An Evaluation Framework for Explainability Approaches in Seq2Seq Machine Translation Models

Anonymous ACL submission

Abstract

The importance of input features in the decision-making of neural network models is a well-explored area of machine learning research. Numerous approaches have been developed to estimate and explain the behavior of these models. Among the models that rely on neural networks, the sequence-to-sequence (seq2seq) architecture is particularly complex. Although general techniques can be applied to these models, the evaluation of explainability methods in this context remains underexplored. In this paper, we propose a novel approach, based on forward simulatabitly, to automatically evaluate explainability methods for transformer-based seq2seq models. The idea is to inject the learned knowledge from a large model into a smaller one and measure the change in the results for the smaller model. We experiment with eight explainability methods using Inseq library to extract the attribution score of input to the output sequence. Then, we inject this information into the attention mechanism of an encoder-decoder transformer model for machine translation. Our results demonstrate that this framework can serve as an automatic evaluation method for explainability techniques and a knowledge distillation process that improves performance. According to our experiments, the attention attribution and value zeroing methods consistently increased the result in three machine translation tasks and composition operators.

1 Introduction

In recent years, natural language processing (NLP) generative models have been extensively developed and applied across a wide range of domains (Chang et al., 2024; Kalyan, 2024). At their core, these models rely on complex neural network architectures, such as the transformer architecture (Vaswani, 2017), which are often referred to as "black boxes" (Dayhoff and DeLeo, 2001; Burkart and Huber, 2021) due to their largely

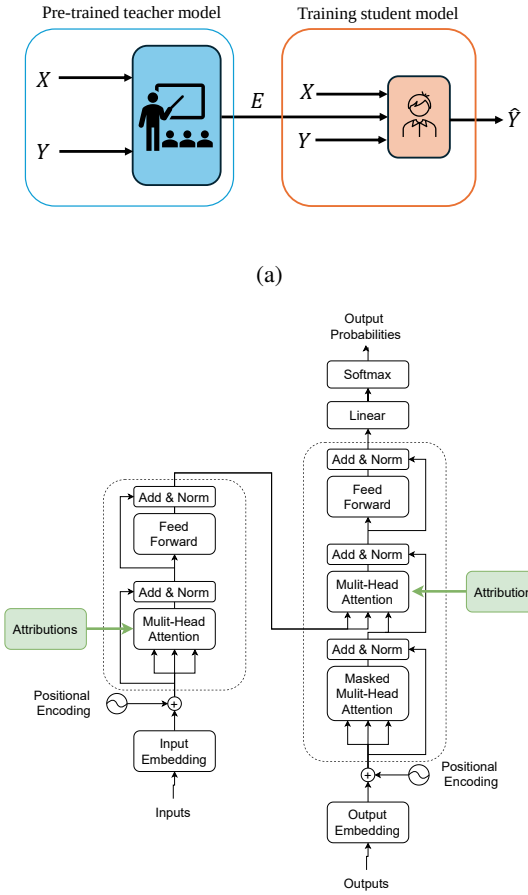


Figure 1: **(a)** shows the overall architecture of our approach. The input sequence and the gold output (X, Y) are given to a teacher model, and their attributions are obtained. Then, a new untrained model is trained using the same (X, Y, E) triples, where E represents the extracted attributions. **(b)** shows two places where we inject the attributions obtained from XAI methods. On the encoder side, the attributions are injected into all layers. In a separate experiment, the attributions are added only to the cross-attention residues.

opaque internal mechanisms. To address this challenge, Explainable AI (XAI) aims to improve the transparency and interpretability of machine learning models. A central objective of this approach is to assess the importance of input features or attributions (Arya et al., 2019; Vieira and Digiampietri, 2022; Saeed and Omlin, 2023), providing insights into how the model processes information or which contribute more to the decision-making of the model. Several of these methods have been developed specifically for NLP models to clarify the mechanisms behind their output classification or generation(Madsen et al., 2022b).

Yet determining which explanation accurately reflects the learned relationship between input and output remains an open question. Since explanations are intended for human interpretation, the validation of XAI methods has predominantly been human-centered (Kim et al., 2024). For the automatic evaluation of XAI methods, some approaches have been developed to assess them in other domains (Nauta et al., 2023). In image classification tasks, techniques such as removing important features identified by XAI methods and retraining the model based on the remaining characteristics (Hooker et al., 2019; Ribeiro et al., 2016a), as well as covering important features (Chang et al., 2018), have been explored. However, such approaches are less common in NLP (Madsen et al., 2022a), where human evaluation remains the dominant method (Madsen et al., 2022b; Leiter et al.). Furthermore, prior research has primarily focused on interpreting the attention mechanism (Moradi et al., 2021; Serrano and Smith, 2019) rather than comparing different explainability methods.

One of the main NLP domains is Sequence-to-sequence (seq2seq) models (Sutskever, 2014), which utilize an encoder-decoder framework and play a central role in neural machine translation, summarization, and dialogue systems. However, their many-to-many mapping, autoregressive nature, and intricate encoding-decoding process make them significantly more challenging to interpret than simpler classification models (Gurrapu et al., 2023). In this context, understanding the causal relationship between input and output tokens provides a framework for explanations of seq2seq models (Alvarez-Melis and Jaakkola, 2017). Typically, explanation methods for seq2seq models have relied on attention mechanisms and perturbation-based analysis to attribute importance to input tokens and assess their impact on model predictions (Moradi et al., 2021). However, attention is one part of this architecture, and the methods that take the whole system into account are less explored.

An important question to explore is whether human evaluations of XAI methods can be approximated by machines. One key approach in human evaluation is ***simulatability*** (Doshi-Velez and Kim, 2017; Hase and Bansal, 2020), which measures how well explanations enable humans to predict a model's behavior after receiving its explanation—a process known as forward simulation. Building on this concept, along with research on feature and attribute importance in explainability evaluation, we hypothesize that each explainability method provides unique information that differs from others. We conjecture that feeding this information to train new models can assess their effect and provide a framework for evaluating this XAI method. For this reason, in this work, we utilize attributions provided by XAI methods and inject them into the transformer architecture (Vaswani, 2017). We train and evaluate Opus-MT (Tiedemann and Thottingal, 2020) machine translation models by applying different matrix composition operations on the encoder and decoder self-attention and cross-attention mechanisms of these in the context of three machine translation datasets. By systematically evaluating various explainability techniques provided by Inseq (Sarti et al., 2023) as well as the four operations to merge the information with the attention, this study provides insights into the strengths and limitations of existing attribution methods.

To summarize, the main contributions of this work are as follows:

- We propose a novel framework for evaluating explainability methods in seq2seq models by injecting attribution-based information into the encoder-decoder architecture.
- We demonstrate that knowledge transfer from a pre-trained model to a model trained from scratch via feature attribution can improve performance, highlighting the role of explainability in model distillation.
- We conduct extensive experiments on multiple strategies for integrating explanations into the Transformer architecture and systematically compare their impact on model performance across different machine translation language pairs.

The findings contribute to the broader discourse on explainable NLP, emphasizing the need for evaluation metrics that balance faithfulness, utility, and alignment in seq2seq model explanations. Furthermore, the proposed knowledge injection framework is not limited to a specific model but can be applied to any encoder-decoder architecture, making it a flexible approach for improving interpretability and performance across various NLP tasks.

2 Related work

2.1 AI Explainability in Seq2seq Models

Explainability in Deep Learning Models. The rise of deep learning models has significantly improved performance in NLP tasks but has also raised concerns about their lack of interpretability (Burkart and Huber, 2021; Madsen et al., 2022b; Vieira and Digiampietri, 2022). Explainable AI (XAI) aims to make machine learning predictions more understandable to humans by providing insights into the decision-making process. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) have been widely adopted to attribute importance to input features (Lundberg and Lee, 2017; Ribeiro et al., 2016a), as well as counterfactual explanations (Chang et al., 2018). However, these methods were primarily designed for models with straightforward input-output mappings and may not directly extend to seq2seq architectures (Jain and Wallace, 2019; Serrano and Smith, 2019).

Challenges in Explaining Seq2seq Models. Seq2seq models, particularly those based on the transformer architecture (Vaswani, 2017), have revolutionized tasks such as machine translation and text summarization by capturing complex dependencies between input and output sequences (Stahlberg, 2020; Shakil et al., 2024). However, their encoder-decoder structure introduces several challenges for explainability methods (Zhao et al., 2024). For instance, **Intermediate representations** pose a challenge as the transformation of inputs through multiple layers makes it difficult to directly correlate input features with outputs (Sutskever, 2014). **Attention mechanisms** are often used to explain decisions in transformer models, but their reliability as faithful explanations has been questioned (Jain and Wallace, 2019; Madsen et al., 2022a). Additionally, **evaluation metrics** used in standard explainability methods may not fully capture the nuances of seq2seq models, necessitating specialized evaluation frameworks (Hase and Bansal, 2020; Nauta et al., 2023).

Advancements in Explainability for Seq2seq Models. To address some of these challenges, recent research has focused on developing explainability methods tailored to seq2seq models (Burkart and Huber, 2021; Zhao et al., 2024). For instance, **FiD-Ex** Framework, introduced by Lakhotia et al. (Lakhotia et al., 2021), improves the faithfulness of explanations in seq2seq models by incorporating sentence markers and fine-tuning on structured datasets. Furthermore, **Inseq** is a Python library that provides a comprehensive tool for analyzing and comparing different explainability methods for generative language models (Sarti et al., 2023). The library offers a range of gradient-based, perturbation-based, and internal representation-based explainability techniques. One of its key features is the ability to take an input-output pair, run it through a pre-trained model, and generate an importance analysis for the output sequence with respect to the input sequence. Recently, **SyntaxShap**, proposed by Amara et al. (Amara et al., 2024), extends Shapley values to account for syntactic dependencies in text generation, providing a syntax-aware interpretability method that improves the alignment between explanations and linguistic structures.

In thi work, we evaluate the explainability of seq2seq models in the context of machine translation tasks. Previous studies, such as (Li et al., 2020), have assessed explanation methods for machine translation using fidelity-based metrics, highlighting the limitations of traditional word alignment approaches. While these metrics provide insights into model behavior, they do not fully capture how explanations contribute to decision-making. To address this, we propose an automatic evaluation framework for explainability methods in transformer-based seq2seq models. Our approach involves injecting learned knowledge from a large model into a smaller one and assessing its impact on performance. Specifically, we first use a pre-trained language model, the Opus-MT model (Tiedemann and Thottingal, 2020), to generate explanations for input-output pairs in the machine translation task. We then train the same model from scratch, incorporating the extracted explainability-based knowledge into its encoder-decoder attention mechanism. By measuring performance changes, we evaluate how effectively these explanations in-

fluence model behavior. To systematically compare different XAI techniques, we use the **Inseq** library[1], applying the eight explainability methods described in the following subsection.

2.2 AI Explainability Methods

XAI methods fall into three main categories (Sarti et al., 2023): gradient-based, internal-based, and perturbation-based methods.

Gradient-based Methods: **Saliency** computes gradients of the output with respect to the input to generate a saliency map, highlighting influential features (Simonyan et al., 2013). **Input(x)Gradient**(IxG) quantifies feature importance by multiplying input values with their gradients, emphasizing features with large gradients (Simonyan et al., 2014). **Layer Gradient(X)Activation**(LGxA) extends Input(x)Gradient by considering the gradient of model output with respect to activations of a specific layer, weighted by those activations. **Integrated Gradients**(IG) computes attributions by integrating gradients along a path from a baseline input to the actual input, ensuring sensitivity and implementation invariance (Sundararajan et al., 2017). **DeepLIFT**(GSHAP) assigns importance scores by comparing a neuron's activation to a reference activation and propagating contribution scores through the network (Shrikumar et al., 2019). **Gradient SHAP** combines SHAP and Integrated Gradients, computing expected gradients from a baseline distribution to the actual input (Lundberg and Lee, 2017).

Internal-based Methods. **Attention** extracts encoder self-attention weights from the model's forward pass (Bahdanau et al., 2016; Vaswani, 2017), using only encoder attention weights.

Perturbation-based Methods. **Value Zeroing**(ValueZeroing) systematically zeros out token value vectors to analyze information integration beyond self-attention, considering contributions from feedforward networks in Transformer encoders (Mohebbi et al., 2023).

While XAI methods offer insights into seq2seq models, their effectiveness varies by task. Next, we present our methodology to systematically evaluate these techniques in machine translation using the **Inseq** library.

[1] https://github.com/inseq-team/inseq

3 Methodology

Inspired by forward simulation (Hase and Bansal, 2020), we design a pipeline to compare different explainability attributions based on their impact on system performance. To analyze the forward simulation of various XAI methods, we use a teacher-student model (Fig. 1a). In the first step, we use the **Inseq** library to extract input-output attributions using the eight explainability algorithms specified in Subsection 2.2. At this stage, the teacher model receives a source-target language pair (X, Y) as input, and the output of **Inseq** is a set of attributions $(X, Y) \rightarrow E$.

All of these attributions, except for the Attention-based ones, are extracted in the shape $e \in \mathbb{R}^{j \times k \times l}$, where j is the input sequence length, k is the output sequence length, and l is the hidden dimension of the model. We aggregate them along the last dimension by averaging the values, resulting in a final shape of $e \in \mathbb{R}^{j \times k}$. With a slight difference, the Attention attributions are extracted in the shape $e \in \mathbb{R}^{j \times k \times n \times h}$, where j and k are the same as before, n represents the number of layers (here, only on the encoder side, $n = 6$), and h is the number of attention heads (8 in this case). We then compute the average along both of the last two axes to obtain a final shape of $e \in \mathbb{R}^{j \times k}$. To handle negative values and normalize the attribution matrices, we apply the MinMaxScaler as follows:

$$\mathbf{e}'_{i,j} = \frac{\mathbf{e}_{i,j} - \min_i(\mathbf{e}_{:,j})}{\max_i(\mathbf{e}_{:,j}) - \min_i(\mathbf{e}_{:,j})}$$

Then, the input to the student model is the triple of $(X, Y, \mathbf{E}')$. In the next step, we apply four different operations on the attention heads as follows:

$$\tilde{\mathbf{A}}^{(h)} = \mathbf{A}^{(h)} + \mathbf{E}' \quad (1)$$

$$\tilde{\mathbf{A}}^{(h)} = \mathbf{A}^{(h)} \odot \mathbf{E}' \quad (2)$$

$$\tilde{\mathbf{A}}^{(h)} = \frac{\mathbf{A}^{(h)} + \mathbf{E}'}{2} \quad (3)$$

$$\tilde{\text{Attention}}(Q, K, V, \mathbf{E}') = f\left(\text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \mathbf{E}'\right) V \quad (4)$$

Where $\mathbf{E}' = \{e'_1, e'_2, \ldots, e'_b\}$ represents the explainability attributions based on the batch size

used for the model, and Q, K, and V are the query, key, and value matrices of the transformer model. f is one of the operators mentioned above in 1-3. The last operation replaces $\mathbf{E}'$ to completely substitute $\frac{QK^\top}{\sqrt{d_k}}$.

Now, we train the student Opus-MT model (Tiedemann and Thottingal, 2020) from scratch with two settings: 1) We apply one of the above-mentioned operators to all layers of the encoder block. 2) We apply the attributions to the cross-attention mechanism between the encoder and decoder blocks. The source and target language pairs remain the same as those used for the teacher model. However, we now infuse the attribution scores into the attention heads of either the encoder or decoder of the model, respectively (Fig. 1b).

4 Experimental Results

4.1 Evaluation Datasets and Metrics

To evaluate the proposed pipeline, we train the Opus-MT model (Tiedemann and Thottingal, 2020) on three datasets. We select two datasets from the same language family: German→English (de-en) and French→English (fr-en). For the third dataset, we choose Arabic→English (ar-en) due to its encoding and linguistic differences from the target language. For de-en and fr-en, we use the WMT14 dataset (Bojar et al., 2014), and for ar-en, we use the UN Parallel Corpus (Ziemski et al., 2016).

We select $200,000$ sentence pairs from each dataset and preprocess them to suit our experimental setup. Given the large number of seq2seq models we train from scratch, we impose constraints to efficiently manage the training process. Specifically, we limit both input and output sequences to a maximum of 128 tokens. Additionally, we discard samples with fewer than three tokens and filter out pairs where the input-to-output length ratio (or vice versa) exceeds 1.7. For the German→English (de-en) and French→English (fr-en) datasets, we further exclude samples with an excessively high normalized Levenshtein distance. Since the validation and test sets of the WMT datasets are relatively small, we select an additional 15,000 samples from their training sets (without overlap with our training data). The UN Parallel Corpus does not include separate validation and test sets, so we extract $15,000$ samples from the main dataset for this purpose.

Throughout our experiments, we use the implementation of BLEU score (Papineni et al., 2002) to evaluate the student models.

4.2 Experimental Settings

We train the Opus-MT models[2] for 20 epochs and apply an early stopping of three consecutive epochs without improvement in validation loss. The model follows an encoder-decoder architecture, with each containing six layers with eight attention heads. The model employs the Swish activation function, as proposed by Ramachandran et al. (2017) (Ramachandran et al., 2017), which has been shown to enhance training dynamics and convergence compared to traditional activation functions like ReLU. To handle preprocessed data and manage training time, we limit the input and output sequence lengths to a maximum of 128 tokens. For training the models, we utilized 20 Nvidia V100 GPUs.

4.3 Results and Discussion

In our analysis, we evaluate the proposed methods through three key comparisons. First, we assess eight XAI methods for their effectiveness in improving translation quality when their attributions are injected into the model. Second, we compare the impact of injecting attributions into encoder self-attention versus cross-attention layer to understand their influence on information flow and source-target alignment. Lastly, we examine the effect of reducing the number of attention heads from eight to four, exploring whether selective attribution enhances model efficiency while maintaining performance. As a baseline, we report the results of training the student model without attribution injection on the three datasets. Table 1 presents the BLEU scores for training the model from scratch for each language pair.

	de-en	fr-en	ar-en
Baseline	22.85	28.79	16.85

Table 1: baseline BLEU score results

Comparison of XAI Methods – This analysis evaluates eight different explainability methods in terms of their impact on translation quality. Table 2 shows the result of the injection of the attribution score to the 8 attention heads of the encoder vs. cross attention part of the Opus-MT model trained from scratch. Across all three language pairs, attention-based and perturbation-based Attention and ValueZeroing tend to have the highest values among the attribution methods. Results suggest that these two mechanisms capture a strong

[2]https://huggingface.co/Helsinki-NLP

de-en Encoder	IxG	Saliency	LGxA	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
add	**23.10**	**27.36**	**23.10**	**27.68**	**23.17**	**23.26**	**31.58**	**33.12**
multiply	21.59	**27.98**	21.85	**27.75**	21.65	21.73	**35.08**	**35.01**
average	**23.18**	**26.65**	**23.18**	**26.51**	**22.90**	**22.99**	**30.47**	**32.27**
replace	21.78	**26.84**	21.75	**26.31**	21.68	21.68	**31.57**	**33.39**
de-en CrossAttention	**IxG**	**Saliency**	**LGxA**	**IG**	**GSHAP**	**DeepLIFT**	**Attention**	**ValueZeroing**
add	22.50	16.82	22.49	19.41	22.83	22.54	14.21	11.99
multiply	7.40	7.57	4.69	8.18	10.32	8.76	3.14	2.27
average	20.06	19.42	20.01	19.72	20.63	**22.87**	14.89	14.96
replace	0.25	0.08	0.21	0.04	0.25	0.38	4.69	0.25

fr-en Encoder	IxG	Saliency	LGxA	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
add	**29.04**	**36.99**	**29.04**	**35.52**	**30.14**	**29.04**	**44.16**	**46.97**
multiply	**29.29**	**38.54**	**29.30**	**35.94**	**29.66**	**28.88**	**49.31**	**49.14**
average	28.68	**36.31**	28.68	**34.16**	**29.55**	**28.84**	**42.62**	**45.43**
replace	28.15	**36.15**	28.15	**34.42**	**29.26**	28.26	**42.77**	**45.35**
fr-en CrossAttention	**IxG**	**Saliency**	**LGxA**	**IG**	**GSHAP**	**DeepLIFT**	**Attention**	**ValueZeroing**
add	24.31	26.50	24.32	23.78	26.53	24.59	24.50	22.25
multiply	14.69	3.66	14.68	6.29	7.49	15.86	5.82	1.62
average	22.12	23.50	28.76	28.76	24.40	20.55	25.75	26.12
replace	0.77	0.06	0.77	0.01	0.70	1.60	5.89	1.63

ar-en Encoder	IxG	Saliency	LGxA	IG	GSHAP	DeepLIFT	Attention	ValueZeroing
add	**32.60**	**38.72**	**32.60**	**30.75**	**33.91**	**33.59**	**46.46**	**46.29**
multiply	**37.06**	**43.74**	**36.78**	**40.28**	**37.59**	**37.03**	**51.53**	**51.64**
average	**27.70**	**34.14**	**27.70**	**30.55**	**28.75**	**26.56**	**46.69**	**41.17**
replace	**36.76**	**43.4**	**36.69**	**40.17**	**37.75**	**36.81**	**49.77**	**51.48**
de-en CrossAttention	**IxG**	**Saliency**	**LGxA**	**IG**	**GSHAP**	**DeepLIFT**	**Attention**	**ValueZeroing**
add	**33.87**	**31.31**	**33.87**	**29.24**	**34.94**	**33.61**	**26.28**	**29.61**
multiply	**17.40**	5.81	**17.41**	12.52	**22.63**	**17.41**	6.17	1.72
average	**18.47**	10.32	**18.47**	12.94	14.38	**18.10**	11.05	13.08
replace	1.81	0.25	1.85	0.01	0.97	1.78	9.56	5.48

Table 2: BLEU score comparison of various attribution methods across different composition strategies (add, multiply, average, replace) to 8 heads applied to encoder and cross-attention modules in neural machine translation models. Results are reported for three language pairs—German–English (de–en), French–English (fr–en), and Arabic–English (ar–en)—with columns corresponding to attribution techniques (IxG, Saliency, LGxA, IG, GSHAP, DeepLIFT, Attention, and ValueZeroing). Scores that beat the baseline model for each setting are boldfaced and the highest BLEU score for each dataset are highlighted in green.

signal relevant to the translation process. Gradient-based methods (IxG, LGxA, IG) generally yield lower scores compared to attention-based methods, but they exhibit consistency across different language pairs, making them more stable for interoperability. DeepLIFT except for the addition operation, decreases the result for de-en and fr-en. French-English (fr-en) consistently exhibits higher attribution scores than German-English (de-en), while Arabic-English (ar-en) shows the highest scores overall across most methods. ValueZeroing and Attention attribution help to double the BLEU score of this language pair. This finding may indicate that the morphological and syntactic complexity of the source language influences the attributions and hence help the result of the student model.

German-English (de-en) Attribution scores are generally the lowest among the three language pairs. This is likely due to the high word reordering requirements in German, which may lead to weaker local alignment between input tokens and model outputs (Avramidis et al., 2019; Macketanz et al., 2021). French-English (fr-en) Attribution scores are higher than de-en, suggesting that French and English have more direct word alignment, which leads to stronger feature attributions. This aligns with linguistic expectations and empiri-

cal evidence (Legrand et al., 2016), as French and English share more lexical and syntactic similarities. Arabic-English (ar-en) This pair exhibits the highest attribution scores, particularly in Attention (46.46–51.53) and ValueZeroing (46.29–51.64). It is possible that Arabic's rich morphology and non-concatenative structure likely cause the model to rely more heavily on attention mechanisms, explaining the higher attribution values.

LGxA and IG perform similarly across all language pairs, suggesting that deeper network activations contribute significantly to attribution results. This reinforces the role of deep-layer interactions in determining translation outputs. Gradient SHAP (GSHAP) and DeepLIFT methods display relatively consistent scores across language pairs, implying that their reliance on perturbation-based techniques makes them less sensitive to specific linguistic properties of the source language.

Overall, Attention and ValueZeroing tend to contribute to the highest scores among attribution methods across all three language pairs. The results suggest these two mechanisms capture a strong signal relevant to the translation process. Gradient-based methods (IxG, LGxA, IG) generally yield lower scores than the other two methods but exhibit consistency across different language pairs, making them more stable for interpretability. French-English (fr-en) consistently exhibits higher attribution scores than German-English (de-en), while Arabic-English (ar-en) shows the highest overall scores across most methods. It may indicate that the morphological and syntactic complexity of the source language influences attributions.

Encoder Self-Attention vs. Cross-Attention – This analysis examines the impact of injecting attribution scores into encoder self-attention layers versus cross-attention layers. In contrast to the encoder self-attention, cross-attention bridges the source and target languages by guiding the decoder's focus on the encoder's output. This mechanism is more sensitive because it manages the alignment between the source input and the target output. Any modification here can directly influence how the source information is integrated into the target generation process. For this reason, the initial hypothesis was that injecting attributes—which describe the relation between the source and target—into the cross-attention layer might enhance the flow of relevant information. However, the experimental results tell a different story.

In most cases, injecting these attributes into cross-attention either blocks or corrupts the flow of information. For example, when we replace the cross-attention weights entirely with the attribute values (using the "replace" operator), the performance degrades drastically to the point where the model essentially fails to learn anything. This suggests that the carefully learned cross-attention weights are critical for proper alignment and that overriding them with attribution values disrupts the fine-grained balance necessary for effective translation.

For addition ($+$) vs. multiplication ($\times$) an interesting trend observed is that the addition operator tends to yield better results than multiplication in the cross-attention context. Adding the attributions seems to augment the existing attention values in a beneficial way, whereas multiplying them often leads to an overly aggressive modification that harms the model's ability to propagate information from the encoder. This sensitivity is particularly pronounced in cross-attention layers. The addition might act as a mild corrective signal that helps the decoder focus better, while multiplication can excessively amplify or diminish the weights, leading to a loss of critical alignment information. The encoder self-attention layers appear to benefit from certain attribution methods (like Attention and ValueZeroing). This suggests that these methods may leverage the rich input representations directly, reinforcing the existing intra-sentence relationships without destabilizing them.

In contrast, the effectiveness of the cross-attention layer is based on an interaction between the representations of the source and the target. Methods like Attention and ValueZeroing, which work well in self-attention due to their direct reliance on the input, do not provide the same improvement when applied to cross-attention. This could be because the cross-attention mechanism requires a more nuanced handling of the inter-sequence relationship—one that the raw input-level attributes cannot fully capture

Effect of Attention Head Reduction (8 Heads vs. 4 Heads) – This analysis investigates how reducing the number of attention heads affects model performance when integrating attribution scores. The experiment of reducing the number of attention heads from eight to four provides valuable insights into the impact of different attribution methods and composition operations on translation performance. By selectively applying attributions to only four heads, we assess whether information flow

can still be effectively captured and whether the model retains its translation quality. The changes in BLEU scores between 8-head and 4-head settings for gradient-based methods are relatively minor. Some methods show slight improvements, while others show slight degradation. This suggests that reducing the number of attention heads does not drastically alter how gradient-based attributions affect the model. A significant jump in BLEU score is observed when applying attributions to only four heads instead of eight. This suggests that focusing attribution influence on fewer attention heads helps refine the attention mechanism, possibly by preventing redundant information flow and reinforcing key attention patterns. Similar to attention modifications, value zeroing shows a major increase in BLEU score when applied to only four heads. This result indicates that reducing the number of modified attention heads all.

For the additive and multiplicative operations, the BLEU score remains relatively stable across 8-head and 4-head settings. Figures 3 and 3 show the result of this comparison on at each place of the attribution injection and datasets. The replacement operation (where the attention values are fully substituted) leads to a performance drop in the 4-head setting, implying that completely overwriting attention information is more detrimental when fewer heads are active. The averaging operation also shows mild degradation in the 4-head setup, suggesting that blending information across fewer heads may not be as effective.

5 Conclusion

In this work, we explore the integration of attribution-based explanations into neural machine translation models, with the aim of evaluating interpretability and translation quality. Our extensive analysis across German–English, French–English, and Arabic–English language pairs included comparing eight XAI methods and various composition strategies—addition, multiplication, averaging, and replacement—for injecting attribution scores into the Transformer's attention mechanisms.

Effectiveness of Attribution Methods can be summarized as: Attention-based and perturbation-based techniques, specifically the Attention and ValueZeroing methods, consistently showed the highest improvements in BLEU scores. In contrast, while gradient-based approaches (e.g., IxG, LGxA, IG) provided stable and consistent attributions, their impact on performance was comparatively modest.

Injecting attribution scores into encoder self-attention layers generally reinforced intra-pairs relationships and improved translation quality. Conversely, modifications in the cross-attention layers, which govern source–target alignment, often disrupted the balance required for effective translation, mainly when we implemented replacement or multiplicative operations. Finally, reducing the number of attention heads from eight to four demonstrated that selective attribution can refine the attention mechanism, mitigate redundancy, and, in some cases, further enhance performance. These results suggest that attribution-based interventions not only serve as valuable tools for model interpretability but can also be leveraged to improve translation outcomes, especially in linguistically challenging scenarios.

Limitations

There are some limitations to this work worth noting. First, we compared attribution information across explainability methods, the majority of which were gradient-based. This choice was primarily due to the computational cost associated with extracting attributions using other methods, particularly perturbation-based approaches such as LIME (Ribeiro et al., 2016b) and reAGent (Zhao and Shan, 2024), which are both time-consuming and resource-intensive. Finally, some of these methods, provided by Inseq, generate attributions for the decoder side of seq2seq models. However, at this stage, we limited our experiments to encoder self-attention and cross-attention between the decoder and encoder.

Another limitation of our study is that our experiments were confined to machine translation and a limited set of language pairs. Future work should extend this pipeline to other sequence-to-sequence models, such as those used for question answering, by integrating it into the Transformer's attention mechanism, thereby enabling a broader evaluation of XAI methods.

Finally, we limited our experiments to a single evaluation metric, the BLEU score. While BLEU provides a quantitative measure of similarity between generated and reference sequences, it does not fully capture semantic adequacy or the overall quality of the generated text. Incorporating additional evaluation metrics, such as those assessing

semantic similarity (e.g., METEOR, BERTScore) or human evaluations based on fluency, coherence, and relevance, could offer deeper insights and a more comprehensive assessment of model performance. Future studies should explore these alternative metrics to ensure a more nuanced evaluation of generated sequences.

References

David Alvarez-Melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943*.

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. 2024. SyntaxShap: Syntax-aware explainability method for text generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4551–4566, Bangkok, Thailand. Association for Computational Linguistics.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. 2018. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Judith E Dayhoff and James M DeLeo. 2001. Artificial neural networks: opening the black box. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 91(S8):1615–1635.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A Batarseh. 2023. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*, 6:1225093.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Katikapalli Subramanyam Kalyan. 2024. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048.

Jenia Kim, Henry Maathuis, and Danielle Sent. 2024. Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence*, 7:1456486.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Scott Yih, Yashar Mehdad, and Srini Iyer. 2021. FiD-ex: Improving sequence-to-sequence models for extractive rationale generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3712–3727, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joël Legrand, Michael Auli, and Ronan Collobert. 2016. Neural network-based word alignment through score aggregation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 66–73, Berlin, Germany. Association for Computational Linguistics.

C Leiter, P Lertvittayakumjorn, M Fomicheva, W Zhao, Y Gao, and S Eger. Towards explainable evaluation metrics for natural language generation (2022). *Preprint at https://arxiv. org/abs/2203.11131*.

Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi. 2020. Evaluating explanation methods for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 365–375, Online. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for german to english and english to german. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2022a. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1731–1751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022b. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, Online. Association for Computational Linguistics.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016a. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016b. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Waddah Saeed and Christian Omlin. 2023. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Hassan Shakil, Ahmad Farooq, and Jugal Kalita. 2024. Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, page 128255.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. Learning important features through propagating activation differences. *Preprint*, arXiv:1704.02685.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Preprint*, arXiv:1312.6034.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of*

the European Association for Machine Translation, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

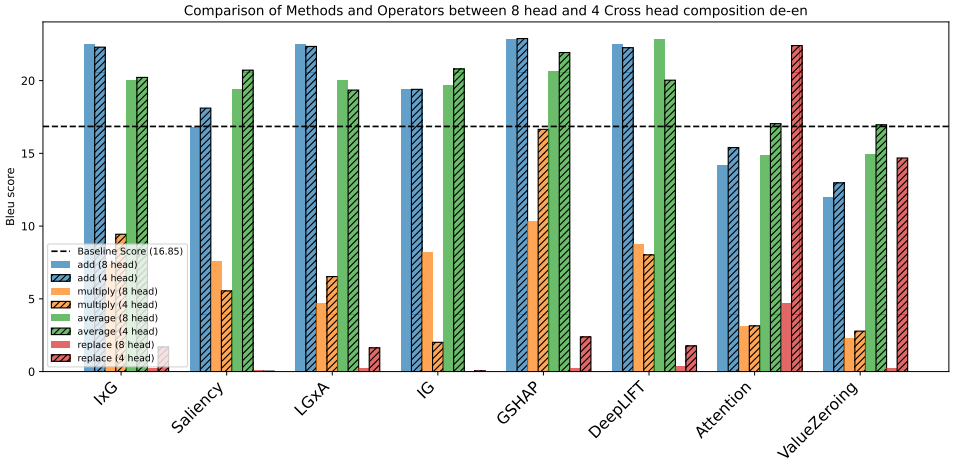
Carla Piazzon Vieira and Luciano Antonio Digiampietri. 2022. Machine learning post-hoc interpretability: A systematic mapping study. In *Proceedings of the XVIII Brazilian Symposium on Information Systems*, pages 1–8.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

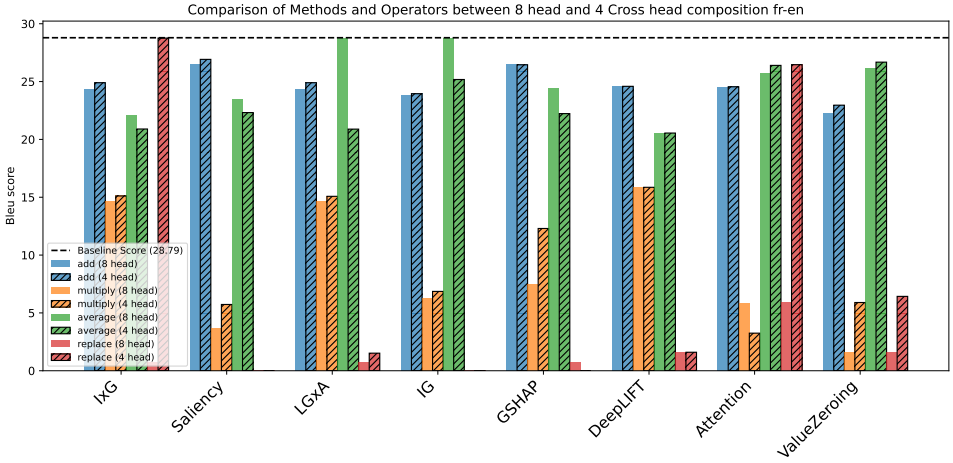
Zhixue Zhao and Boxuan Shan. 2024. Reagent: A model-agnostic feature attribution method for generative language models. *Preprint*, arXiv:2402.00794.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

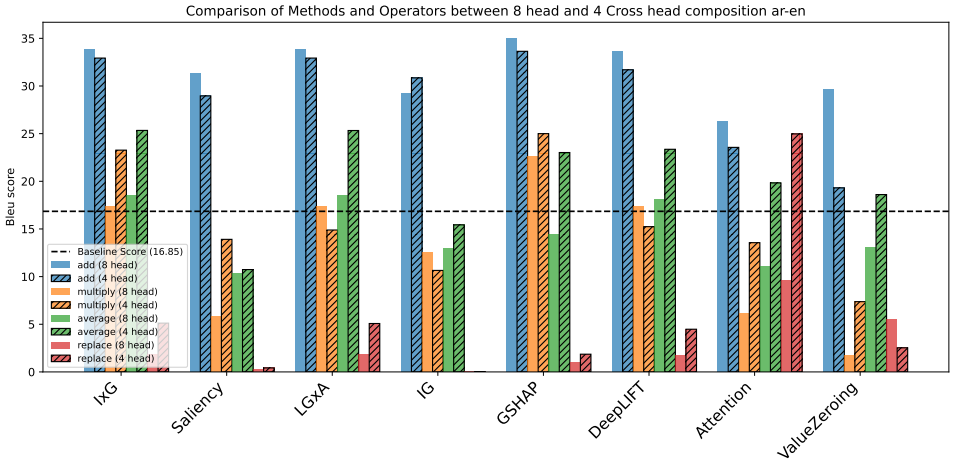
6 Appendix



(a) de-en

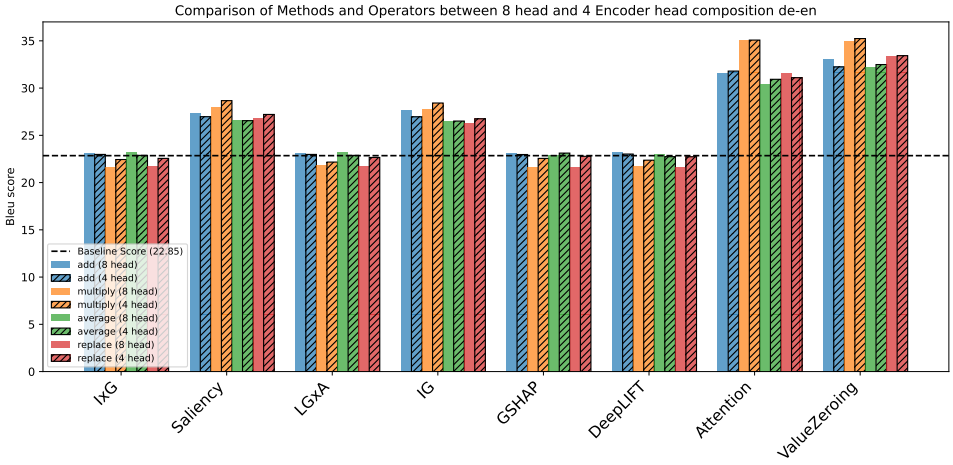


(b) fr-en

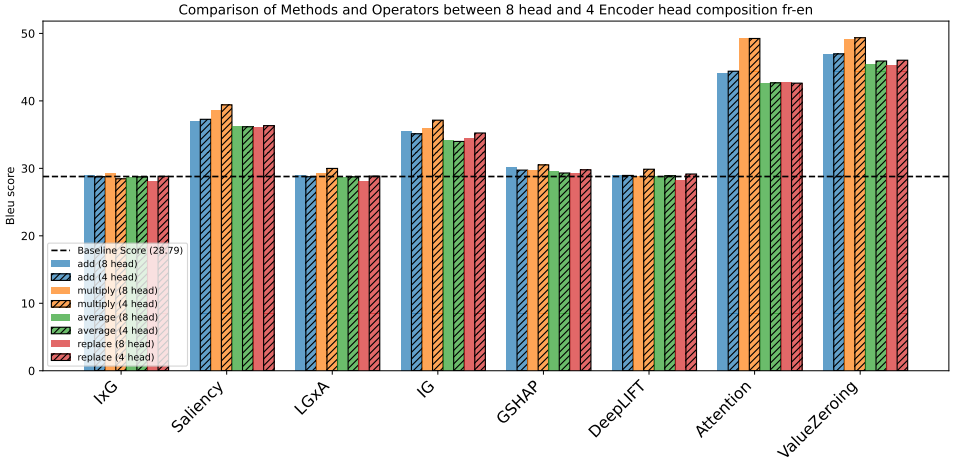


(c) ar-en

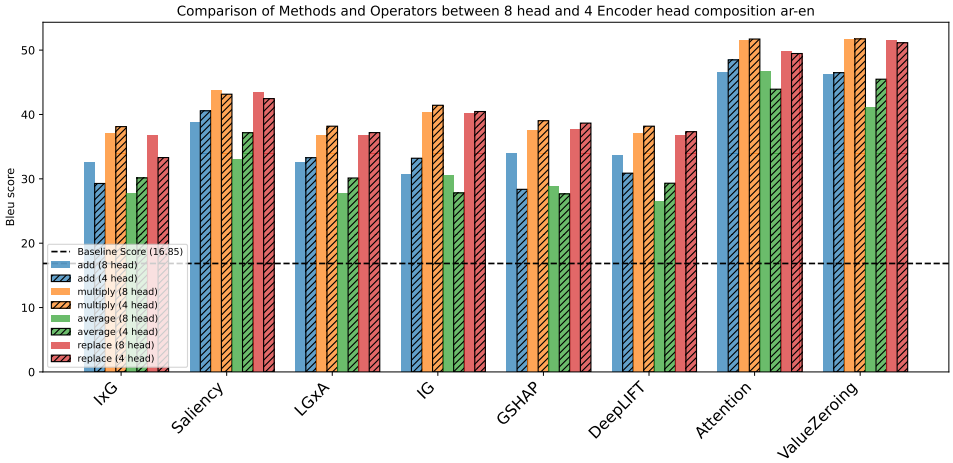
Figure 3: Result of Attribution Composition of Cross-Attention Weights on the de-en (a), fr-en(b) and ar-en(c) Datasets: Comparing 4-Head (Striped Bars) and 8-Head (Plain Bars) Configurations.



(a) de-en



(b) fr-en



(c) ar-en

Figure 4: Result of Attribution Composition of Encoder Self-Attention Weights on the de-en (a), fr-en(b) and ar-en(c) Datasets: Comparing 4-Head (Striped Bars) and 8-Head (Plain Bars) Configurations.