Interpreting Cross-Attention in Transformer-Based Machine Translation

TARA SHUKLA and TIFFANY DEANE

ACM Reference Format:

Tara Shukla and Tiffany Deane. 2025. Interpreting Cross-Attention in Transformer-Based Machine Translation. 1, 1 (May 2025), 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In recent years, advancements in the field of machine translation (MT) have made great strides to facilitate more accurate and efficient cross-linguistic communication. Since its introduction in 2017, the Transformer model [8] has played a pivotal role in these rapid developments, revolutionizing the fields of neural machine translation (NMT) and natural language processing (NLP) as a whole. This Transformer architecture is at the core of several modern and widely used language translation software (e.g. Google Translate and DeepL) and extends its applications in text generation and question answering platforms (e.g. OpenAI's GPT-3 and BERT) as well. At the heart of the Transformer model is the attention mechanism, which, in the context of translation, allows it to determine and focus on the most relevant parts of an input sequence at each step of generating the output sequence. In essence, the attention mechanism aims to mimic human attention by enabling the model to view words in context and weigh their importance relative to one another. Attention mechanisms come in two forms: self-attention and cross-attention. In this project, we primarily investigate cross-attention, which allows models to relate words from an input sequence to words in an output sequence, effectively identifying connections between words in different languages.

Although transformer-based models have been instrumental in the rapid development of NMT, the complexity of their structures and of machine translation tasks have made their internal decision-making processes relatively unclear. This lack of transparency has made it difficult for humans to understand how exactly models might arrive at particular translations [11]. In order to establish trustworthiness and measure faithfulness of model attention to

Authors' Contact Information: Tara Shukla, ts6796@princeton.edu; Tiffany Deane, tdeane@princeton.edu.

Manuscript submitted to ACM

human word-to-word alignment, the ability to look under the hood is imperative. In this project, we seek to increase the interpretability of transformer-based models by visualizing cross-attention during model interference and observing patterns that arise across layers and heads. Our research aims to answer the following key research questions:

- How do cross-attention patterns compare to traditional word alignment?
- How do cross-attention patterns correspond to parts of speech?
- Do cross-attention weights correlate with dependency relations in predictable ways?

To explore these questions, we use a pre-trained transformer-based MT model, MarianMT, and run French-to-English sentence translation. After visualization, we then explore how attention patterns compare with those in classical aligners like `awesome_align`, and how they vary between linguistic structures, like parts of speech and dependency relations. In doing this, we aim to understand how transformer-based machine translation models make cross-linguistic connections, which could prove useful in more effective error diagnosis and model improvement.

2 Literature Review

Developed by Google Researchers in 2017, the Transformer model has now become the dominant architecture in NMT. In their now famous paper "Attention is All You Need," Vaswani et al. present an encoder-decoder structure using stacked self-attention and fully connected layers for the encoder and decoder. In essence, the encoder functions to map an input sequence of tokens in a source language to a sequence of contextual representations or "weights." It consists of N identical layers, each containing a multi-head self-attention mechanism sub-layer, and a feed-forward network sub-layer [8]. The multi-head self-attention mechanism assigns weights to each token of the input sequence according to the following attention function: $Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V$ where Q, K, V are vectors of queries, keys, and values, respectively. The queries and keys have dimension d. Each token in the input sequence has a distinct Q, K, and V, and the function compares Q with K to determine each words' relevance to another, such that higher attention weights indicate greater relevance to the query token. The decoder also consists of N identical layers and the same two sub-layers, but has an additional multi-head attention mechanism sub-layer [8]. This sub-layer is the cross-attention block, where the attention function uses the encoded input tokens as K and V, and the generated output tokens as Q. In short, the cross-function mechanism in the Transformer model allows the decoder to attend to each encoded token based on its relevance, or weight, in order to inform the generation of each token in the output sequence. Three years prior to the development of the Transformer, Bahdanau et al. first proposed the attention mechanism for use in RNN-based sequence-to-sequence models, and visualized attention weights to show that the model learned to align source and target words [1]. This research revolutionized the field of NMT, enabling models to focus only on

information relevant to the generation of the next target word, and thus generating better results when translating longer sentences. At this time, this led to the common belief that attention weights could be interpreted as direct word alignment, adding a degree of interpretability to the decision-making process. However, since the introduction of the attention mechanism, subsequent research has challenged its reliability. In their paper, "Six Challenges for Neural Machine Translation," Koehn and Knowles argue that though there is a need for an alignment mechanism between words in different languages, the attention model should not always be considered as such. They observed that the attention mechanisms may align words that do not correspond with human intuition or with alignment obtained by the traditional word alignment method, `fast_align` [6]. In response to the introduction of the attention-based Transformer model in 2017, Jain and Wallace (2019) published "Attention is not Explanation," in which they found that, although they contribute to improved NLP performance, attention weights largely do not provide meaningful explanations for model predictions. They observe that attention-based models may encode random interactions between individual tokens, suggesting that the relationship between attention and a model's output sequence is not as direct as previously thought [5]. In their study, "What does Attention in Neural Machine Translation Pay Attention to?" Ghader and Monz conclude that attention does not always correlate with traditional alignment, but is able to capture useful information other than direct word-to-word alignments [3]. Despite their reservations about the faithfulness of attention as alignment, Ghader and Monz observe that attention patterns do vary based on the POS tags of the target words, suggesting similarity with traditional human word-alignment methods. Higher correlations between nouns and the concentrated pattern of attention show that it may be useful in generating nouns [3]. Similarly, Voita et al. found that in multi-head self-attention, certain heads in the encoder consistently capture linguistic roles, such as subject-object relations [10]. Htut et al. more explicitly extracts dependency relations between words in the pre-trained transformer model, BERT, using the attention weights of each layer/head in order to investigate the extent to which they are accurately captured with respect to ground-truth Universal Dependency [4]. Their results suggest that in BERT, certain heads track specific dependency types. In our work, we will employ a similar method of dependency analysis. Raganato and Tiedemann investigate the information that is learned by attention through layer-wise analysis, similarly concluding that certain heads mark dependency relations, and that early layers capture more syntactic information, while higher layers align with more semantics [7]. In Vig and Belinkov's research, they interpret the focus of attention in transformer-based model GPT-2 in relation to dependency at different depths by visualizing attention for individual layers and heads [9]. In our study, we draw inspiration from their method of individual instance visualization in order to investigate the specialization of heads and evolution across layers in cross-attention.

3 Data and Resources

3.1 Data

For our data, we used a subset of the Helsinki-NLP/opus_books dataset on HuggingFace, specifically focusing on their French-to-English translation pairs. We sampled 100 French sentences from the corpus to run our corpus-level cross-attention analysis of the model. The Helsinki-NLP/opus_books dataset is part of a larger OPUS collection, which contains translated texts from different domains. We chose the books subset in order to ensure syntactic correctness and variation in sentence length, structure, and topic; this allows us to analyze cross-attention in varied linguistic contexts .

3.2 Tools and Resources

We utilized the Marian Neural Machine Translation framework via the HuggingFace Transformers library. Specifically, we employed the pretrained French-to-English model, Helsinki-NLP/opus-mt-fr-en. We utilized Google Colab with an A100 GPU.

Our additional analysis tools included Python packages such as NumPy, Pandas, Matplotlib, and Seaborn, which we used for data manipulation, statistical analysis, and creation of visualizations. In additino, we used NLTK and Spacy for linguistic analysis, particularly in tokenization, dependency parsing, alignment between tokenizers, and POS tagging. Our analysis of cross-lingual word alignment was done using awesome-align, an open-source machine translation alignment tool based on pre-trained multilingual BERT [2].

4 Approach and Implementation

4.1 Approach

Our research investigates the relationship between cross-attention in NMT, and different linguistic phenomena. We create an analysis pipeline to examine how attention patterns in transformer-based translation correlate with word alignment, part-of-speech, and dependency structure.

- **Word Alignment:** We analyze how cross-attention weights, averaged across heads and layers, correspond to word-level alignments between source and target languages. In doing so, we provide a high-level analysis of how attention mechanisms implicitly map source to target words. This is similar to the approach taken by Ghader and Monz in [3], but instead of conducting an entropy analysis of attention calculation, we count cross-attention as being 'focused' on a word if the word lies in the 75th percentile of cross-attention across all source tokens. This is an approach inspired by the Maximum Attention Weights (MAX) framework proposed in [4] by Htut et.al.

- **POS Correspondence:** Then, we analyze the relationship between cross-attention patterns in layers, and POS categories. We calculated the top three POS that each layer paid attention to most often during dataset inference. This parallels the layer-wise analysis used in [7] by Raganato and Tiedemann.

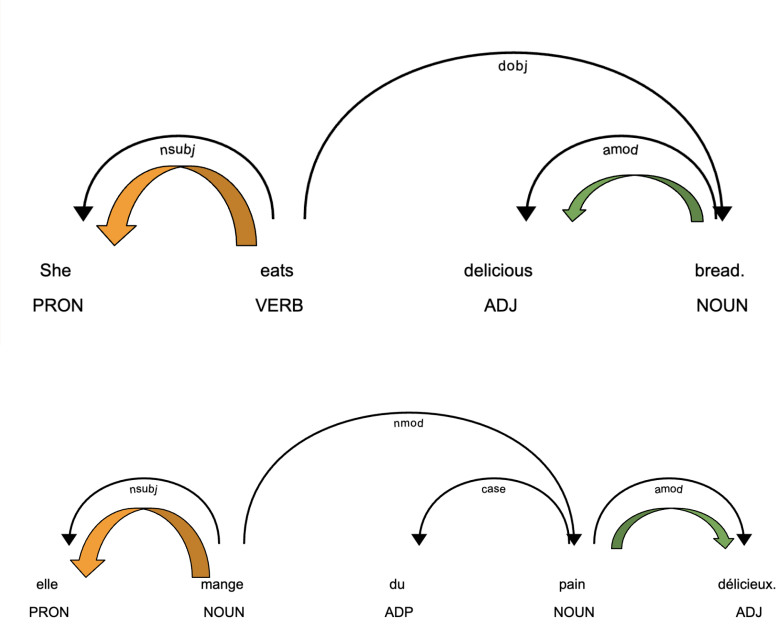


Fig. 1. Dependency parsing approach: visualized

- **Dependency Structure Transfer:** Finally, we study how cross-attention mechanisms handle the transfer of syntactic dependency structures from source to target language. To do so, we utilize cross-lingual word alignment to check if attention mechanisms for each head, averaged across layers, will focus on the aligned source word of the target word to which they are dependent. Referencing Figure 1, we see a simple example of parallel dependencies between aligned words. Our analysis seeks to recover the attention focus on, for instance, "Elle", when generating "eats". Our per-head approach here is inspired by findings in [7], in which the authors found that heads often marked syntactic dependency relations.

Overall, our multifaceted approach gives us insights into how NMT models capture and parse linguistic information in cross-attention; this has implications for informing architectural improvement strategies and interpretability methods.

4.2 Implementation

To implement our approach, we set up a pipeline that included data processing, model loading, inference, and the three analysis steps above. First, we focused on sentence-level visualizations of cross-attention patterns: this included a sentence-level alignment heatmap and line mappings of word-to-word attention patterns across layers and heads. Then, we conducted a corpus-level analysis with the OPUS books dataset described earlier. First, we run model inference to generate translations, preserving the cross-attention matrices from all decoder layers. Then we run `awesome-align` to get cross-lingual word alignments; this involved utilizing the `Spacy Alignment` module to shift between `Marian MT` tokenization used in inference, and the `spacy` tokenization used in `awesome_align`. We used `spacy` POS and dependency parsing to map attention weights to parts of speech and dependency.

5 Results and Discussion

5.1 Cross-Attention Across Heads and Layers

First, we conducted a higher-level analysis of cross-attention between source and target words in different heads and layers of the transformer model. In Figure 2, target (generated) words are on the left, and the words they pay attention to, the source words, are on the left; line weights correspond to proportion of attention given to the source token at that generation step. The holistic visualization in Figure 2 showcases broader patterns. For each head, lower layers seem to have many heads with diffused attention patterns, suggesting that they capture spread-out syntactical relationships and general patterns. Meanwhile, upper layers seem to display more refined and sparse attention patterns, perhaps focusing on semantics or meaningful context.

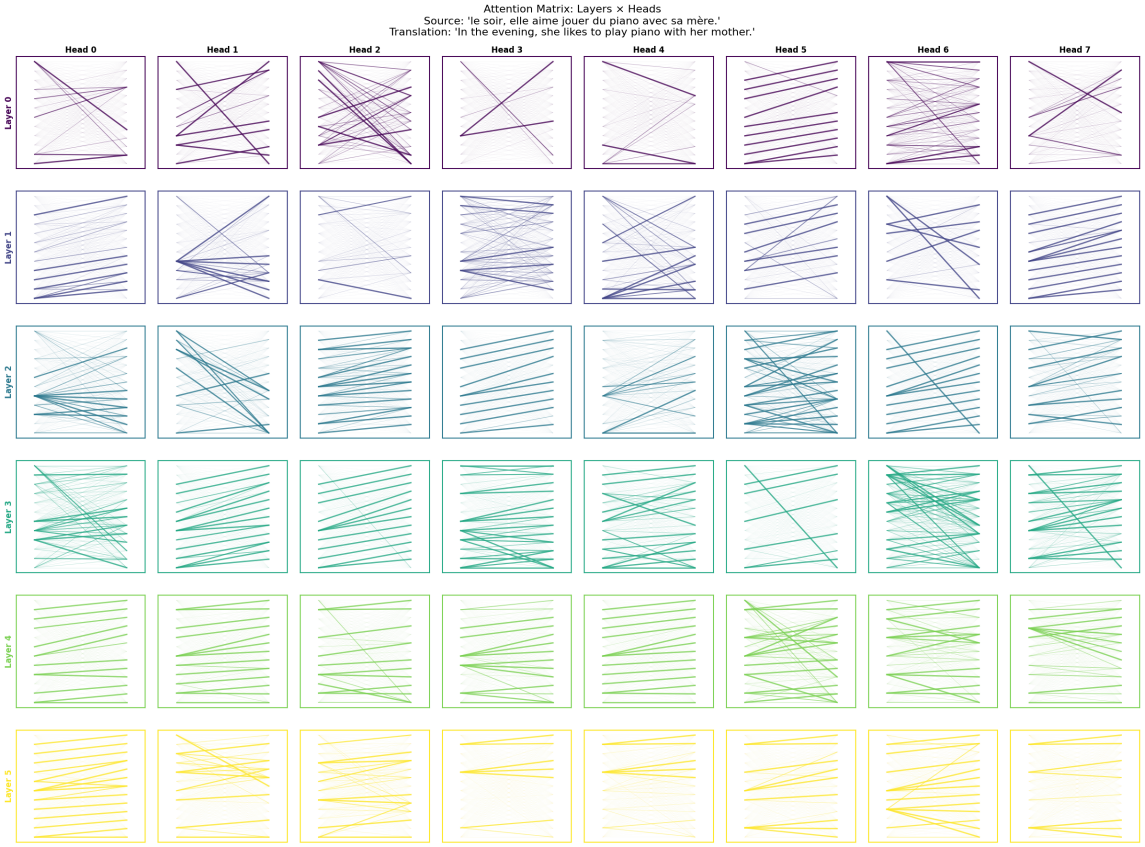
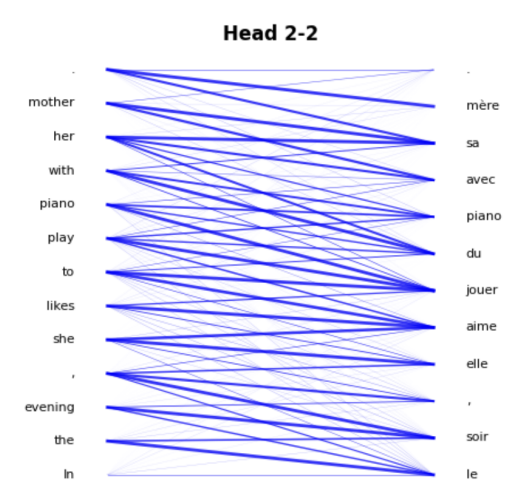
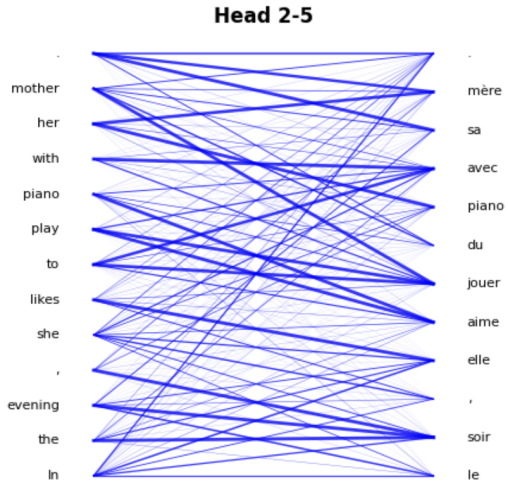


Fig. 2. Visualization of attention head layer architecture and information flow patterns

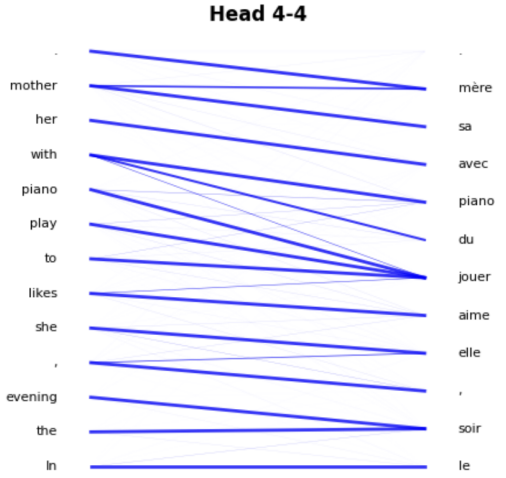
In Figure 3, we highlight specific layer-head attentions in more detail. We note that the labels are in the format `[Layer][Head]`. Thus, we see in lower heads there is attention to close-context words and broad attention across many tokens. For instance, in Figure 3a, attention is focused on close-context words in the source text. In Figure 3c, attention is spread out across local context. Meanwhile, in upper layers, such as in Figure 3b and 3d, attention is focused on the next token and contextual key words, suggesting that higher layers capture longer-range syntactic relationships as well as overall semantics and meaning. This is consistent with the findings in [7], and has implications for the interpretability of neural MT models. In the next sections, we move towards generalizing these cross-attention findings across a corpus.



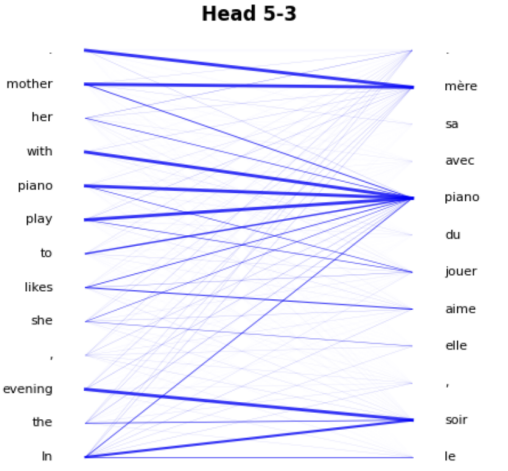
(a) Small context attention pattern



(b) Diffused attention pattern



(c) Next token attention pattern



(d) Topic-based attention pattern

Fig. 3. Visualization of cross-attention in specific transformer heads and layers

5.2 Cross-Attention and Alignment

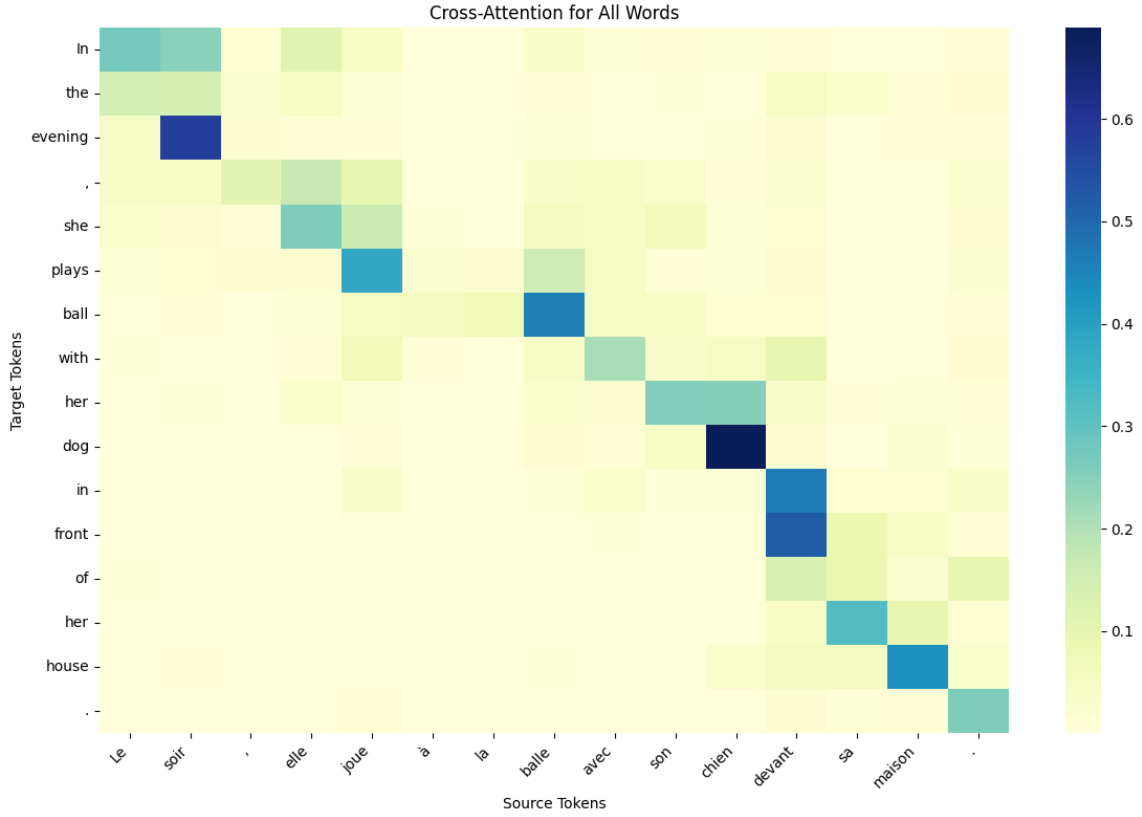


Fig. 4. Visualization of cross-attention alignment patterns across transformer architecture

Now, we look at the relationship between cross-attention and word alignment in neural machine translation. In 4, we see a heatmap of cross-attention averaged across all heads and layers for a simple sample sentence. The visualization demonstrates a predominantly diagonal pattern of high attention weight, indicating strong correlations between source and target tokens that are aligned translations of each other. The heatmap shows particularly strong attention for content words like "evening," "ball," "dog," "front," and "house," while function words exhibit more diffused attention patterns. This visualization supports our hypothesis that cross-attention mechanisms implicitly learn meaningful alignment information during translation. However, our corpus-level analysis reveals important limitations to this correspondence. Only 3% of our dataset shows exact matches between the highest attention weight and the linguistically determined alignment, suggesting that raw attention weights alone are insufficient for precise word alignment. More encouragingly, 40% of alignments fall within the 75th percentile of attention-paid words, indicating that cross-attention

Head	Most Attention-Paid POS
0	PUNCT AUX CCONJ
1	VERB CCONJ PUNCT
2	PUNCT CCONJ NOUN
3	CCONJ NOUN VERB
4	CCONJ NOUN VERB
5	NOUN CCONJ VERB
6	PUNCT NOUN VERB
7	CCONJ ADV NOUN

Table 1. Most-Attention-Paid POS Tags by Head, Averaged Across Layers

does capture alignment information but in a distributed rather than concentrated manner. These findings suggest that cross-attention serves as a useful but imperfect proxy for linguistic alignment, with attention being one of several factors influencing the model's translation decisions.

5.3 Cross-Attention and POS

Our analysis of attention distribution across different heads reveals clear patterns corresponding to part-of-speech tagging. In Table 1, we see that attention heads in the translation model develop distinct preferences for different specific POS tags. The table shows the top 3 POS that the heads pay attention to during inference of our whole dataset.

Notably, coordinating conjunctions (CCONJ) receive substantial attention across nearly all heads, suggesting a critical role in maintaining cross-lingual structural coherence. Punctuation (PUNCT) dominates in heads 0, 2, and 6, potentially serving as structural anchors that help align sentence boundaries and internal segmentation. Meanwhile, content words show interesting distributions, with nouns (NOUN) receiving significant attention in five heads (2, 3, 4, 5, and 6) and verbs (VERB) appearing prominently in four heads (1, 3, 4, and 5). This functional differentiation across heads suggests

Name	Dependency	Layer 0	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
amod	Adjectival Modifier	0.46	0.49	0.60	0.63	0.61	0.65
advmod	Adverbial Modifier	0.45	0.35	0.54	0.52	0.57	0.55
det	Determiner	0.45	0.51	0.62	0.64	0.67	0.67
dobj	Direct Object	0.30	0.29	0.34	0.35	0.28	0.30
nsubj	Nominal Subject	0.36	0.35	0.42	0.50	0.54	0.54
poss	Possession Modifier	0.36	0.42	0.62	0.58	0.58	0.60
prep	Prepositional Modifier	0.46	0.38	0.41	0.44	0.35	0.37
root	Root	0.37	0.44	0.57	0.58	0.68	0.65

Table 2. Percent of focused attention on dependency relation, by layer.

the model implicitly learns to decompose translation into linguistically motivated sub-operations. This specialization aligns with Vig, et.al's observation in [9], that "most tags are disproportionately targeted by one or more attention heads." We extend their work by replicating this finding in a translational setting via analysis of cross-attention, instead of self-attention.

5.4 Cross-Attention and Dependency

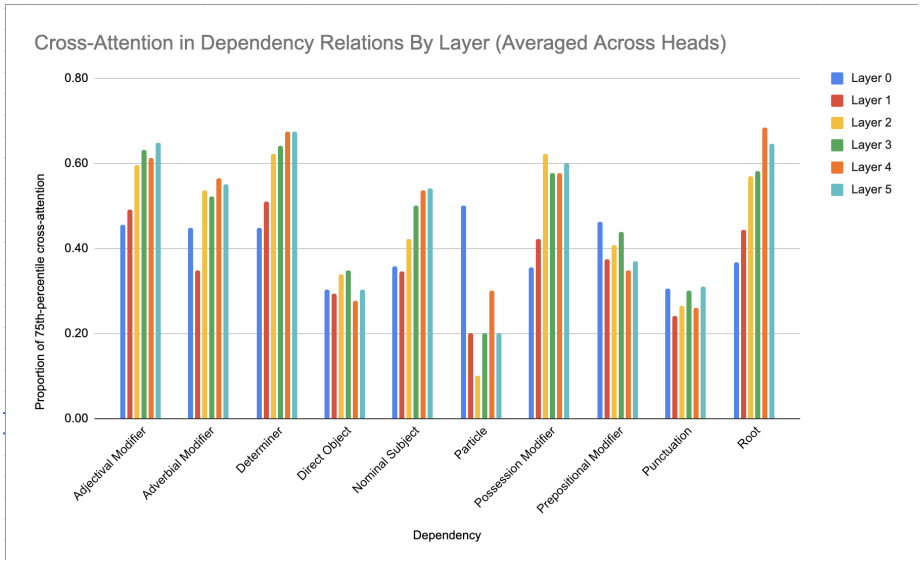


Fig. 5. Dependency parsing structure demonstrating syntactic relationships between words in a sentence

Finally, we conducted an analysis of how cross-attention correlates with dependency relationships: specifically those in which the target word is the dependent. In Table 2, we show the percentage of times in which a dependency relationship is given 75-th percentile attention weight during generation. We see that overall, dependency relationships are usually

given high attention, with the accuracy in common dependencies ranging from 0.3 to 0.67. A full version of this table is available in the Appendix.

For a more specific by-layer analysis, we look at Figure 5. We see that different layers have different levels of cross-attentional focus on various dependencies. In general, we see that for many dependencies, lower layers are outperformed by higher layers; the exclusions to that rule are particles, punctuation, and prepositional modifiers. We hypothesize that this is because these dependencies are usually close in the sentence to the cross-lingually aligned source token: in other words, they correlate well with positional attention.

Determiners, roots, and adjectival modifiers show consistently strong attention focus, reaching 0.67, 0.65, and 0.65 accuracies respectively in the highest layer. They also display a marked increase in attention from lower to higher layers, indicating the model progressively refines its understanding of syntactic structure.

In several cases, middle layers outperform the highest layers in paying cross-attention to dependencies: this pattern aligns with Vig et al.'s claim in [9] that "attention aligns with dependency relations most strongly in the middle layers," though our overall findings suggest that often higher layers will outperform lower ones.

This overall trend supports the hypothesis that lower layers establish basic positional relationships, while higher layers refine these into meaningful syntactic connections. Moving up layers, the model can transform positional attention into more linguistically-informed structural alignment. This layered specialization demonstrates how neural translation models progressively construct cross-lingual syntactic mappings.

5.5 Limitations

As much of our analysis involves comparing cross-attention based alignments with the traditional word aligner, `awesome_align`, it's important to acknowledge the limitations inherent in this baseline technique. `awesome_align`, like other statistical machine translation aligners, often struggles with nuanced language like idioms, figurative language, and sarcasm (Hovy et al., 2013). As it is reliant on pre-trained multilingual embeddings, limited or biased training data or fine-tuning corpora could lead to alignments that do not correctly reflect inter-lingual correlation. Due to subword tokenization strategies in models like BERT, in traditional models like `awesome_align`, whole words may be broken down into smaller parts. An additional step is needed to move from subword alignment to the whole word alignment, during which there is potential for the nuances of subword alignment to be lost as they are combined into whole words (Devlin et al., 2019). Because our visualizations show attention weights between whole source and target words, comparing these to converted words, where information may have been lost, has the potential to lead to misinterpretation. While `awesome_align` provides a useful benchmark of attention weights, even these are

not necessarily faithful to the nuance of human word-to-word alignments, and must be considered when drawing conclusions about the faithfulness of cross-attention alignments in respect to them.

In the similar vein, our analysis of dependency structure transfer in cross-attention mechanisms relies on cross-lingual word alignment to check that the attention focus in our model is representative of word dependency relations. Thus, the validity of our analysis is solely reliant on the accuracy of the word alignment tool we use as our benchmark, and if it contains any flaws, our investigation of whether attention corresponds to dependency will also be inaccurate.

Another potential limitation may lie in our method of considering attention "focused" or not. We count cross-attention as being "focused" on a word if it lies in the 75th percentile of cross-attention. This carries the risk of oversimplification, ignoring the continuous nature of attention distribution once this threshold is reached. When attention is distributed across several source words that each contribute to a translated target word, this categorization may fail to capture the nuance of alignment.

6 Conclusion and Future Work

In this project, we use the cross-attention mechanism intrinsic to the pre-trained transformer-based MT model, MarianMT in order to gain insight into learned cross-lingual alignments. Our generated visualizations of the decoder's cross-attentions weights revealed that the cross-attention mechanism serves as an implicit aligner between source and target words to a certain degree, and that different attention heads exhibit distinct alignment patterns. Specific heads tend to focus on semantically related words across the French and English sentence pairs, suggesting a degree of specialization within the multi-head attention architecture. Other heads, however, display more distributed attention, potentially suggesting broader contextual dependencies or syntactic relationships necessary for fluent translation. These findings suggest that some aspects of dependency parsing information is implicitly learned by cross-attention. Our initial observations suggest the potential of attention visualizations as a means of error diagnosis. By analyzing attention maps of the various target tokens, we can develop a deeper understanding of which source tokens the model deems most relevant for translation, which in turn may be beneficial for identifying instances where translation errors can be attributed to misalignment.

There are several directions in which our research may be expanded, including a deeper linguistic analysis of attention weights. This may involve analyzing attention weights based on finer-grained parts-of-speech, syntax, and semantics, and determining if specific heads or layers lean towards attending to certain linguistic features across languages. Future work might also include more quantitative analysis, such as devising metrics to measure focus and strength of attention, in order to make more statistical comparisons across heads, layers, and linguistic structures. It may also prove useful to extend our visualization to the translation of corpora within a specific domain, like the medical

or legal fields, for example, in order to observe how attention patterns are influenced by domain-specific vocabulary and syntax. Additionally, investigating attention patterns more thoroughly among decoder layers could provide beneficial insight and understanding into the hierarchy of the translation process.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

[3] Hamidreza Ghader and Christof Monz. 2017. What does Attention in Neural Machine Translation Pay Attention to?. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Greg Kondrak and Taro Watanabe (Eds.). Asian Federation of Natural Language Processing, Taipei, Taiwan, 30–39. https://aclanthology.org/I17-1004/

[4] Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do Attention Heads in BERT Track Syntactic Dependencies? (2019). arXiv:arXiv:1911.12246

[5] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).

[6] Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872* (2017).

[7] Alessandro Raganato and Jörg Tiedemann. 2018. An Analysis of Encoder Representations in Transformer-Based Machine Translation. (Nov. 2018), 287–297. https://doi.org/10.18653/v1/W18-5431

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (2017). arXiv:arXiv:1706.03762

[9] Jesse Vig and Yonatan Belinkov. 2019. Analyzing the Structure of Attention in a Transformer Language Model. (2019). arXiv:arXiv:1906.04284

[10] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418* (2019).

[11] Zijie Zhou, Junguo Zhu, and Weijiang Li. 2024. Towards understanding neural machine translation with attention heads' importance. *Appl. Sci. (Basel)* 14, 7 (March 2024), 2798.

7 Appendix

Dependency	Layer 0	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
Adjectival Clause	0.42	0.42	0.42	0.33	0.25	0.33
Adjectival Complement	0.50	0.42	0.50	0.50	0.58	0.58
Adjectival Modifier	0.46	0.49	0.60	0.63	0.61	0.65
Adverbial Clause Modifier	0.31	0.40	0.37	0.34	0.29	0.31
Adverbial Modifier	0.45	0.35	0.54	0.52	0.57	0.55
Attribute	0.43	0.43	0.43	0.29	0.14	0.43
Auxiliary	0.51	0.45	0.56	0.63	0.57	0.56
Case Marking	0.44	0.56	0.67	0.67	0.56	0.56
Clausal Complement	0.28	0.23	0.18	0.18	0.18	0.23
Clausal Subject	0.31	0.40	0.40	0.40	0.40	0.40
Compound	0.47	0.53	0.73	0.73	0.73	0.60
Compound	0.47	0.53	0.73	0.73	0.73	0.60
Conjunct	0.25	0.17	0.25	0.19	0.23	0.19
Coordinating Conjunction	0.33	0.23	0.27	0.27	0.27	0.21
Dative	0.60	0.40	0.60	0.40	0.60	0.20
Determiner	0.45	0.51	0.62	0.64	0.67	0.67
Direct Object	0.30	0.29	0.34	0.35	0.28	0.30
Marker	0.39	0.39	0.46	0.68	0.57	0.64
Negation Modifier	0.32	0.32	0.41	0.36	0.50	0.55
Nominal Subject	0.36	0.35	0.42	0.50	0.54	0.54
Noun Phrase as Adverbial Modifier	0.38	0.31	0.46	0.38	0.38	0.38
Numeric Modifier	0.36	0.73	0.82	0.82	0.73	0.82
Object of Preposition	0.41	0.37	0.42	0.41	0.41	0.38
Open Clausal Complement	0.45	0.45	0.50	0.45	0.27	0.32
Particle	0.50	0.20	0.10	0.20	0.30	0.20
Passive Auxiliary	0.00	0.33	1.00	0.67	1.00	1.00
Passive Nominal Subject	0.33	1.00	1.00	1.00	1.00	0.67
Possession Modifier	0.36	0.42	0.62	0.58	0.58	0.60
Predeterminer	0.50	0.50	0.50	0.50	0.50	0.00
Prepositional Complement	0.46	0.54	0.62	0.54	0.62	0.46
Prepositional Modifier	0.46	0.38	0.41	0.44	0.35	0.37
Punctuation	0.30	0.24	0.26	0.30	0.26	0.31
Quantifier Modifier	0.50	0.75	0.75	0.75	0.75	1.00
Relative Clause Modifier	0.30	0.33	0.41	0.30	0.33	0.33
Root	0.37	0.44	0.57	0.58	0.68	0.65

Table 3. Cross-attentions on dependency relationships; by layer. Full Table.