translation, how it can be adapted on the task of Multimodal Machine Translation and test whether the use of visual features can improve the task. In particular the aims are to:

- Implement a Transformer architecture and manage its hyperparameters in order to achieve a good result for a text-only Neural Machine Translation.
- Adapt the architecture in different ways to exploit visual modalities.
- Compare the performance of monomodal and multimodal systems by using relevant evaluation metrics for the task.
- Visualize the quality of the results and the attention scores at each layer of the Transformer.

1.4 Delimitations

Several discussions were made considering the used architectures for both language and image models. For the translation model a recurrent neural network architecture with attention was considered to be used as a baseline, while for the visual data a CNN built from scratch to extract the image features. Due to time limitations only the Transformer architecture was chosen to be used as our baseline and the pre-extracted image features provided by the data set. Another idea was to create a larger data set to be used for this task, with audio features as well, but we are leaving that for future work.

1.5 Ethical Considerations and Gender Aspects

A lot of discussion has been made around the lack of interpretability in neural networks and their ethical implications, often focused on privacy, bias and discrimination.

Neural machine translation and language systems have managed to generate human like text that has raised a lot safety concerns. NMT systems are difficult to control, since the programmer cannot easily give specific guidelines to the machine for the translation and doesn't know that an unwanted text may occur in the output translation until it actually occurs.

Natural language training data inevitably reflect biases present in our society and there is a lot of concern about widespread used systems that blindly use machine learning and amplify gender and racial bias [17]. Gender biases can be easily obtained when training data do not contain equal amount of sentences that refer to women and men or contain sentences with gender role analogies, e.g. *woman-homemaker, man-computer programmer* [18]. In that case, the translation quality is decreased and the model has the risk to output text that propagates gender stereotypes, especially when there is no information about the gender in the source language. In this work, a related example appears during the result investigation in 4.5, where it seems that most girls appearing in sentences are wearing outfits of color pink. We suggest the removing of these types of biases as future work.

2. NEURAL MACHINE TRANSLATION

Neural Machine Translation applies an artificial neural network in order to perform the task of translating a source language sentence to a target language sentence. The most common model used in NMT is the Sequence-to-Sequence that involves two recurrent neural networks (RNN), which is basically an encoder-decoder architecture. Recurrent neural networks manage to process input sequences of any length while not increasing the size of the model since the same weights are applied to every timestep of the input. In practice, it is hard to access information from many steps back due to vanishing and exploding gradients problems.

2.1 NMT with Sequence-to-Sequence

Sequence-to-sequence models (Seq2Seq) introduced in [19] and [20] consist of an Encoder RNN and a Decoder RNN. The Encoder RNN takes as input a sequence $x = (x_1, x_2, ..., x_n)$ of any arbitrary length and creates a representation of this sequence. The Encoder has as many hidden states as the input symbols in the sequence and at every time step t each hidden state h_t can have an optional output y and is computed based on the input of that step x_t and the previous hidden state h_{t-1}.

$$h_t = f(x_t, h_{t-1}), \qquad h_t \in \mathbb{R}_n \tag{2.1}$$

In the case of NMT, the Encoder takes as input the words of a sentence in the source language sequentially and updates the hidden states at every time step according to 2.1. An end of sequence token is fed to the encoder at the end of each sentence and when this token is met the representation of the sentence is created. This representation is a context vector c that is basically a summary of the whole input sentence.

The Decoder RNN conditions on this output fixed-length representation produced by the Encoder to generate a new variable-length sequence $y = (y_1, y_2, ..., y_T)$. During training, the Decoder takes as input a start of sentence token and the words of the sentence in the target language sequentially, while using the produced context vector c as its initial hidden state. e.

In every step the model produces a probability distribution of the next word $\hat{y_i}$ to compute the loss. The total loss will be the average of all the losses of every step. Back-propagation happens end-to-end, one end being the loss and the other being the beginning of the encoder, so the gradients flow throughout the whole system w.r.t this total loss. During inference, the Decoder takes the previous generated word as input in each step. Hence, the hidden state of the decoder at time t is the following:

$$h_t = f(y_{t-1}, h_{t-1}, c) \tag{2.2}$$

The whole network architecture is jointly tuned to maximize the performance of the translation.

The problem with the Seq2Seq architecture is that the created encoding of the source sentence by the Encoder RNN is representing the whole sentence in a single vector. This means that all the information about the source sentence is forced to be captured in this

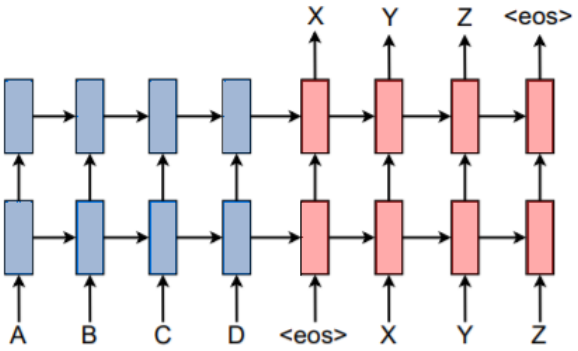

Figure 1: Neural Machine Translation with Seq2Seq architecture [1].

fixed-length vector, which is too much pressure for the vector to be a good representation, especially for longer sentences [21]. This problem introduced a technique called Attention mechanism which nowadays has taken the place of the conventional encoder-decoder architectures as the baseline in neural machine translation and many other tasks. We describe the Attention mechanism in the next sub-chapter 2.2.

2.2 Attention Mechanism

Attention was introduced in [1] in order to deal with the bottleneck in improving the performance of the basic encoder–decoder architecture for longer input sentences. The authors of the paper suggest an extension of the Seq2Seq architecture where the decoder can focus in relevant parts of the encoder when performing the prediction as shown in Figure 2.

The basic idea of the attention mechanism is that on each step of the decoder there is a direct connection to the encoder by computing an attention score between the current decoder hidden state and every encoder hidden state. In a more formal way, on a time step t of the decoder we have a hidden state $s_t \in \mathbb{R}^h$ and N encoder hidden states $h_1, ..., h_N \in \mathbb{R}^h$. The attention scores are computed by taking the dot-product of the decoder hidden state with each one of the encoder hidden states. The result gives a vector e^t (2.3) of the same length as the input sentence N, that gives one score per source word.

$$e^t = [{s_t}^T h_1, ..., {s_t}^T h_N] \in \mathbb{R}^N \tag{2.3}$$

After computing the attention scores, a softmax function is applied on the score vector to get an attention distribution α^t (2.4), in order to take a weighted sum of the encoder hidden states that gives the final attention output c_t (2.6). The attention output is finally concatenated with the decoder hidden states to proceed with the whole prediction as in the conventional model (2.7).

$$\alpha^t = \sigma(e^t) \in \mathbb{R}^N \tag{2.4}$$

where

$$\sigma(x)_i = \frac{exp(x_i)}{\sum_{k=1}^{N} exp(x_k)}, \quad x \in \mathbb{R}^N \tag{2.5}$$

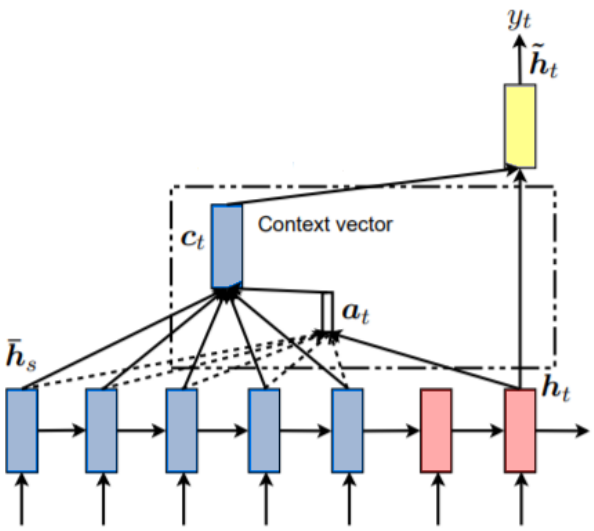

Figure 2: Neural Machine Translation with attention in Seq2Seq architecture [1].

$$c_t = \sum_{i=1}^{N} {\alpha_i}^t h_i \in \mathbb{R}^h \tag{2.6}$$

$$[c_t; s_t] \in \mathbb{R}^{2h} \tag{2.7}$$

A more general definition of attention is the following: Given a set of vector values V and a vector query Q, attention is a technique to compute a weighted sum of the values dependent on the query. In the case of the Seq2Seq model, each decoder hidden state is the query and that query attends on all the encoder hidden states which are the values. Hence, attention can be seen as a way to obtain a fixed-size representation of an arbitrary set of representations V by using the query Q, where the computed weighted sum is a selective summary of the information in the values. There are a few variants of attention like basic dot-product attention, additive attention and multiplicative attention.

There are several reasons that have placed encoder-decoder architectures with attention as the current baseline instead of the conventional Seq2Seq architecture described in 2.1. As mentioned previously, attention addressed successfully the bottleneck problem of having a single vector as a representation of an entire source sentence, by allowing the decoder to focus on specific parts of this sentence when performing the translation. This way, extra context and interpretability are provided. As the size of the input sentence grows, simple encoder-decoder architectures miss information by only using the final representation. Attention-based models have the ability to efficiently translate long sentences and experiments indeed confirm the intuition behind it [22]. Furthermore, the direct connections between decoder and encoder provided by the attention mechanism over many time steps can help with the vanishing gradient problem, that appears in neural networks and especially in RNNs because of the Backpropagation Through Time (BPTT) [23].

2.3 Transformer architecture

Inspired by the attention mechanism the Transformer architecture was proposed to replace recurrent layers and use attention in order to create meaningful representations of

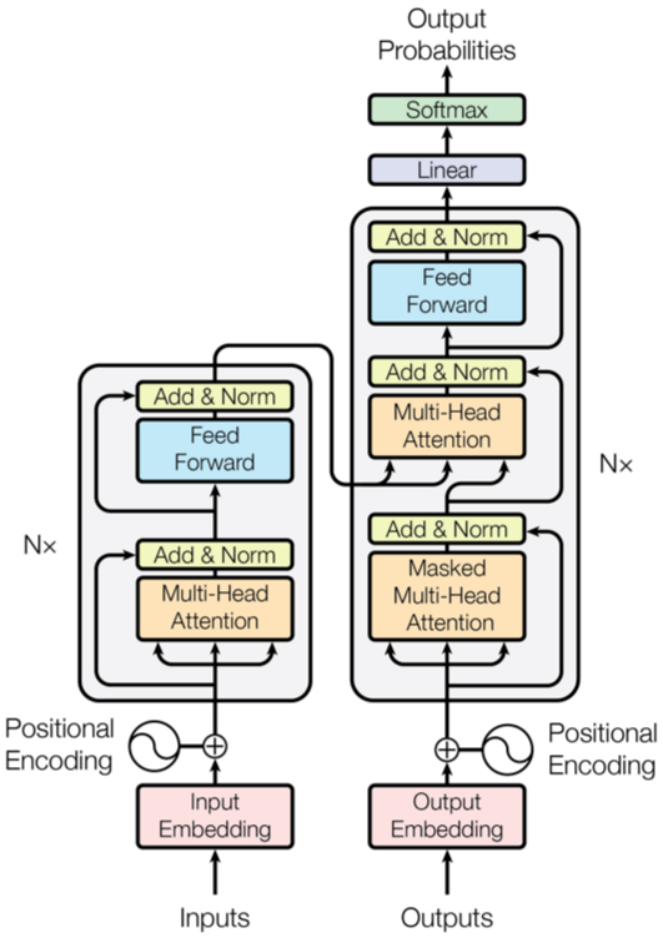

Figure 3: The original Transformer architecture presented in "Attention is All You Need" [2].

the inputs [2]. The Transformer model has achieved state of the art results for WMT 2014 English-to-German and English-to-French translation tasks while opening the path to architectures that have achieved incredible performances for various NLP tasks [24], [25], [26].

The Transformer architecture consists of an encoder-decoder structure as most of the neural sequence transduction models. The encoder creates a sequence of continuous representations $z = (z_1, ..., z_n)$ from the input sequence of symbol representations $(x1, ..., xn)$. This new representation of the input is given to the decoder that generates an output sequence of symbols $(y1, ..., ym)$ one element at a time. After the generation of one output element the model takes this generated element as additional input for the next generation. The overall Transformer architecture consists of N identical encoder and decoder layers. Each encoder layer passes the input to a multi-head self-attention mechanism layer and then into a feed-forward neural network that sends out the output to the next encoder. Around each of these two sub-layers there is a residual connection [27], followed by a layer normalization [28]. The decoder is similar to the encoder, with an additional multi-headed attention sub-layer that attends to the encoder, between the self-attention layer and the feed-forward network. Again, we have residual connections around each sub-layer and a use of masking in the initial multi-head self-attention layer (See Fig: 3).

2.3.1 Transformer Sublayers and Components

In this subsection we explain each component of the Transformer separately to give a better understanding of the architecture.

2.3.1.1 Positional-Encoding

In recurrent neural networks data are given sequentially to the model, thus the network recognizes the position and the order that each word occurs. Since there is no recurrence in the Transformer architecture there should be a way in order to determine the position of each word. To this end, positional encodings are added to the embeddings of the encoder and decoder inputs which are basically vectors that contain information about the position and give the network a sense of the order of the words. The equation used for the positional encoding is the following:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{2.8}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{2.9}$$

2.3.1.2 Attention in the Transformer

Attention is used in various ways in the Transformer architecture. Both encoder and decoder use **Multi-head Self-Attention** to create different representations of their inputs. The attention function used in the model is called **Scaled Dot-Product Attention** (see eq. 2.10) that is a dot-product attention as described in 2.2 except that it uses a scaling factor to avoid small gradients in the softmax function. The dot-product between the queries and keys of the same dimension d_k is computed and then is divided by $\sqrt{d_k}$. Then a softmax function σ is applied and the result is multiplied with the values of dimension d_u. The computation happens between matrices Q, K and V that contain the queries, keys and values respectively.

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.10}$$

Both encoder and decoder apply this scaled dot-product attention between h different linear projections of their queries, keys and values in h different layers or heads in parallel. Then, the outputs from these heads are concatenated and then again linearly projected. This is the multi-head self-attention mechanism. The self-attention means that we get a different representation subspace of the same sentence that attends on itself in the different heads. So, every position of the encoder or decoder can attend to all positions of the previous head. For the decoder a masking is used before the application of the softmax function to prevent leftward information flow in the decoder and preserve the auto-regressive property.

In the decoder, except from the multi-head self-attention layer where the decoder input attends on itself we have another multi-head attention layer which is the encoder-decoder attention, as seen in Seq2Seq models, that helps the decoder focus on the source. The queries come from the previous decoder layer, while the keys and values come from the

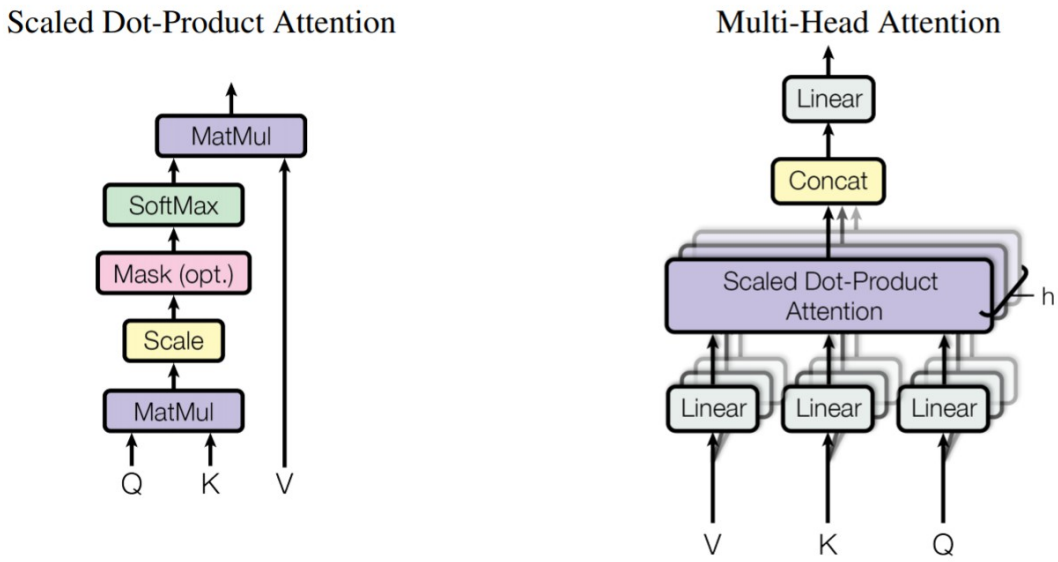

Figure 4: On the right we can see the different attention heads containing the scaled dot-product attention shown on the left. The image is retrieved from the original paper ”Attention is all you need” [2].

encoder output. Thus, every position of the decoder attends on all positions of the input. The scaled dot-product attention and multi-head attention are shown in Figure 4.

2.3.1.3 Position-wise Feed-Forward Network

A fully connected feed-forward network is used in every encoder and decoder layer after the multi-head attention. This feed-forward network is applied to every position separately and identically. The original paper states that this can be seen as two convolutions with kernel size 1 and in this work we use this approach.

3. IMPLEMENTATION

3.1 Multi-30k Dataset

We use the Multi-30k dataset [29] introduced for the WMT16 Shared Task on Multimodal Translation and Image Description [5]. The dataset consists of 29000 training, 1014 validation and 1000 test sentences that describe an image for the task of English to French, German and Czech translation. The text is pre-processed to lowercase, normalise punctuation and tokenise the sentences. For the multimodal task the dataset provides images and pre-extracted image features from a 50 layer residual neural network [27].

Textual Data: The data were created by extending the Flickr30K Entities dataset [30] in the following way: for each image, one of the English descriptions was selected and manually translated into German, French, and Czech by human translators. For English-German, translations were produced by professional translators who were given the source segment only or the source segment and image. For English-French, translations were produced via crowd-sourcing where translators had access to source segment, the image and an automatic translation created with a standard phrase-based system as a suggestion to make translation easier. For English-Czech, the translations were produced by crowd-sourcing where translators had access to the source segment and the image. For the inference four test sets are provided: `test_2016_flickr`, `test_2017_flickr`, `test_2017_mscoco` and `test_2018_flickr`. We use the English sets as our source and the other languages as our targets.

Visual Data: For the multimodal model we use pre-extracted global averaged pooled image features of dimension 2048 from a pre-trained ResNet50 [27] on Imagenet [31], provided by the organizers of the competition. To exploit the image features we linearly project them to dimension 512 to fit with the model dimension.

3.2 Transformer Architecture Variations

As our NMT baseline we use the Transformer architecture described in 2.3 implemented in PyTorch. To add the image features we adapt the model in three ways:

1. We concatenate the encoder output with the image features and then feed the concatenated result in the multi-head attention. The architecture is shown in Figure 5. For the concatenation, we linearly project the image features from dimension 2048 to 512 to have the same dimension as the encoder output. The result gives us a dimension of 1024, so we apply again a linear projection to get the same dimension as the dimension of the model which is 512. The idea comes from [16] that concatenates the visual features with the context vector of an attentive RNN.

2. We use the architecture proposed in [13], where they add another multi-head attention layer in their Transformer with normalization and residual connection between the layer that attends to the encoder and the feed-forward network, that attends on the visual features. To this end, we linearly project the image features to have the same dimension as the whole model. The model is shown in Figure 6.

3. We multiply the target word embeddings in the decoder with the image features (see Fig:7), as done in [16] for an RNN encoder-decoder architecture.

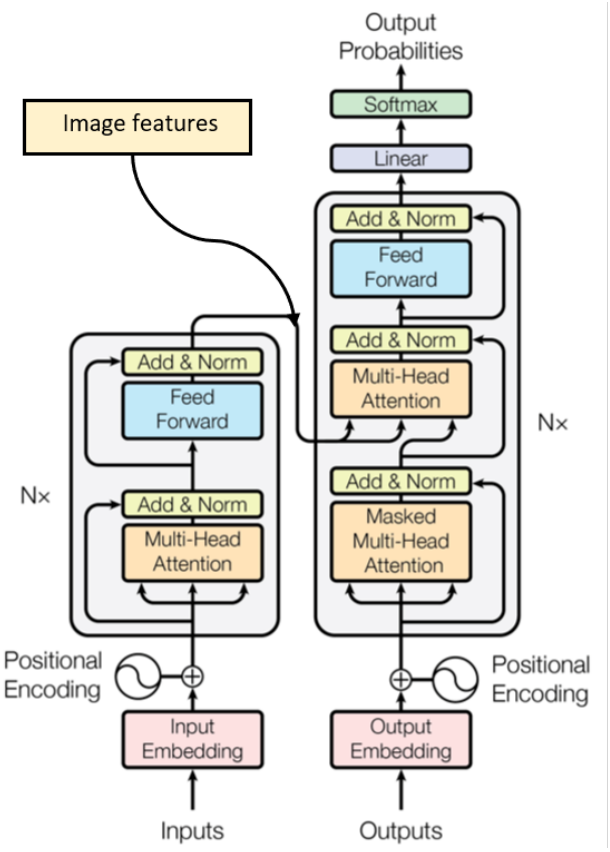

Figure 5: Multimodal Transformer with concatenated image features with encoder output.

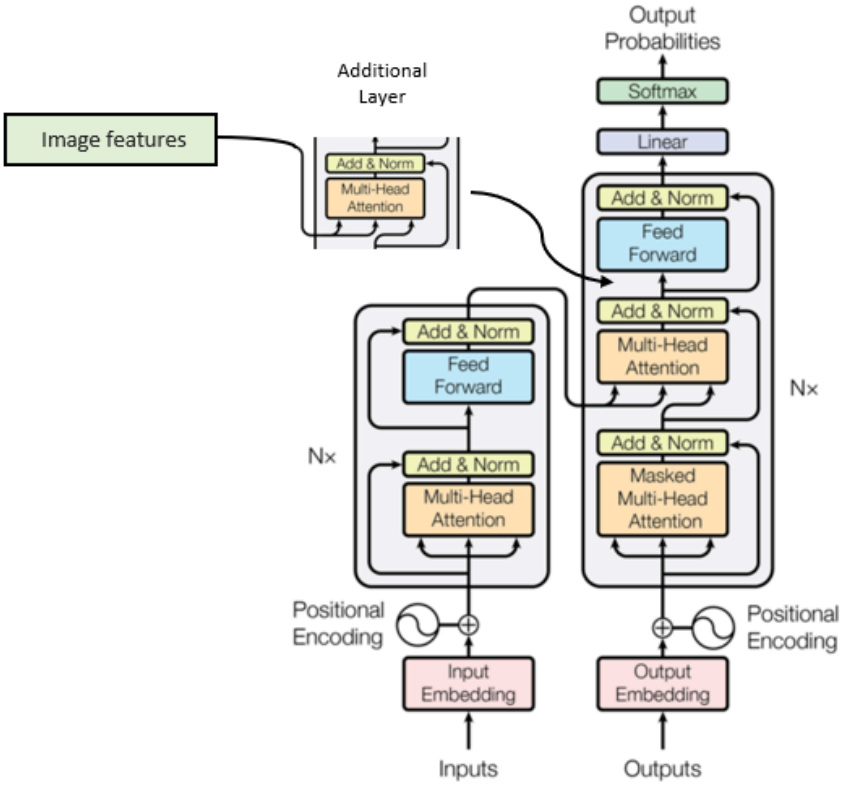

Figure 6: Multimodal Transformer with additional multi-head attention layer for the image features.(Image retrieved from [2])

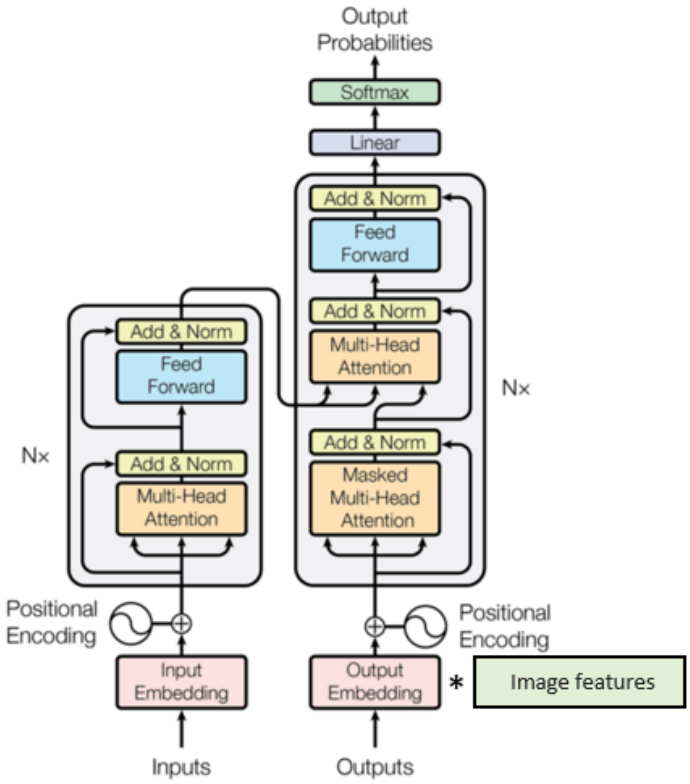

Figure 7: Multimodal Transformer where the target embeddings are multiplied with the image features. (Image retrieved from [2])

3.3 Training and Hyperparameters

We train our model on a Tesla V100-SXM2 32GB. For every training process we use early stopping and stop training when the validation accuracy does not improve for more than 10 epochs. We start training the text-only model for the English-to-French translation with $N = 6$ layers, $h = 8$ heads, model dimension $d_{model} = 512$, hidden layer dimension $d\_hid\_inner = 1024$, $dropout = 0.1$ [32] and $4000 = warmup$ steps, as stated in the original paper [2]. We use Adam optimizer [33], with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. Label smoothing of value $ls = 0.1$ is used, too. The learning rate is varied in a way that it is increased linearly for the first $warmup$ steps training steps, and decreased thereafter proportionally to the inverse square root of the step number. The used formula is the following:

$$lr = d_{model}^{-0.5} \cdot min(step^{-0.5}, step \cdot warmup^{-1.5}) \tag{3.1}$$

The result shows that the model does not generalize, which is reasonable since the Transformer is meant to be trained with millions of data while our data set only contains 30000 pairs. We start decreasing the number of layers to reduce the parameters of the model and plot the validation accuracy for the different layers. We notice that the best validation accuracy is given for N = 3 layers, as shown in Figure (8). Thus, we decide to continue the implementation using 3 layers of identical encoders and decoders instead of 6. Decreasing the number of heads had no effect on the performance of the model.

For 3 layers of identical encoders and decoders, although the validation accuracy is the highest among the others, we notice a slight overfit shown in Figure 9. We experiment with the hyperparameters and manage to reduce the overfit by increasing the dropout to

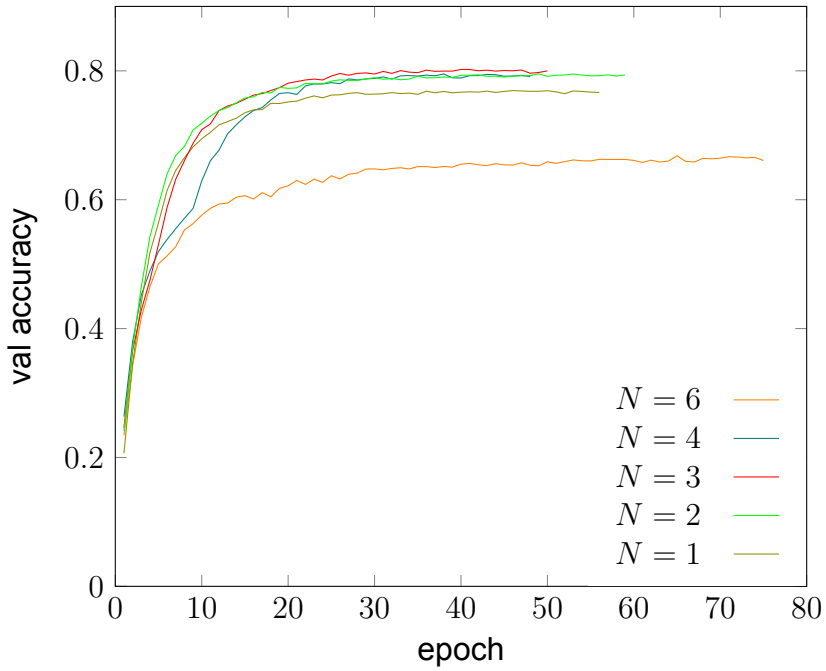

Figure 8: Validation accuracies for different number of N layers of identical encoders and decoders

0.3. We show the results in Figure 10. We train all the multimodal models with the same parameters.

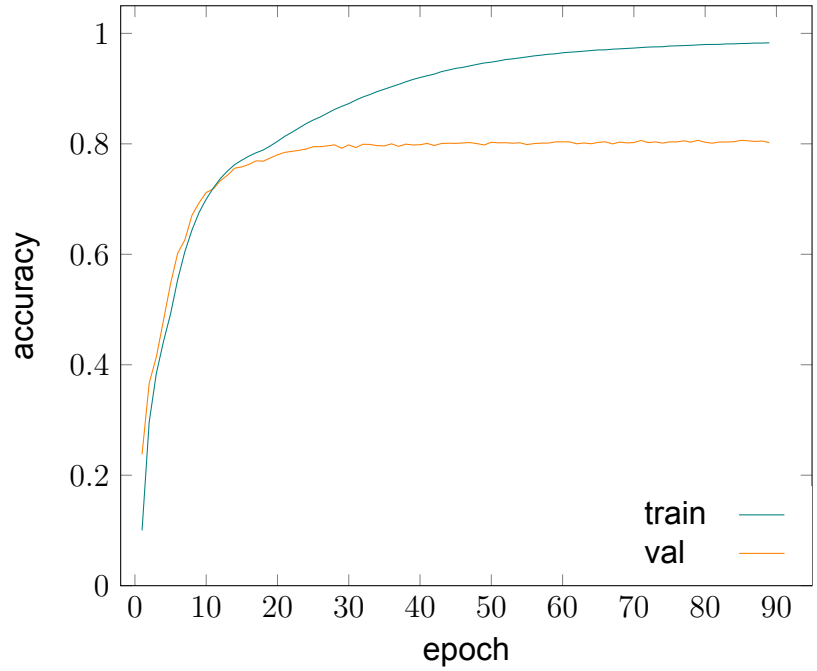

Figure 9: Training and validation accuracy for N=3 layers of identical encoders and decoders with dropout 0.1.

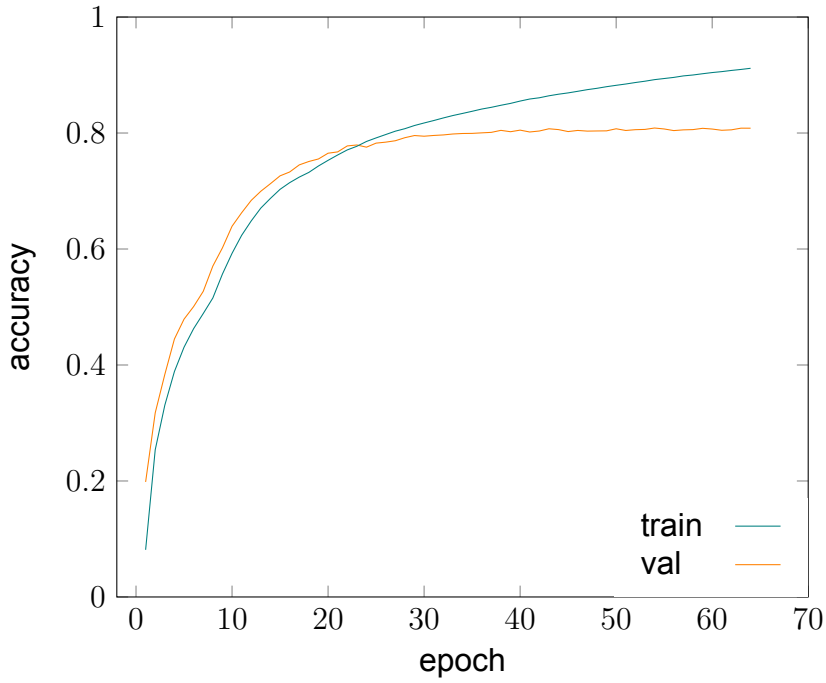

Figure 10: Training and validation accuracy for N=3 layers of identical encoders and decoders, dropout=0.3. It is obvious that there is much less overfit when increasing the dropout than the previous set of parameters shown in Fig. 9.

4. EVALUATION

4.1 Accuracy and Perplexity

We present the training and validation accuracy and perplexity for each system that was trained in every language. We compute accuracy by taking the number of correct words divided by the total number of words, while perplexity is the exponential of our loss. We present the results for all systems in Tables 1, 2 and 3. For English-to-French presented in Table 1, the last system $T_{emb*img}$ where the image features are multiplied with the target embeddings achieves the highest validation accuracy, while our baseline T_{text} has the lowest accuracy and the biggest overfit between the models. For both English-to-German and English-to-Czech the $T_{conc(text,img)}$ model performs best in terms of validation. The $T_{img\_attn\_layer}$ gives the lowest validation accuracy for German and Czech, but lower perplexity in comparison to the text-only network.

4.2 Evaluation Metrics

We evaluate out results using BLEU [34], METEOR [35] and TER [36] metrics. We use the `MultEval` [37] to compute the mentioned metrics as suggested in the shared task. We perform the translation of the `test_2016_flickr` from English to all three target languages French, German and Czech, while for the `test_2017_flickr` and `test_2017_mscoco` we translate only for English to French and German since we don't have a ground truth for Czech to evaluate our results. To evaluate our results on the English test set of 2018 `test_2018_flickr` we are not given a ground truth, but there is the option of submitting the predictions on the competition [7]. We don't submit the results on the competition because of time restrictions but we leave that for future work. The results are presented and discussed in the next chapter 4.

Tables 4, 5 and 6 summarize the average BLEU, METEOR and TER scores obtained by our systems for the English-to-French test sets. In all cases the text-only baseline system has achieved the best scores, while the Transformer with the extra multi-head attention layer seems to be very close to the baseline results. Metric scores are averages over multiple runs.

The results for the English-to-German translations shown in Tables 7, 8 and 9 suggest overall the baseline performs better than the multimodal models. The text-only and the Transformer with the additional layer perform equally well for the `test_2017_flickr` in

Table 1: Accuracy and perplexity for the training and validation of every system for English-to-French.

EN$\rightarrow FR$	Train Acc %	Val Acc %	Train PPL	Val PPL
T_{text}	91.291	80.836	4.41	2.88
$T_{conc(text,img)}$	88.117	80.941	4.96	2.68
$T_{img\_attn\_layer}$	87.843	80.843	5.09	2.62
$T_{emb*img}$	90.162	81.292	4.60	2.67

Table 2: Accuracy and perplexity for the training and validation of every system for English-to-German.

EN$\rightarrow DE$	Train Acc %	Val Acc %	Train PPL	Val PPL
T_{text}	85.327	71.451	5.37	4.30
$T_{conc(text,img)}$	82.054	72.154	6.02	3.90
$T_{img\_attn\_layer}$	83.730	71.437	5.71	4.24
$T_{emb*img}$	85.373	71.321	5.35	4.12

Table 3: Accuracy and perplexity for the training and validation of every system for English-to-Czech.

EN$\rightarrow CS$	Train Acc %	Val Acc %	Train PPL	Val PPL
T_{text}	78.792	65.477	7.03	6.24
$T_{conc(text,img)}$	72.118	66.069	9.17	5.51
$T_{img\_attn\_layer}$	76.861	65.141	7.59	6.05
$T_{emb*img}$	77.671	65.689	7.28	5.99

Table 8 according to the BLEU score, while for the other two scores the text-only achieves the highest METEOR and the multimodal the lowest TER score.

As a reference to compare our results we should mention the best scores between the submissions of the shared task. The best scores for the `test_2016_flickr` are given by [11], that achieves BLEU, METEOR and TER scores of 53.2, 34.2 and 48.7 respectively for the EN$\rightarrow$ DE [6] and for EN$\rightarrow$ FR BLEU of 52.6 and METEOR of 55.3. As mentioned in [38], the best performance for the `test_2017_flickr` is given by [16] where they get BLEU 55.9, METEOR 72.1 and TER 28.4 scores for the En$\rightarrow$ DE while for the En$\rightarrow$ FR they get BLEU 33.4, METEOR 54.0 and TER 48.5 scores. The same system achieves the best performance for the `test_2017_mscoco` with BLEU, METEOR and TER scores of 45.9, 65.9 and 34.2 respectively for the EN$\rightarrow$ FR and 28.7, 48.9 and 52.5 for the EN$\rightarrow$ DE. We should mention that these scores are achieved by unconstrained submissions that use additional data that we don't.

4.3 Example Translations

We present several predicted translations for the different systems. The results for the multimodal systems, although not perfect, seem to be nicer with less uknown output tokens. Figure 11 reveals inconsistencies in the dataset since the source mentions a violin case, although in the image we see a guitar that is correctly used for the German language. In Figure 15, all systems have output the same correct prediction. In Section 4.4 we visualize the attention scores of the different layers and heads of Encoders and Decoders for every system for image 15. The predicted examples are presented in Figures 11 - 15.

Table 4: Evaluation metrics of the four systems for English-to-French translation of the `test_2016_flickr`.

EN→ FR `test_2016_flickr`	**System**	**Avg**	$\overline{s}_{\textbf{sel}}$	p**-value**
BLEU ↑	T_{text}	56.8	0.9	-
	$T_{conc(text,img)}$	57.0	0.8	0.69
	$T_{img\_attn\_layer}$	57.4	0.9	0.28
	$T_{emb*img}$	**58.1**	0.9	0.01
METEOR ↑	T_{text}	71.1	0.7	-
	$T_{conc(text,img)}$	71.4	0.7	0.48
	$T_{img\_attn\_layer}$	71.6	0.6	0.20
	$T_{emb*img}$	**71.9**	0.9	0.03
TER ↓	T_{text}	27.2	0.6	-
	$T_{conc(text,img)}$	27.6	0.8	0.45
	$T_{img\_attn\_layer}$	26.9	0.6	0.52
	$T_{emb*img}$	**26.4**	0.6	0.06

Table 5: Evaluation metrics of the three systems for English-to-French translation of the `test_2017_flickr`.

EN→ FR `test_2017_flickr`	**System**	**Avg**	$\overline{s}_{\textbf{sel}}$	p**-value**
BLEU ↑	T_{text}	**49.3**	0.9	-
	$T_{conc(text,img)}$	48.7	0.9	0.22
	$T_{img\_attn\_layer}$	48.6	0.9	0.18
	$T_{emb*img}$	**49.3**	0.9	0.94
METEOR ↑	T_{text}	65.0	0.7	-
	$T_{conc(text,img)}$	64.7	0.7	0.34
	$T_{img\_attn\_layer}$	64.5	0.7	0.16
	$T_{emb*img}$	**65.1**	0.7	0.75
TER ↓	T_{text}	33.2	0.7	-
	$T_{conc(text,img)}$	33.4	0.7	0.57
	$T_{img\_attn\_layer}$	33.7	0.7	0.27
	$T_{emb*img}$	**32.7**	0.7	0.36

Table 6: Evaluation metrics of the three systems for English-to-French translation of the `test_2017_mscoco`.

EN→ FR `test_2017_mscoco`	**System**	**Avg**	$\bar{s}_{\text{sel}}$	p**-value**
BLEU ↑	T_{text}	38.2	1.2	-
	$T_{conc(text,img)}$	**39.5**	1.2	0.08
	$T_{img\_attn\_layer}$	39.0	1.1	0.22
	$T_{emb*img}$	39.3	1.2	0.12
METEOR ↑	T_{text}	57.0	1.0	-
	$T_{conc(text,img)}$	57.8	1.0	0.18
	$T_{img\_attn\_layer}$	**58.1**	0.9	0.05
	$T_{emb*img}$	57.9	1.0	0.08
TER ↓	T_{text}	39.6	1.0	-
	$T_{conc(text,img)}$	**38.5**	0.9	0.09
	$T_{img\_attn\_layer}$	38.6	0.9	0.10
	$T_{emb*img}$	38.8	0.9	0.18

Table 7: Evaluation metrics of the three systems for English-to-German translation of the `test_2016_flickr`.

EN→ DE `test_2016_flickr`	**System**	**Avg**	$\bar{s}_{\text{sel}}$	**p-value**
BLEU ↑	T_{text}	35.6	0.8	-
	$T_{conc(text,img)}$	**35.7**	0.8	0.90
	$T_{img\_attn\_layer}$	35.4	0.8	0.62
	$T_{emb*img}$	34.4	0.8	0.04
METEOR ↑	T_{text}	53.6	0.7	-
	$T_{conc(text,img)}$	**54.5**	0.7	0.08
	$T_{img\_attn\_layer}$	54.0	0.7	0.47
	$T_{emb*img}$	53.1	0.7	0.24
TER ↓	T_{text}	43.0	0.7	-
	$T_{conc(text,img)}$	**42.4**	0.7	0.25
	$T_{img\_attn\_layer}$	42.7	0.7	0.51
	$T_{emb*img}$	43.0	0.7	0.91

Table 8: Evaluation metrics of the three systems for English-to-German translation of the `test_2017_flickr`.

EN→ DE `test_2017_flickr`	**System**	**Avg**	$\bar{s}_{\text{sel}}$	p**-value**
BLEU ↑	T_{text}	27.8	0.8	-
	$T_{conc(text,img)}$	**28.3**	0.8	0.44
	$T_{img\_attn\_layer}$	27.4	0.8	0.51
	$T_{emb*img}$	27.1	0.8	0.24
METEOR ↑	T_{text}	46.3	0.6	-
	$T_{conc(text,img)}$	**46.6**	0.6	0.57
	$T_{img\_attn\_layer}$	46.2	0.6	0.76
	$T_{emb*img}$	45.8	0.7	0.28
TER ↓	T_{text}	51.7	0.8	-
	$T_{conc(text,img)}$	**50.4**	0.7	0.02
	$T_{img\_attn\_layer}$	51.4	0.8	0.64
	$T_{emb*img}$	52.3	0.8	0.34

Table 9: Evaluation metrics of the three systems for English-to-German translation of the `test_2017_mscoco`.

EN→ DE `test_2017_mscoco`	**System**	**Avg**	$\bar{s}_{\text{sel}}$	p**-value**
BLEU ↑	T_{text}	24.1	1.1	-
	$T_{conc(text,img)}$	**24.9**	1.2	0.33
	$T_{img\_attn\_layer}$	23.9	1.1	0.78
	$T_{emb*img}$	23.6	1.1	0.47
METEOR ↑	T_{text}	41.8	0.9	-
	$T_{conc(text,img)}$	**43.1**	0.9	0.04
	$T_{img\_attn\_layer}$	42.1	0.9	0.63
	$T_{emb*img}$	42.2	0.9	0.54
TER ↓	T_{text}	56.2	1.6	-
	$T_{conc(text,img)}$	**53.4**	1.1	0.02
	$T_{img\_attn\_layer}$	56.1	1.3	0.83
	$T_{emb*img}$	54.8	1.0	0.33

Figure 11: Example translations of a `test_2016_flickr` sentence that describes the orchestra image for French and German.

SRC: small orchestra playing with open violin case in front.

TGT FR: un petit orchestre jouant avec un étui de violon ouvert devant.

T_{text}: un petit orchestre jouant avec un <unk> devant l' <unk>

$T_{conc(text,img)}$: un petit orchestre jouant avec un violon ouvert devant

$T_{img\_attn\_layer}$: un petit orchestre jouant avec un violon ouvert devant lui deux personnes.

$T_{emb*img}$: un petit orchestre jouant avec un violon ouvert devant une <unk>.

TGT DE: kleines orchester spielt im freien mit einem gitarrenkoffer auf dem boden.

T_{text}: ein kleines orchester spielt mit offenen geöffneten geöffneten geöffneten geöffneten geöffneten geöffneten geöffneten geöffneten .

$T_{conc(text,img)}$: ein kleines orchester spielt mit offenen <unk> vor einer <unk>.

$T_{img\_attn\_layer}$: ein kleiner orchester spielt mit offenen geige.

$T_{emb*img}$: kleine orchester spielt mit offenen <unk> vor einem orchester.

Figure 12: Example predictions of an English translated sentence in French and German with the corresponding image.

SRC: a girl in white and a girl in green walk past a blue car wash station.

TGT FR: une fille en blanc et une fille en vert passent devant une station de lavage de voiture bleue .

T_{text}: une fille en blanc et une fille en vert passent devant une voiture bleue.

$T_{conc(text,img)}$: une fille en blanc et une fille en vert passent devant une station de voiture bleue.

$T_{img\_attn\_layer}$: une fille en blanc et une fille en vert passent devant une voiture bleue dans le long d' une gare bleue.

$T_{emb*img}$: une fille en blanc et une fille en vert passent devant une station de voiture bleue en train de <unk> des <unk>.

TGT DE: ein mädchen in weiß und eines in grün gehen an einer blauen autowaschanlage vorbei .

T_{text}: ein mädchen in weißer und ein mädchen in grüner kleidung gehen an einem blauen auto vorbei .

$T_{conc(text,img)}$: ein mädchen in weiß und ein mädchen in grün gehen an einer blauen <unk> vorbei .

$T_{img\_attn\_layer}$: ein mädchen in weiß und ein mädchen in grün gekleidet , gehen an einer blauen <unk> vorbei .

$T_{emb*img}$: ein mädchen in weiß und ein mädchen in grün gehen an einem blauen <unk> vorbei .

Figure 13: Predicted translations of a sentence corresponding to an image of two dogs for En→ Fr and En→ De

SRC: two dogs are nuzzling each other nose to nose .

TGT FR: deux chiens nez à nez se donnent des coups de museau.

T_{text}: deux chiens se <unk> mutuellement pour se <unk>.

$T_{conc(text,img)}$: deux chiens se <unk> le nez au nez.

$T_{img\_attn\_layer}$: deux chiens se battent l' un à l' autre.

$T_{emb*img}$: deux chiens se battent l' un contre l' autre le nez.

TGT DE: zwei hunde beschnuppern sich gegenseitig nase an nase.

T_{text}: zwei hunde <unk> sich einander an der nase.

$T_{conc(text,img)}$: zwei hunde <unk> einander um die nase zu <unk>.

$T_{img\_attn\_layer}$: ein kleiner orchester spielt mit offenen geige.

$T_{emb*img}$: zwei hunde <unk> sich die nase zu der nase.

Figure 14: Predictions for EN$\rightarrow FRandEN \rightarrow DEofasentencedescribingtheimage.$

SRC: the man in a japanese cooking suit is preparing a meal for two people.

TGT FR: l' homme en tenue de cuisinier japonais prépare un repas pour deux personnes.

T_{text}: l' homme dans un <unk> préparant un repas pour deux personnes.

$T_{conc(text,img)}$, $T_{img\_attn\_layer}$, $T_{emb*img}$: l' homme en costume japonais prépare un repas pour deux personnes.

TGT DE: der mann im japanischen kochgewand bereitet ein essen für zwei personen zu.

T_{text}: der mann in einer japanischen <unk> bereitet ein essen vor.

$T_{conc(text,img)}$: der mann in einem japanischen anzug bereitet sich für zwei menschen eine mahlzeit vor .

$T_{img\_attn\_layer}$: der mann in einem japanischen anzug bereitet in einem japanischen anzug eine mahlzeit zu .

$T_{emb*img}$: der mann in der japanischen <unk> bereitet essen zu und bereitet ein gericht vor.

Figure 15: Example translation of an image of the `test_2016_flickr` for all languages.

SRC: a worker in an orange vest is using a shovel.

TGT FR: un ouvrier en gilet orange utilise une pelle.

T_{text}, $T_{conc(text,img)}$, $T_{img\_attn\_layer}$, $T_{emb*img}$: un ouvrier en gilet orange utilise une pelle.

TGT DE: ein arbeiter in einer orangefarbenen weste arbeitet mit einer schaufel.

T_{text}, $T_{img\_attn\_layer}$, $T_{emb*img}$: ein arbeiter in orangefarbener weste benutzt eine schaufel.

$T_{conc(text,img)}$: ein arbeiter in einer orangefarbenen weste benutzt eine schaufel.

TGT CS: dělník v oranžové vestě používá lopatu.

T_{text}, $T_{img\_attn\_layer}$, $T_{emb*img}$: dělník v oranžové vestě používá lopatu.

$T_{conc(text,img)}$: pracovník v oranžové vestě používá lopatu.

Table 10: Evaluation metrics of the four systems for English-to-Czech translation of the `test_2016_flickr`.

EN→ CS `test_flickr_2016`	**System**	**Avg**	$\overline{s}_{\textbf{sel}}$	p**-value**
BLEU ↑	T_{text}	27.5	0.8	-
	$T_{conc(text,img)}$	28.0	0.8	0.37
	$T_{img\_attn\_layer}$	**28.5**	0.8	0.07
	$T_{emb*img}$	28.2	0.8	0.20
METEOR ↑	T_{text}	27.0	0.4	-
	$T_{conc(text,img)}$	27.2	0.4	0.49
	$T_{img\_attn\_layer}$	27.5	0.4	0.05
	$T_{emb*img}$	**27.6**	0.4	0.01
TER ↓	T_{text}	48.1	0.9	-
	$T_{conc(text,img)}$	49.6	1.3	0.23
	$T_{img\_attn\_layer}$	46.9	0.9	0.23
	$T_{emb*img}$	**46.3**	0.7	0.01

4.4 Attention Visualization

We visualize the attention scores for every layer of Encoder and Decoder for our 4 systems for the French prediction of Figure 15. We use the `seaborn` python library to plot the attention heatmaps. In each row we present 4 heads of the multi-head attention of a layer. Layers are presented from top to bottom, with layer 1 on top and layer 3 at the bottom of every figure. The heatmaps show the evolution of the representations passing through the different layers and how every head and step in the encoding and decoding attends on different word creating new representation subspaces.

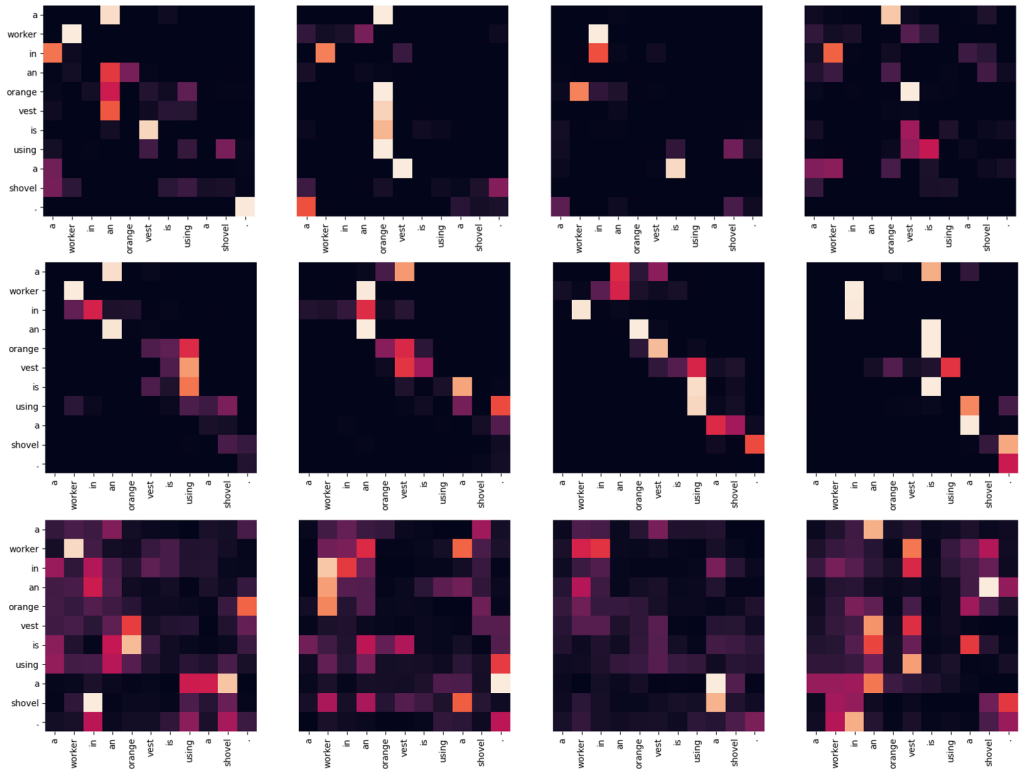

Figure 16: Encoder Layers with 4 heads of the baseline T_{text}. At the top row we see layer 1, layer 2 in the middle and layer 3 at the bottom.

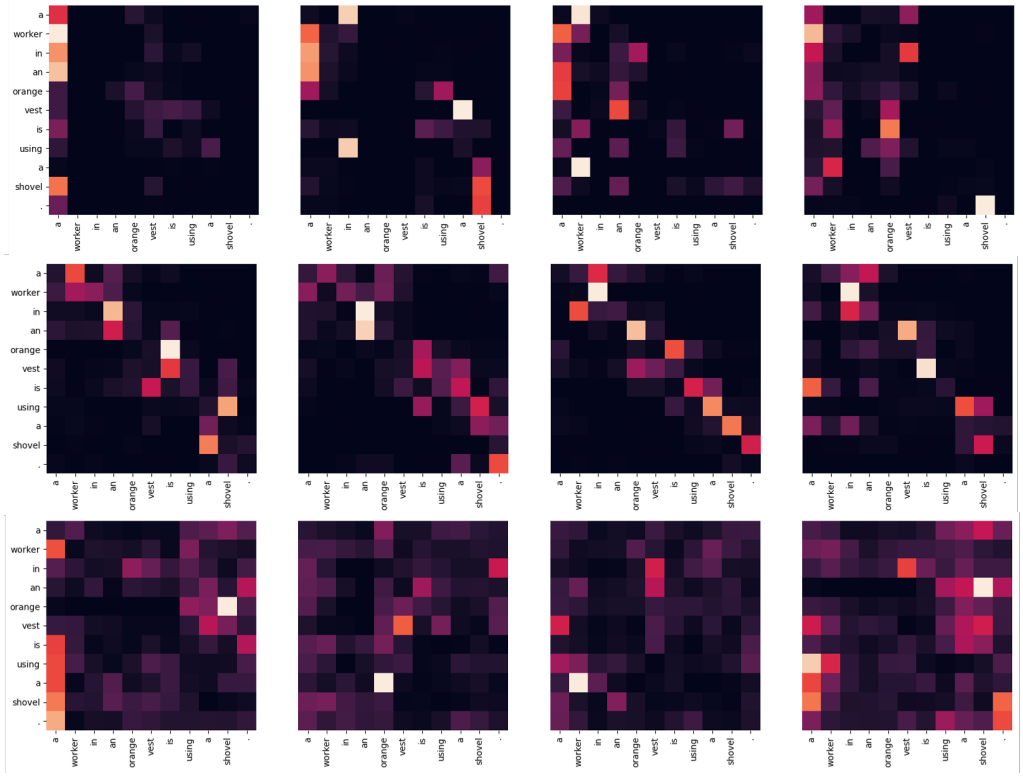

Figure 17: Encoder Layers for the multimodal system $T_{conc(text,img)}$.

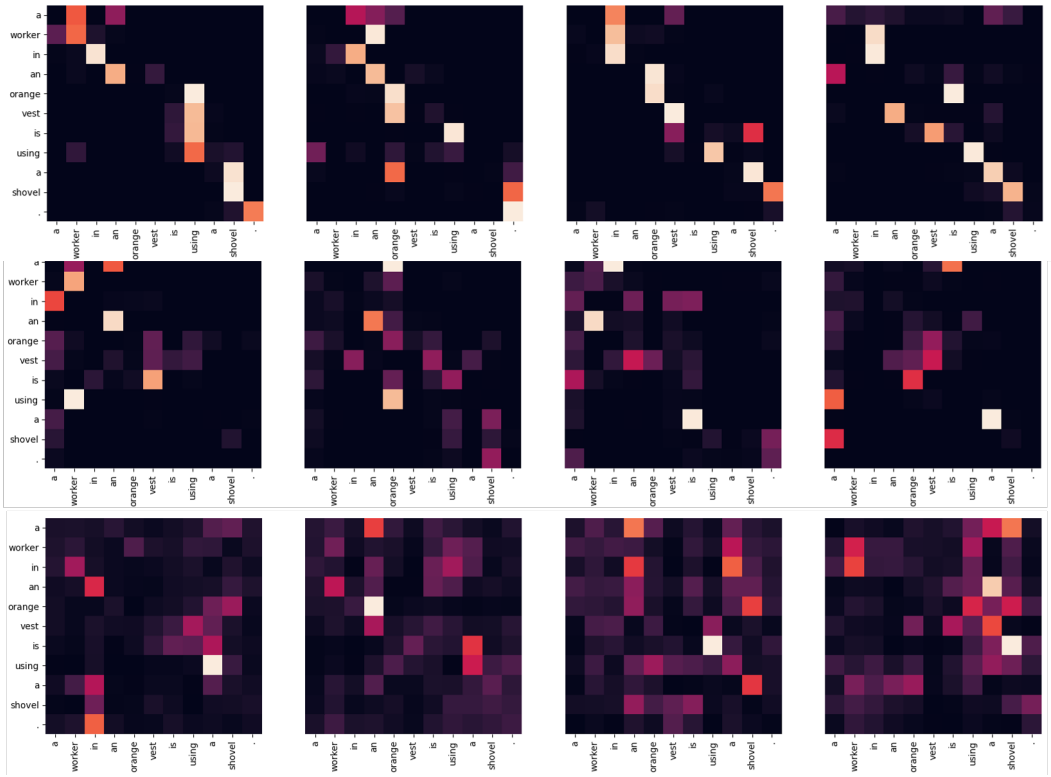

Figure 18: Encoder Layers for the multimodal system $T_{img\_attn\_layer}$.

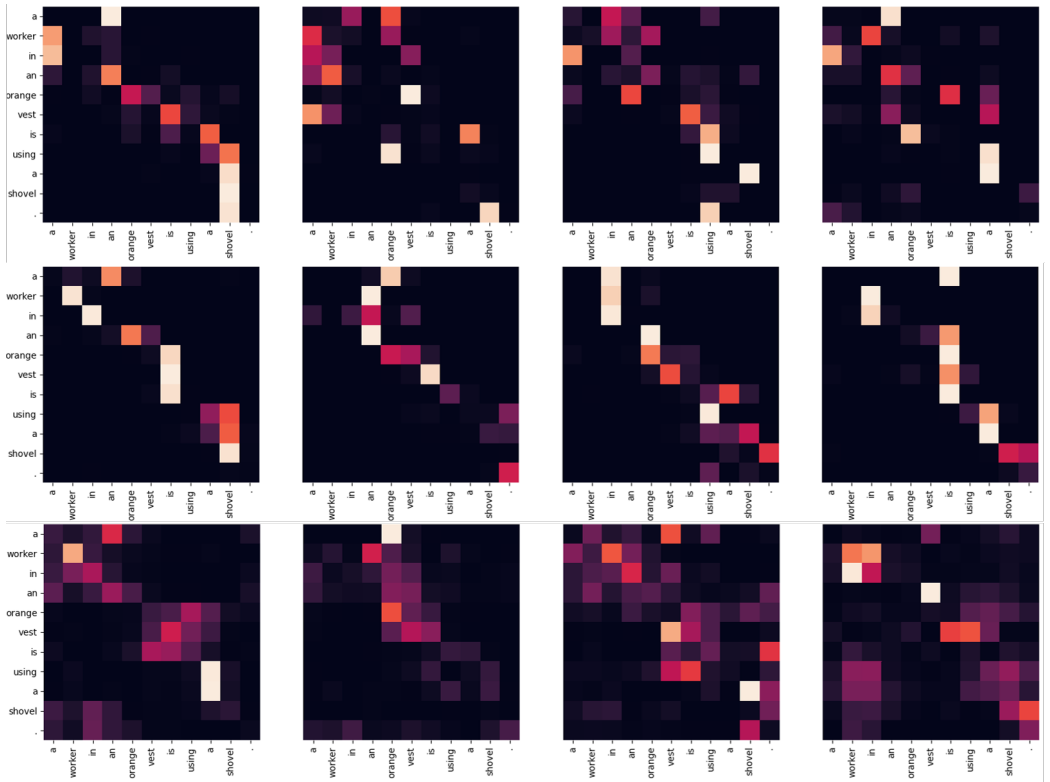

Figure 19: Encoder Layers for the multimodal system $T_{emb*img}$.

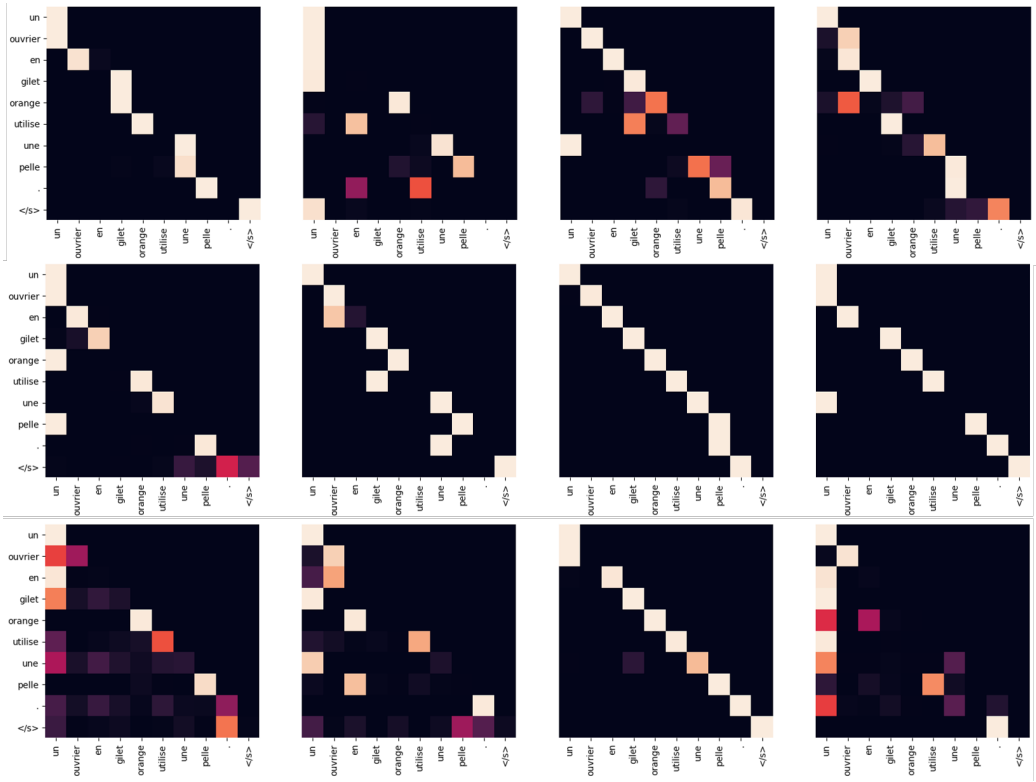

Figure 20: Decoder Self Layers with 4 heads of the baseline T_{text}. At the top row we see layer 1, in the middle layer 2 and layer 3 at the bottom.

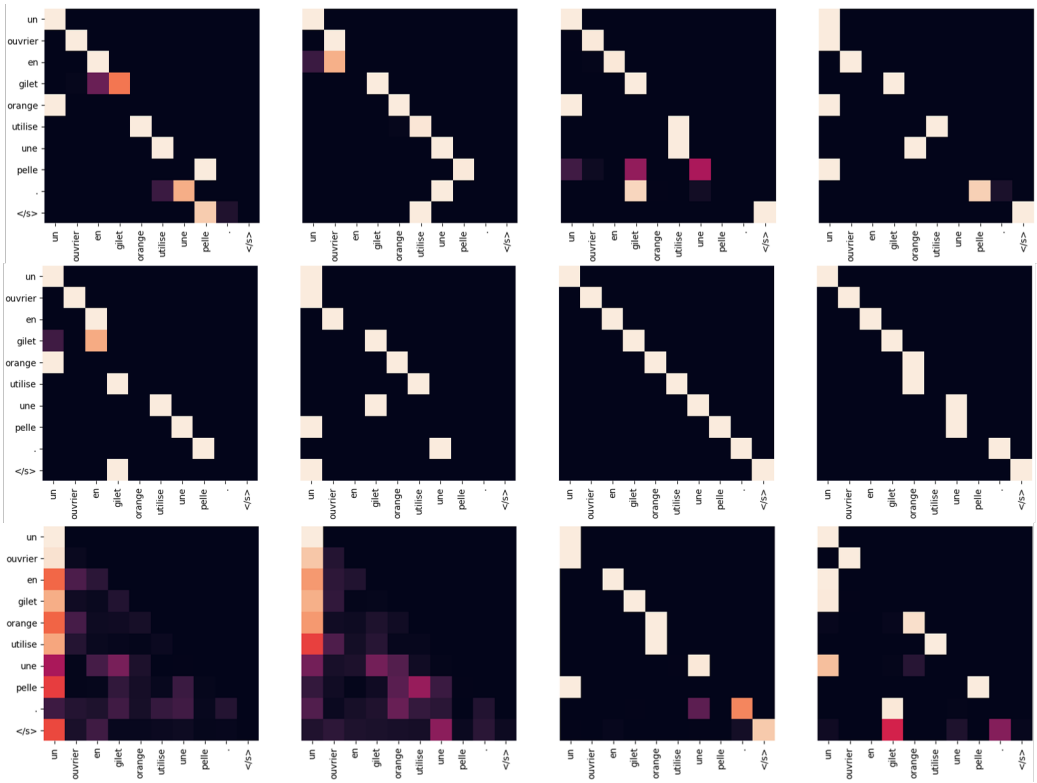

Figure 21: Decoder Self Layers for the multimodal system $T_{conc(text,img)}$.

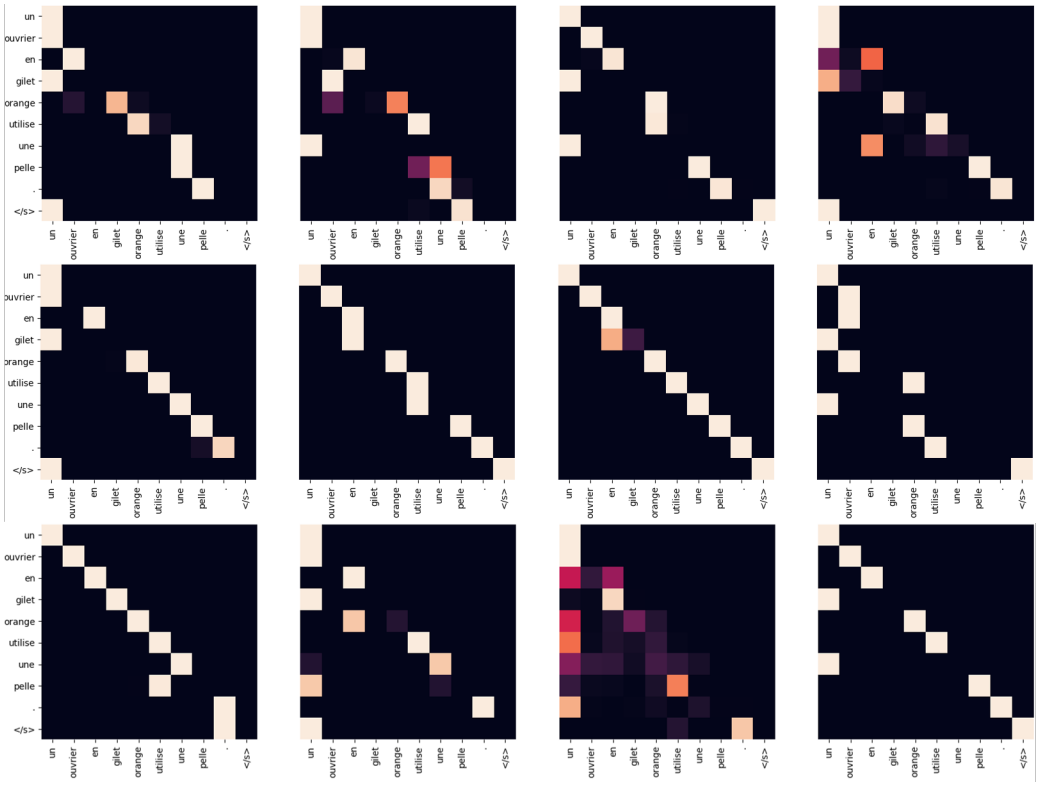

Figure 22: Decoder Self Layers for the multimodal system $T_{img\_attn\_layer}$.

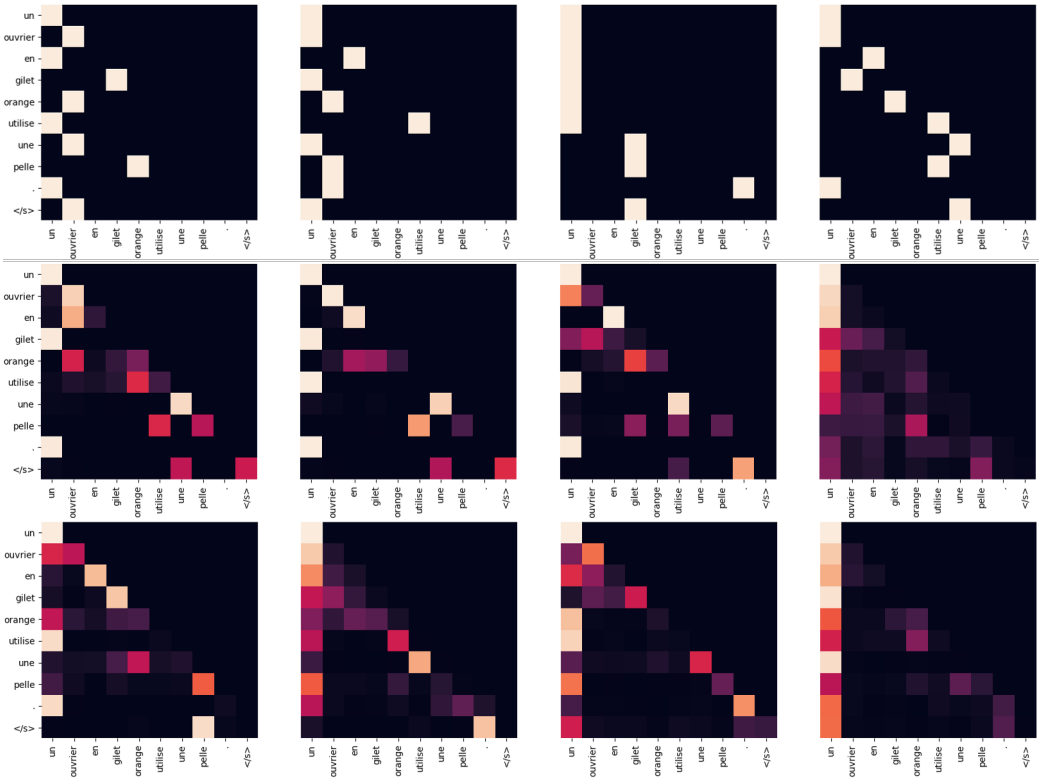

Figure 23: Decoder Self Layers for the multimodal system $T_{emb*img}$.

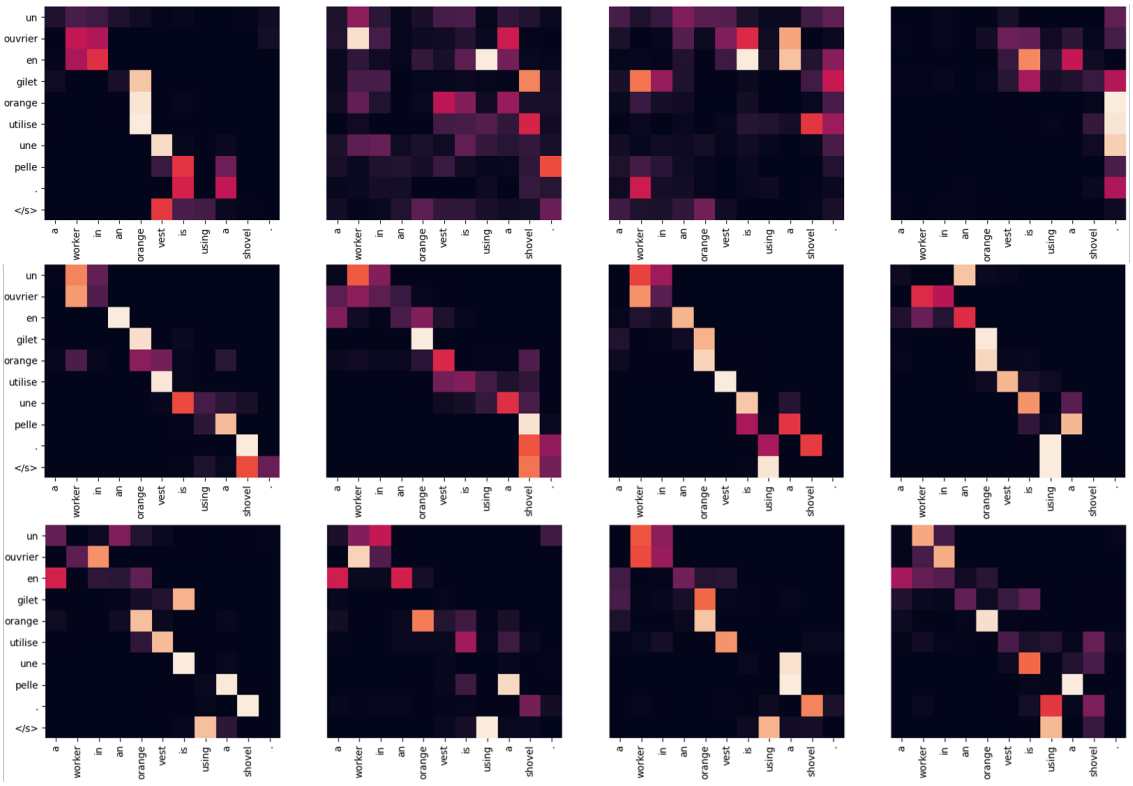

Figure 24: Decoder Source Layers 1-3 (top-bottom) with 4 heads of the baseline T_{text}.

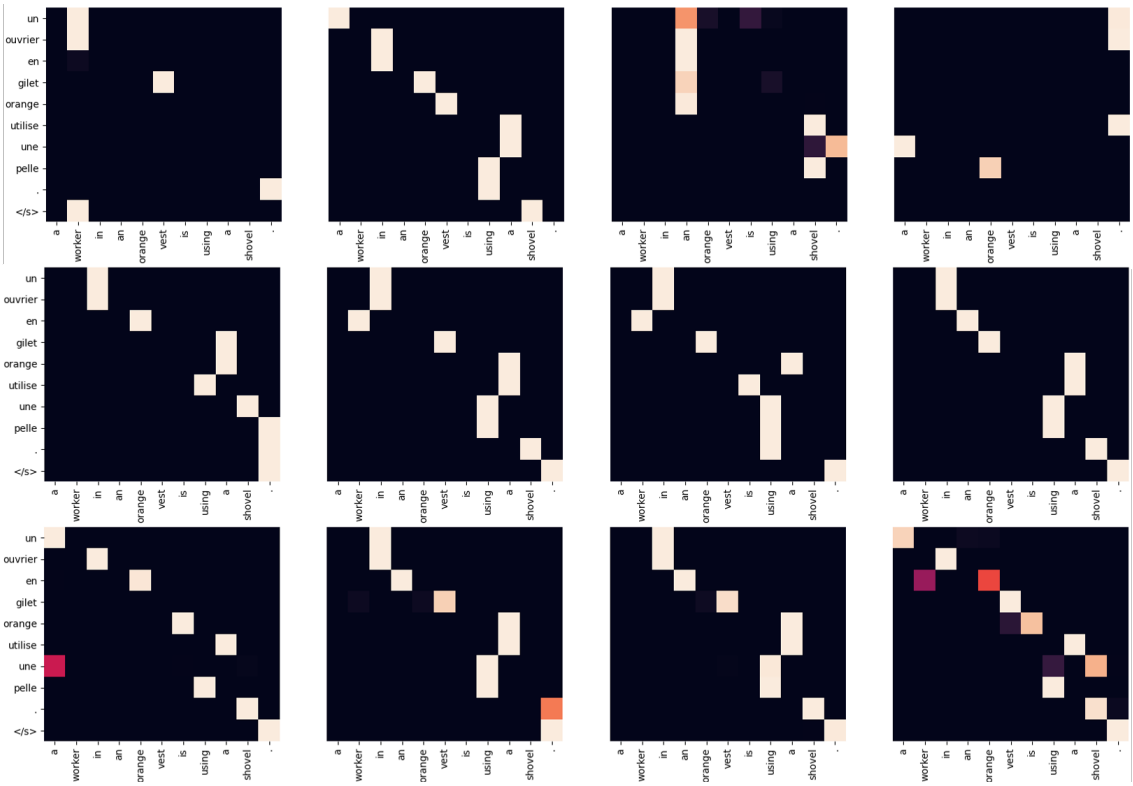

Figure 25: Decoder Source Layers for the multimodal system $T_{conc(text,img)}$.

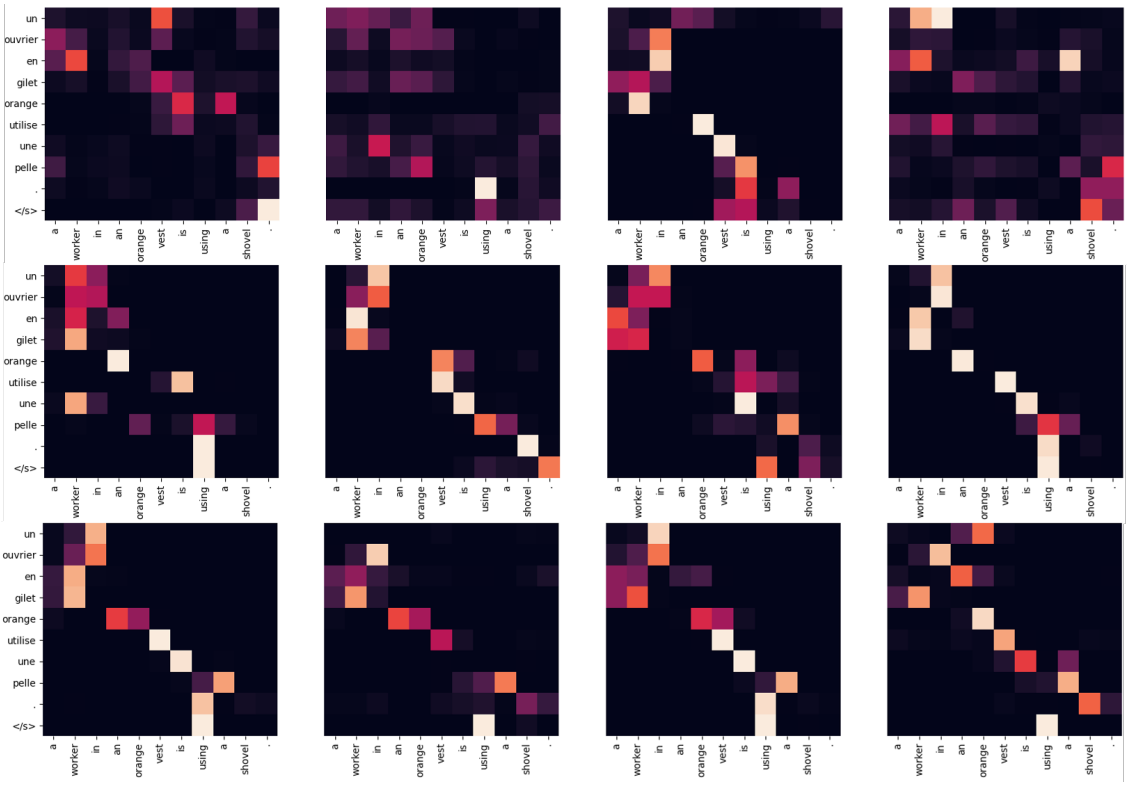

Figure 26: Decoder Source Layers for the multimodal system $T_{img\_attn\_layer}$.

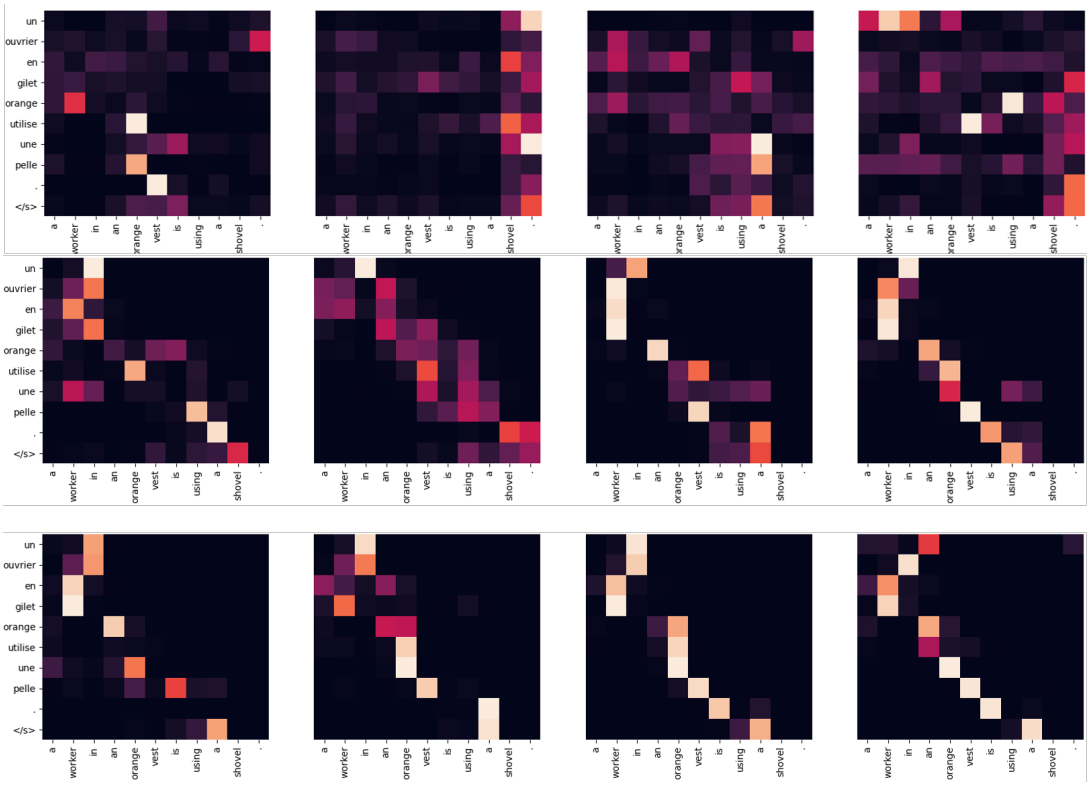

Figure 27: Decoder Source Layers for the multimodal system $T_{emb*img}$.

4.5 Color Masking

The results presented in 4.2 suggest that the the image features used to enhance the translation are only marginally beneficial or even completely unnecessary. To evaluate whether our multimodal models take advantage of the visual modalities we apply a masking with a special [c] token on color words in our source train and test 2016 flickr for English-to-French as suggested in [39]. The masking happens in around 2.5% of the words. Although there are a few failures, the predictions show promising results and partially prove that in many cases where the text-only model fails and depends only on biases to predict the correct color, the multimodal models succeed. In Figure 28 we can see the gender bias of the training set, as it seems that most of the appearing sentences with girls are with pink outfits. Two of the systems are relied on this bias, while two of the multimodal networks seem to recognize that the outfit of the little girl is blue. This can be interpreted as that the visual features are actually used to improve the translation and that the improvement is only marginal, because of the nature of the only existing dataset that is small and descriptive which makes the text sufficient to perform the translation. Examples of the masked predictions are shown in Figures 28 - 35.

Figure 28: Example translation with mask token in the source sentence.

SRC: a little girl in a [c] outfit is climbing on metal railings in the street.

TGT: une petite fille en tenue bleue monte sur des barreaux métalliques dans la rue.

T_{text}: une petite fille en tenue rose escalade du métal dans la rue.

$T_{conc(text,img)}$: une petite fille en tenue bleue escalade des sur des métalliques dans la rue.

$T_{img\_attn\_layer}$: une petite fille en tenue bleue escalade sur des de métal dans la rue.

$T_{emb*img}$: une petite fille en tenue rose escalade des métalliques dans la rue.

Figure 29: Translation results for masking the color [red] that refers to the word ”shirt”.

SRC: a woman in a [c] shirt raising her arm to the passing crowd below.

TGT: une femme en t-shirt rouge saluant la foule qui passe en bas.

T_{text}: une femme en t-shirt **noir** levant son bras vers la foule qui s’en contrebas.

$T_{conc(text,img)}$: une femme en t-shirt rouge levant son bras vers la foule en contrebas.

$T_{img\_attn\_layer}$: une femme en t-shirt rouge lève son bras devant la foule en contrebas.

$T_{emb*img}$: une femme en t-shirt bleu lève son bras vers la foule qui passe en contrebas.

Figure 30: Translation results for masking the color [red] that refers to the word ”chairs”.

SRC: a woman reads a book while sitting in a row of [c] chairs.

TGT: une femme lit un livre tout en étant assise dans une rangée de fauteuils rouges.

T_{text}: une femme lit un livre, assise en rang d’une rangée de chaises bleues.

$T_{conc(text,img)}$: une femme lit un livre tandis qu’elle est assise en rangée de chaises rouges.

$T_{img\_attn\_layer}$: une femme lit un livre assis en rang dans une rangée de chaises rouges.

$T_{emb*img}$: une femme lit un livre, assise en rang de chaises bleues.

Figure 31: Translation results for masking the color [red] that refers to the word ”vest”.

SRC: a woman in a [c] vest working at a computer.

TGT: une femme en gilet rouge travaillant sur un ordinateur.

T_{text}: une femme en veste jaune travaillant sur un ordinateur.

$T_{conc(text,img)}$: une femme en gilet rouge travaillant sur un ordinateur.

$T_{img\_attn\_layer}$: une femme en veste jaune travaillant sur un ordinateur.

$T_{emb*img}$: une femme en veste **noire** travaillant sur un ordinateur.

Figure 32: Color masking results for masking the word [green] that refers to the ”arm band”.

SRC: girl with [c] arm band , hair ties, and earrings stands outside.

TGT: une fille avec un brassard vert, des rubans dans les cheveux et des boucles d’oreilles est debout dehors.

T_{text}: une fille avec un bras rose, des cheveux et des se tient debout dehors.

$T_{conc(text,img)}$: une fille avec le bras d’un groupe de cheveux verts, des et des sont debout dehors.

$T_{img\_attn\_layer}$: une fille avec un bras rose, des cheveux, et des se tient à l’ extérieur.

$T_{emb*img}$: une fille avec un bras jaune, des cheveux, et des est debout dehors.

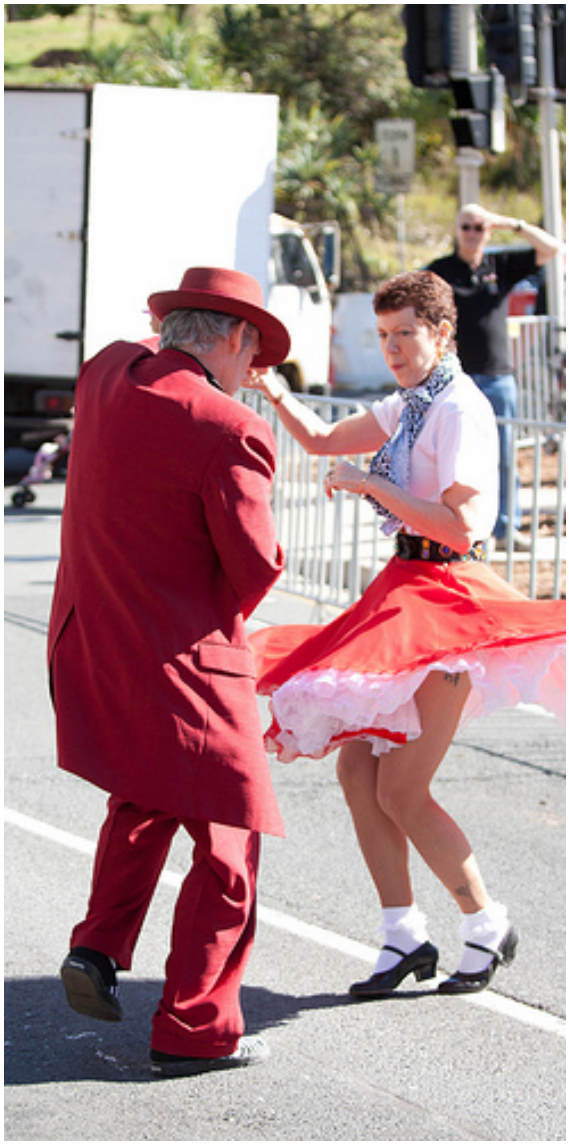

Figure 33: Color masking results for masking the word [red] that refers to the ”suit”.

SRC: a man with a [c] suit is dancing with a lady.

TGT: un homme avec un costume rouge danse avec une femme.

T_{text}: un homme avec un costume **noir** danse avec une femme.

$T_{conc(text,img)}$: un homme avec un costume rouge danse avec une femme.

$T_{img\_attn\_layer}$: un homme avec un costume **noir** danse avec une femme.

$T_{emb*img}$: un homme avec un costume **noir** danse avec une femme.

Figure 34: Color masking results for masking the word [brown] that refers to the ”hair” of the woman.

SRC: a woman with [c] hair sitting on a bench outside a cafe.

TGT: une femme aux cheveux bruns assise sur un banc devant un café.

T_{text}: une femme aux cheveux **noirs** assise sur un banc devant un café.

$T_{conc(text,img)}$: une femme aux cheveux bruns assise sur un banc devant un café.

$T_{img\_attn\_layer}$: une femme aux cheveux bruns assise sur un banc devant un café.

$T_{emb*img}$: une femme aux cheveux bruns assise sur un banc devant un café.

Figure 35: Masking translation results for masking the word [white] that refers to the ”dog”.

SRC: a [c] dog on mountainside turns to face something offstage, sky in background.

TGT: un chien blanc sur le versant d’une montagne se tourne pour faire face à quelque chose hors cadre, avec le ciel en arrière-plan .

T_{text}: un chien **noir** sur une se retourne pour se faire face à quelque chose, avec un ciel en arrière-plan.

$T_{conc(text,img)}$: un chien blanc sur le visage tourne pour aller quelque chose de, avec un ciel en arrière-plan.

$T_{img\_attn\_layer}$: un chien blanc se retourne face à quelque chose, le ciel, avec un ciel en arrière-plan.

$T_{emb*img}$: un chien blanc se retourne pour se tourne quelque chose, avec le ciel en arrière-plan.

5. CONCLUSION AND FUTURE WORK

Neural Machine Translation systems take only text into account and disregard other types of modalities. In this work, we investigated several ways of using image features extracted from pre-trained neural networks as additional input to a state-of-the-art neural machine translation system as the Transformer in order to ground and improve the quality of the translation. We considered four systems to perform the task:

1. a text-only Transformer architecture as our baseline system,
2. a Transformer that exploits the visual features by concatenating them with the encoder output to feed afterwards to the decoder,
3. a Transformer with an additional multi-head attention layer between the source multi-head layer and the feed-forward network, that attends on the image features,
4. a Transformer with multiplied target embeddings with image features.

The used code is available in the following github repository:

`https://github.com/koninik/mmt`

Our results suggest that even though performance measures and scores do not dramatically improve in comparison to the text-only system and there is no winning system, the resulted translations in many cases seem to be nicer and more human-like.

We further evaluate whether the image features are exploited by the multimodal systems by masking words that refer to colors. The results from this application seem promising since the text-only architecture fails in predicting the correct color and and seems to be relied only on biases that appear during training.

Variations of the same model were performed each one separately giving decent results in terms of performance. There is a possibility for future work in combining all multimodal cases in one. The predictions also showed that there are several inconsistencies in the ground truth and several gender biases. It would be definitely worth removing biases by data augmentation or other measures or even create a completely new dataset that takes all these aspects into account.

ABBREVIATIONS - ACRONYMS

SMT	Statistical Machine Translation
NMT	Neural Machine Translation
MMT	Multimodal Machine Translation
RNN	Recurrent Neural Networks
Seq2Seq	Sequence-to-sequence
BPTT	Backpropagation Through Time (BPTT)

BIBLIOGRAPHY

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[3] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," 2016.

[4] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," 2015.

[5] "ACL 2016 FIRST CONFERENCE ON MACHINE TRANSLATION (WMT16)." http://www.statmt.org/wmt16/multimodal-task.html, 2016.

[6] L. Specia, S. Frank, K. Sima'an, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, (Berlin, Germany), pp. 543–553, Association for Computational Linguistics, Aug. 2016.

[7] "Multimodal translation shared task." https://competitions.codalab.org/competitions/19917#learn_the_details-overview, 2018.

[8] I. Calixto, D. Elliott, and S. Frank, "DCU-UvA multimodal MT system report," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, (Berlin, Germany), pp. 634–638, Association for Computational Linguistics, Aug. 2016.

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

[10] J. Hitschler, S. Schamoni, and S. Riezler, "Multimodal pivots for image caption translation," 2016.

[11] O. Caglayan, L. Barrault, and F. Bougares, "Multimodal attention for neural machine translation," 2016.

[12] S.-A. Grönroos, B. Huet, M. Kurimo, J. Laaksonen, B. Merialdo, P. Pham, M. Sjöberg, U. Sulubacak, J. Tiedemann, R. Troncy, and R. Vázquez, "The MeMAD submission to the WMT18 multimodal translation task," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, (Belgium, Brussels), pp. 603–611, Association for Computational Linguistics, Oct. 2018.

[13] J. Helcl, J. Libovický, and D. Variš, "Cuni system for the wmt18 multimodal translation task," 2018.

[14] R. Zheng, Y. Yang, M. Ma, and L. Huang, "Ensemble sequence level training for multimodal mt: Osubaidu wmt18 multimodal machine translation system report," 2018.

[15] C. Lala, P. S. Madhyastha, C. Scarton, and L. Specia, "Sheffield submissions for WMT18 multimodal translation shared task," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, (Belgium, Brussels), pp. 624–631, Association for Computational Linguistics, Oct. 2018.

[16] O. Caglayan, M. García-Martínez, A. Bardet, W. Aransa, F. Bougares, and L. Barrault, "Nmtpy: A flexible toolkit for advanced neural machine translation systems," *The Prague Bulletin of Mathematical Linguistics*, vol. 109, p. 15–28, Oct 2017.

[17] D. Saunders and B. Byrne, "Reducing gender bias in neural machine translation as a domain adaptation problem," 2020.

[18] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," 2016.

[19] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014.

[20] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[21] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” 2014.

[22] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” 2015.

[23] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” 2012.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019.

[26] A. Radford, “Improving language understanding by generative pre-training,” 2018.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016.

[29] D. Elliott, S. Frank, K. Sima’an, and L. Specia, “Multi30k: Multilingual english-german image descriptions,” 2016.

[30] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *TACL*, vol. 2, pp. 67–78, 2014.

[31] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[33] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014.

[34] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” pp. 311–318, 2002.

[35] M. Denkowski and A. Lavie, “Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems,” in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, (Edinburgh, Scotland), pp. 85–91, Association for Computational Linguistics, July 2011.

[36] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A study of translation edit rate with targeted human annotation,” in *In Proceedings of Association for Machine Translation in the Americas*, pp. 223–231, 2006.

[37] J. H. Clark, C. Dyer, A. Lavie, and N. A. Smith, “Better hypothesis testing for statistical machine translation: Controlling for optimizer instability,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (Portland, Oregon, USA), pp. 176–181, Association for Computational Linguistics, June 2011.

[38] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia, “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), pp. 215–233, Association for Computational Linguistics, Sept. 2017.

[39] O. Caglayan, P. Madhyastha, L. Specia, and L. Barrault, “Probing the need for visual context in multimodal machine translation,” 2019.