Measuring and Improving Faithfulness of Attention in Neural Machine Translation

Pooya Moradi, Nishant Kambhatla, Anoop Sarkar
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
{pmoradi, nkambhat, anoop}@sfu.ca

Abstract

While the attention heatmaps produced by neural machine translation (NMT) models seem insightful, there is little evidence that they reflect a model's true internal reasoning. We provide a measure of faithfulness for NMT based on a variety of stress tests where attention weights which are crucial for prediction are perturbed and the model should alter its predictions if the learned weights are a faithful explanation of the predictions. We show that our proposed faithfulness measure for NMT models can be improved using a novel differentiable objective that rewards faithful behaviour by the model through probability divergence. Our experimental results on multiple language pairs show that our objective function is effective in increasing faithfulness and can lead to a useful analysis of NMT model behaviour and more trustworthy attention heatmaps. Our proposed objective improves faithfulness without reducing the translation quality and has a useful regularization effect on the NMT model and even improve translation quality in some cases.

1 Introduction

How trustworthy are our neural models? This question has led to a wide variety of contemporary NLP research focusing on (a) different axes of interpretability including *plausibility* (or interchangeably *human-interpretability*) (Herman, 2017; Lage et al., 2019) and *faithfulness* (Lipton, 2018; Jacovi and Goldberg, 2020b), (b) interpretation of the neural model components (Belinkov et al., 2017; Dalvi et al., 2017; Vig and Belinkov, 2019), (c) explaining the decisions made by neural models to humans (using explanations, highlights, rationales, etc.) (Ribeiro et al., 2016; Li et al., 2016; Ding et al., 2017; Ghaeini et al., 2018; Bastings et al., 2019; Jain et al., 2020), and (d) evaluating different explanation methods from different perspectives (Samek et al., 2016; Mohseni and Ragan, 2018; Poerner et al., 2018; Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Li et al., 2020).

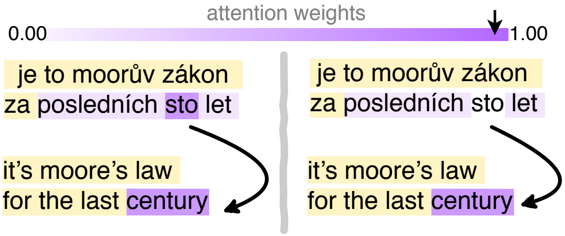


Figure 1: An example of *unfaithful* attention weights produced during a Cs-En translation. Note in the left attention heatmap, the attention is on the word sto while the decoder generates century. However, in the right heatmap, sto is not attended to at all but century is still produced as the output. This shows unfaithful behavior.

All of these approaches make NLP neural models more trustworthy. In this work, we focus on faithfulness which intuitively provides the extent to which an explanation accurately represents the true reasoning behind a prediction. It is particularly important for NLP practitioners who wish to debug their neural models and improve them. Faults of a neural model cannot be identified if the neural model does not provide a faithful and trustworthy description of what it is doing.

However, the formal definition of faithfulness and the proper approach to its evaluation are still contended in the literature. Jacovi and Goldberg (2020b) emphasize distinguishing faithfulness from human-interpretability in interpretability research by providing several clarifications about the terminology used by researchers. They describe the following conditions on the evaluation of how well a research project tackles the notion of faithfulness:

- Be explicit: provide a measurable evaluation of faithfulness.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2791–2802
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

- Human judgements are not relevant because we are interested in model internals.
- Do not match against gold labels (e.g. AER) because faithfulness of both correct and incorrect decisions made by the model are equally important.
- No model is "inherently" faithful. We need to measure faithfulness not as a binary aspect of a model (it is faithful or not) but rather as a gray-scale measure.
- A more faithful system is a necessary but not sufficient condition for model interpretation by humans, c.f. Jacovi and Goldberg (2020a).

Aligned with these criteria, we study faithfulness of NLP neural models, specifically NMT models. We provide a faithfulness measure that is computed based on a variety of stress tests where attention weights that are crucial for prediction are perturbed. We expect from a faithful model to change its prediction under such tests (Figure 1). We quantify faithfulness based on how often the model outputs changed. The proposed metric is defined based on discrete changes in the output. It is not differentiable and cannot be simply included in the loss function of NMT to be optimized. We propose a novel differentiable objective based on probability divergence and study its effect on the discrete faithfulness measure. Our findings show that our objective is effective in increasing faithfulness and can lead to a useful analysis of NMT model behaviour and more trustworthy attention heatmaps. We assert that faithfulness is a good property to have in a model whether or not it will be useful for downstream interpretation. A model that is faithful can be trusted better as a component in a larger end-to-end neural model.

Contributions We seek to improve faithfulness of NMT models. To this end, we make the following contributions in this work:

- We propose a measure for quantifying faithfulness in NMT.
- We introduce a novel learning objective based on probability divergence that rewards faithful behavior and which can be included in the training objective for NMT.
- We provide empirical evidence that we can improve faithfulness in an NMT model. Our approach results in more a more faithful NMT model while producing better BLEU scores.[1]

We chose to study the impact of faithfulness in NMT because it is under-studied in terms of interpretability. Most previous work has focused on document or sentence-based classification tasks where attention models are not as directly useful as in NMT models. Attention is also more challenging in terms of faithfulness in the context of NMT models due to the substantial impact of the decoder component. While Transformers (Vaswani et al., 2017) generally produce better NMT models, they rely on multiple heads for attention. Defining an *overall* faithfulness measure in this case is challenging as different heads possibly have different faithfulness. Before addressing this more complicated problem, we first focus on the simpler single-head attention models. However, we expect larger and overparameterized models to get worse in terms of faithfulness because the language model in the decoder gets stronger in guessing the next word which, as we shall discuss in more detail later, tends to make attention less faithful.

2 Faithfulness in NMT Models

Intuitively, a faithful explanation should reflect the true internal reasoning of the model. Although there is no formal definition for faithfulness, a common approach in the community is to design stress tests to perturb the model parameters chosen in such a way that the model's decision should change if the model is faithful (Jacovi and Goldberg, 2020b). A common stress test is the *erasure* test in which the most-relevant part of the input is removed (Arras et al., 2017). In the context of NMT, at decoding time step t the attention component assigns attention weights α_t, attending to the source word at position $m_t = \text{argmax}_i\, \alpha_t[i]$ (or the k-best attended-to words in the source). These weights are often implicitly or explicitly regarded as an interpretation for the model's prediction at the time step t (Tu et al., 2016; Mi et al., 2016; Liu et al., 2016; Wang et al., 2016; Lee et al., 2017; Ding et al., 2017; Ghaeini et al., 2018). The erasure stress test for evaluating faithfulness offered by α_t is done by setting $\alpha_t[m_t]$ to zero and observing whether or not the output changes. It is worth noting that erasure is only one of the possible stress tests for evaluating faithfulness. Passing

[1] Our code is available at https://github.com/sfu-natlang/attention_regularization

more stress tests implies a more faithful model as it is properly reacting to more adversaries by changing its decision. In this paper we consider three intuitive stress test cases:

ZeroOutMax: (Arras et al., 2017): Here we remove attention from the most important token according to the attention weights by setting $\alpha_t[m_t] = 0$.

Uniform: (Moradi et al., 2019): In this stress test all attention weights are set to be equal, $\alpha_t = \frac{1}{m}\vec{1}$, where m is the length of the source sentence. This is to confuse the model about which part of the input is the most important one.

RandomPermute: (Jain and Wallace, 2019): In this stress test we randomly permute attention weights several times until a change in the model output is observed. We ensure that m_t, the most important token according to attention, is always changed. We set $\alpha'_t = \texttt{random\_permute}(\alpha_t)$ such that $\mathrm{argmax}_i\, \alpha'_t[i] \neq m_t$

Many prior studies of attention (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) have used a binary measure: either attention is faithful or it is not. These studies typically are about whether attention has the potential to be useful in terms of accuracy and faithful in terms of model behaviour. In many cases, especially in the case of NMT models, attention is clearly useful and by and large it must be faithful. The question is can we measure the faithfulness and improve faithfulness. It is more natural to have a gray-scale notion of faithfulness for evaluation (Jacovi and Goldberg, 2020b). Following this reasoning, we define $F(M)$ as faithfulness of attention heatmaps in model M as the following equation:

$$F(M) = \frac{\text{\# tokens passing stress tests}}{\text{\# tokens}} \qquad (1)$$

$F(M)$ is a number between 0 to 1 measuring the percentage of output tokens during inference which passed the stress tests, i.e., they changed in the presence of adversarial attention. This metric can also be regarded as a measure of trust we can assign to the attention heatmap to fully reflect the internal reasoning of the NMT model.

3 Approach

The conventional objective function in a sequence-to-sequence task is a cross-entropy loss $\mathcal{F}_{acc}$:

$$\mathcal{F}_{acc}(\theta) = -\frac{1}{|S|} \sum_{(X,Y)\in S} \log p(Y|X;\theta) \qquad (2)$$

where S is the training data and X and Y are source sentence and the correct translation respectively. This training objective does not explicitly model the interpretability aspects (e.g. faithfulness) of the network and it remains unoptimized during training.

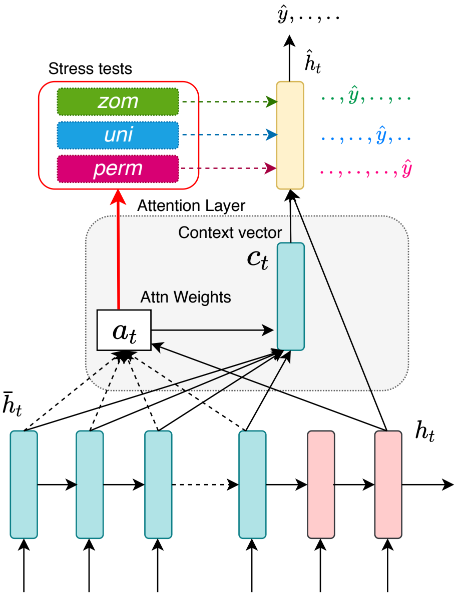


Figure 2: Using *ZeroOutMax, Uniform,* and *RandomPermute* stress tests, we generate adversaries to the attention weights. When adversarial attention weights are used, in a faithful model we expect the probability of the original output ($\hat{y}$) to drop significantly. We use this criterion to define a faithfulness objective function.

Faithfulness Objective In an effort to develop a model that is *right for right reason*, Ross et al. (2017) change the loss function of their classifier to model both *right answers* and *right reasons* instead of only the former. They achieve this by introducing a regularizing term that tends to shrink irrelevant gradients. In a similar spirit, we change our objective to account for the NMT model's faithfulness as well as the cross-entropy score against the reference translations:

$$\mathcal{F} = \mathcal{F}_{acc} + \lambda_{faith}\mathcal{F}_{faith} \qquad (3)$$

$\mathcal{F}_{faith}$ is an additional component that rewards the model for having more faithful attention. The parameter λ_{faith} regulates the trade-off between faithfulness and accuracy objectives. Our proposed metric for faithfulness is calculated based on discrete changes in the output under adversarial attention. It is not differentiable and cannot be simply used as $\mathcal{F}_{faith}$ to be optimized. Thus, we propose a novel differentiable objective which mimics a faithful behavior hoping it improves the discrete faithfulness metric.

While here we intend to improve faithfulness using adversarial attention weights, it is important to note that making the *model* more *robust* is *not* our main goal. Robustness of a model is its resilience against adversarial input or small perturbations such as typos. Whether or not our approach results in a more robust model is a separate research question that we have not focused on.

3.1 Divergence-based Faithfulness Objective

Consider a predictive model g_θ in which an intermediate calculation is later employed to justify predictions:

$$\hat{y} = \arg\max_y p(y|x) = \arg\max_y g_\theta(x, IC(x), y) \tag{4}$$

where $IC(x)$ is the intermediate calculation on the input. A concrete example for $IC(x)$ would be the context vector calculated by the attention mechanism.

Hypothesis If there exists an intermediate calculation $IC'(x)$ that conveys a contradictory post-hoc attention compared to $IC(x)$, then $IC(x)$ cannot be regarded as faithful for predicting $\hat{y}$. If $IC(x)$ is faithful, we expect the model to *diverge* from predicting $\hat{y}$ when $IC'(x)$ is employed instead.

Based on our hypothesis, we propose a divergence-based objective which mimics behavior of a faithful explanation under stress test:

$$\mathcal{F}_{faith} = \log p(\hat{y}|x, IC'(x)) \tag{5}$$

This objective is a negative loss that should be minimized. The minimum of this objective is achieved when the probability of the original prediction approaches zero under the stress test which is the ideal. Thus, it promotes reduction in output probability under an adversarial intermediate calculation (Figure 2). It is worth noting that this objective can be potentially employed in models where outputs are modeled as soft probabilities and thus is not limited to NMT. To put model under various stress tests we manipulate the context vector during training time by changing the attention weights and feed it to the decoder to calculate the probability. More precisely:

$$\begin{aligned}\mathcal{F}_{faith} = {} & \lambda_{zom} \log p(\hat{y}|x, IC'_{zom}(x)) \\ & + \lambda_{uni} \log p(\hat{y}|x, IC'_{uni}(x)) \\ & + \lambda_{perm} \log p(\hat{y}|x, IC'_{perm}(x))\end{aligned} \tag{6}$$

where IC'_{zom}, IC'_{uni} and IC'_{perm} are *ZeroOutMax*, *Uniform* and *RandomPermute* methods (see Sec. 2) to manipulate attention weights, respectively. $\lambda_{\{method\}}$ parameters regulate the contribution of each objective. We use the term $\mathcal{F}_{all}$ when all $\lambda_{\{method\}}$s in Eq. (6) are non-zero. Moreover, we use the term $\mathcal{F}_{\{method\}}$ when $\lambda_{\{method\}}$ is set to 1 and other regularization weights are zero.

3.2 On Attention Sparsity

Do the models trained with the faithfulness objective have sparser attention weights? Sharper attention in a model M might correlate with an intensified contribution of the most-attended source hidden state on the prediction resulting in higher faithfulness.

To measure sparseness of the attention, we take an average over the normalized entropy of attention distribution for each output token during inference on test data. We use normalized entropy which is in range [0,1] to account for the fact that the range of the entropy for each output token depends on the length of the corresponding source sentence.

$$AvgEnt = \frac{1}{\sum_{i=1}^{|S|} \left|\hat{Y}_i\right|} . \sum_{i=1}^{|S|} \sum_{j=1}^{\left|\hat{Y}_i\right|} NormEnt(\alpha_{ij}) \tag{7}$$

$$NormEnt(P) = -\sum_i \frac{P_i \log P_i}{\log N} \tag{8}$$

Here α_{ij} is the attention distribution for the output token j in the generated translation of source sentence i, and P is a discrete probability distribution. In Eq. (8) low entropy indicates a sharper distribution.

Attention Entropy Regularization Alongside investigating sparsity of the models trained by the faithfulness objective, we also train a model in which sparsity in attention is directly optimized. We used attention entropy regularization (Zhang et al., 2018):

$$\mathcal{F}_{ent} = \mathcal{F}_{acc} + \lambda_{ent} \sum_{i=1}^{|S|} \sum_{j=1}^{\left|\hat{Y}_i\right|} Ent(\alpha_{ij}) \tag{9}$$

where entropy of attention weights is added to the cross-entropy loss (2) as a regularization term.

4 Experimental Setup

4.1 Datasets

We use the Czech-English (Cs-En) dataset from IWSLT2016[2] and the German-English (De-En) dataset from IWSLT2014[3]. For the Czech-English dataset we use dev2010, tst2010, tst2011, tst2012, and tst2013 as the test data. For the German-English dataset we use dev2010, tst2010, tst2011, dev2012, and tst2012 as the test data. We used Moses (Koehn et al., 2007) to tokenize the dataset.

4.2 Architecture and Hyperparameters

We use OpenNMT (Klein et al., 2017) as our translation framework. We employ a 2 layer LSTM-based encoder-decoder (Sutskever et al., 2014; Cho et al., 2014) model with global attention (Luong et al., 2015). Dimension of the hidden states and the word embeddings for both source and target languages are set to 500. Vocabulary size for both the source and target language is set to 50000. We remove sentences with more than 50 tokens from the training data. We use Adam (Kingma and Ba, 2014) for training our models and we set the learning rate to 0.001. Models are trained until convergence. Our models have around 82M parameters. We optimize the hyperparameters of our models using the validation set. The baseline model is trained using Eqn. (2) and we call it $\mathcal{F}_{baseline}$. λ_{ent} in Eq. (9) is set to 0.04. We refer to the objective as $\mathcal{F}_{all}$ when λ_{zom}, λ_{uni}, and λ_{perm} are set to 0.5, 0.375, and 0.125 respectively. λ_{faith} is set to 1.

4.3 Training Difficulties

Our first attempts at using the modified objective function in Eq. (3) trained poorly. We observed that it was difficult for the model to learn the faithfulness constraint without having already learned to assign a reasonable probability to correct translations. To address this problem, we first train the NMT model using the standard unmodified objective function and then fine-tune this trained model by switching the objective function to Eq. (3).

One caveat is that the value of faithfulness loss can be arbitrarily large and interfere with the learning because cross-entropy error converge to infinity as the probability approaches to zero. However, we found that as long as the original output token received a small probability, there was no need for further reduction in it. Empirically, we clip losses more than 2.5 to 2.5 for each output token.

5 Results and Discussion

5.1 Impact on Faithfulness

To measure the effectiveness of the proposed objectives, we choose the best model in terms of provided faithfulness but within the 0.5 BLEU score of the maximum achieved BLEU score in the validation set. The reason is that we prefer a model that is both accurate and with faithful attention-based explanations. Table 1 shows the performance of the different faithfulness objective functions when generating content words and function words across different attention manipulation methods in the Czech-English (Cs-En) and German-English (De-En) datasets.

Results indicate that the proposed divergence-based objective has been effective in increasing the faithfulness metric. $\mathcal{F}_{all}$ is the most effective objective for increasing faithfulness when all stress tests are included in Eq. (1). When using $\mathcal{F}_{all}$, faithfulness of attention-based explanations for content words is increased 78% to 89%, while that of the function words is from 33% to 82%(see *All* column in Table 1). The same reductions are from 76% to 89% for content works and from 32% to 86% for function words in De-En dataset. These results establish the effectiveness of our proposed objectives to increase the faithfulness metric.

It is worth noting that increase in faithfulness of attention-based explanations for function words is much more than that of content words. This can be attributed to the fact the function words are mostly generated using the target-side information in the decoder (Tu et al., 2017; Moradi et al., 2019) and manipulating attention does not have much effect on generating them. However, our proposed faithfulness objective ($\mathcal{F}_{faith}$) seems to tighten the dependence of the decoder on the attention component. This results in much more increase in faithfulness for function words compared to such content words.[4] We also plot faithfulness over different

[2] https://sites.google.com/site/iwsltevaluation2016/

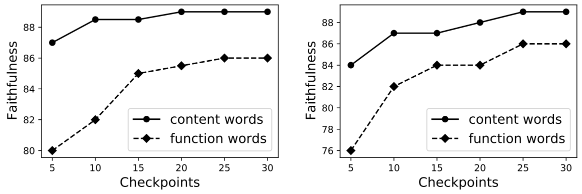
[3] https://sites.google.com/site/iwsltevaluation2014/

[4] If this dependence is not desired, it is possible not to penalize function words in the faithfulness objective. However, relying on attention for generating function words can be helpful, not necessarily for interpretability but for dealing with long-range dependencies (Vaswani et al., 2017) and, as a result, better translations.

	Objective	Content Words				Function Words			
		ZOM	*Uniform*	*RandPerm*	*All*	*ZOM*	*Uniform*	*RandPerm*	*All*
Cs-En	$\mathcal{F}_{baseline}$	83%	90%	94%	78%	46%	48%	64%	33%
	$\mathcal{F}_{zom}$	91%	93%	98%	86%	84%	87%	95%	74%
	$\mathcal{F}_{uni}$	84%	98%	97%	83%	56%	98%	91%	54%
	$\mathcal{F}_{perm}$	86%	95%	96%	83%	74%	97%	98%	71%
	$\mathcal{F}_{all}$	**91%**	**99%**	**98%**	**89%**	**83%**	**98%**	**98%**	**82%**
	$\mathcal{F}_{ent}$	78%	90%	94%	73%	46%	48%	64%	33%
De-En	$\mathcal{F}_{baseline}$	81%	90%	93%	76%	45%	48%	64%	32%
	$\mathcal{F}_{zom}$	91%	95%	98%	87%	87%	95%	97%	82%
	$\mathcal{F}_{uni}$	81%	98%	91%	80%	60%	100%	95%	58%
	$\mathcal{F}_{perm}$	85%	95%	97%	82%	74%	97%	98%	72%
	$\mathcal{F}_{all}$	**91%**	**98%**	**98%**	**89%**	**87%**	**100%**	**99%**	**86%**
	$\mathcal{F}_{ent}$	81%	90%	93%	76%	47%	47%	64%	33%

Table 1: Faithfulness metric for the generated content and function words through different objectives. Columns are different stress-tests included in the Eq.(1).

checkpoints in Figure 3. It indicates that progress in faithfulness is much faster for function words compared to content words.



(a) Cs-En

(b) De-En

Figure 3: Progress in faithfulness over different checkpoints. It increases much faster in function words compared to content words.

5.2 Effect of Training With Single Adversary on Passing Other Stress Tests

An interesting observation in Table 1 is that training with an adversary has positive effects on the model for passing stress tests from other types of adversaries. As an example, in Table 1 the column *Uniform* is the faithfulness metric when only Uniform test is employed in Eq. (1). When using this metric, we can observe that training a model with $\mathcal{F}_{perm}$ increased faithfulness from 90% to 95% for content words and from 48% to 97% for function words. We can see such effect for the German-English dataset as well. This observation indicates that training with each adversary can be beneficial for making model tolerant against other types of stress tests. It seems that training with each adversary strengthens the dependence of the decoder on the attention component which can be beneficial for passing other stress tests.

5.3 POS-tag Analysis

In addition to categorizing tokens into function and content words, we also analyze the effect of our proposed objective within different universal part-of-speech (POS) tags (Petrov et al., 2012) in Table 2. Our proposed objective has increased faithfulness in each POS tag and in our both datasets. Tokens with less lexical meaning are the ones affected the most as explained in Sec. 5.1. As expected, punctuations (PUNC) and particles (PRT) tags have benefited the most from increase in the faithfulness. Interestingly numbers (NUM tag) have the lowest increase in faithfulness. One reason might be that they already had a high initial faithfulness and this has made further increase less likely.

	De-En		**Cs-En**	
Tag	Baseline	Ours	Baseline	Ours
PUNC	0.19	**0.70**	0.28	**0.66**
PRON	0.42	**0.78**	0.35	**0.75**
VERB	0.47	**0.80**	0.50	**0.81**
ADP	0.30	**0.75**	0.40	**0.65**
DET	0.35	**0.74**	0.38	**0.70**
PRT	0.13	**0.63**	0.17	**0.50**
ADV	0.66	**0.80**	0.63	**0.79**
NOUN	0.63	**0.87**	0.64	**0.85**
ADJ	0.68	**0.87**	0.69	**0.85**
NUM	0.84	**0.86**	0.79	**0.86**
X	0.67	**0.78**	0.55	**0.80**

Table 2: Faithfulness metric within different part-of-speech (POS) tags.

5.4 Regularization Effect

The model checkpoints used in Tables 1 were selected based on maximum increase in faithfulness without sacrificing accuracy. To investigate if the proposed objective can have a general positive side effect in terms of accuracy, we train three independent models using the $\mathcal{F}_{baseline}$ and $\mathcal{F}_{all}$ objectives.

To make it fair for the baseline, we also add additional steps of training for the baseline model as well to isolate the benefit of adding the faithfulness objective. Table 3 contains the average BLEU score of the trained models. It indicates that the model trained with $\mathcal{F}_{all}$, has +0.7 and +0.4 increase in BLEU score compared to the baseline for the Czech-English and German-English language pairs respectively.

	Objective	BLEU%
Cs-En	$\mathcal{F}_{baseline}$	19.68
	$\mathcal{F}_{all}$	**20.4**
De-En	$\mathcal{F}_{baseline}$	24.85
	$\mathcal{F}_{all}$	**25.21**

Table 3: BLEU score of the baseline and the model trained with $\mathcal{F}_{all}$. Pairwise bootstrap resampling (Koehn, 2004) resulted in a p-value < 0.01 which indicates the statistical significance of the observed difference.

Improved BLEU scores for the faithful model can be due to two reasons: 1) the faithfulness objective can be seen as a regularization term which prevents the model from relying too much on the target-side context and the implicit language model in the decoder, which results in increased contribution of attention on the decoder and reducing some bias in the model. 2) penalizing the model for the lack of connection between justification and prediction forces the model to learn better translations by forcing it to justify each output in a *right answer for the right reason* paradigm. Figure 4 shows some examples of how our proposed model can produce better translations.

5.5 Do the New Models Have Sparser Attention?

Table 4 shows the average entropy and average normalized entropy for the baseline, the proposed model ($\mathcal{F}_{all}$), and the model trained with attention entropy regularization respectively. Evidently, the proposed model has not increased sparsity. On the other hand attention entropy regularization has been very effective in making attention weights sparser. But Table 1 indicates that attention entropy regularization has not been effective in increasing faithfulness. This suggests that sharper attention weights only affect the context vector and do not contribute to increased dependence of the decoder on attention.

src	es ist alles hier es ist alles online
ref	it 's all here it 's all on the web
base	it 's all right it 's all online .
ours	it 's all here it 's all online .
src	die erste ist , dass wir uns nicht weiterentwickeln werden .
ref	the first is that we will not evolve .
base	the first is that we will not move forward .
ours	the first is that we will not evolve .
src	sie drängten wasser aus dem land heraus und hinaus in den fluss
ref	they pushed water off the land and out into the river
base	they kept running water from the land and out in the river
ours	they pushed water out of the country and out in the river .
src	anstatt hunderte von kilometern entfernt im norden
ref	instead of hundreds of miles away in the north
base	instead of hundreds of miles away from north america
ours	instead of hundreds of miles away from north

Figure 4: These examples show some cases where the more faithful model trained using our faithfulness objective produces better translations compared to the baseline model. In each of these cases, perturbing the attention weights has no effect on the baseline model output. The faithful model is able to focus on the source side when needed in order to produce a more accurate translation.

	Model	*AvgEnt*	*AvgNormEnt*
Cs-En	$\mathcal{F}_{baseline}$	0.69	0.23
	$\mathcal{F}_{all}$	0.84	0.27
	$\mathcal{F}_{ent}$	**0.35**	**0.11**
De-En	$\mathcal{F}_{baseline}$	0.89	0.29
	$\mathcal{F}_{all}$	1.0	0.32
	$\mathcal{F}_{ent}$	**0.43**	**0.14**

Table 4: Average entropy and average normalized entropy of the baseline, the proposed model ($\mathcal{F}_{all}$), and the model trained with attention entropy regularization.

6 Related Work

Attention and Different Axes of Interpretability While several studies have focused on understanding the semantic notions captured by attention (Ghader and Monz, 2017; Vig and Belinkov, 2019; Clark et al., 2019), evaluating attention as an interpretability approach has garnered a lot of interest. From the faithfulness perspective, (Jain and Wallace, 2019; Serrano and Smith, 2019) show that for instances in a data set there can be adversarial attention heatmaps that do not change the output of the text classifier. In other words, adversarial attention leads to no decision flip in each instance. They use this to claim that attention heatmaps are not to be trusted, or unfaithful. Wiegreffe and Pinter (2019) argue against per-instance modifications at test time for two reasons: 1) in classification tasks attention may not be useful so perturbing attention is misleading. This is not true for NMT since attention is very useful in NMT. 2) they train an adversarial attention model (e.g. uniform attention) chosen to produce attention weights distant from the original attention weights while at the same time trying to minimize classification error. They show that such adversarial attention models are not as accurate as models with attention. In our work we acknowledge that attention is useful and faithful to some extent and we aim to improve faithfulness of NMT models.

While most of these works provide evidence that attention weights are not always faithful, Moradi et al. (2019) confirm similar observations on the unfaithful nature of attention in the context of NMT models. Li et al. (2020) is one of the few papers examining attention models in NMT. However, they are focused on the task of identifying relevant source words to explain the output translations selected by the NMT model. They look for optimal proxy models that agree with the NMT model such that the relevant source words picked as an explanation by a proxy model exhibits similar behaviour to the target model. They use the notion of *fidelity* over proxy models and evaluate several alternative proxy models using empirical risk minimization. Attention weights are evaluated alongside other proxy models for this task. In contrast, our work is about improving the faithfulness of NMT models and we focus on the internal state of the NMT model rather than proxy models. They use human references, e.g. AER, for evaluating fidelity. As discussed earlier, evaluation of faithfulness cannot involved human judgements or reference data. It is possible that our faithful NMT models are also better at fidelity, but that is an open question.

While prior works have mostly failed to explicitly distinguish faithfulness from plausibility in their arguments, Jacovi and Goldberg (2020a,b) focus on formalizing faithfulness and addressing evaluation of faithfulness separately from plausibility respectively. Subramanian et al. (2020) have investigated the concept of faithfulness in neural modular networks (NMN) which are employed for modeling compositionality. They question the faithfulness of the structure of the network modules describing the true abstract reasoning of the model. Similar to us, they attempt to quantify faithfulness and improve upon it. However their contributions like training with an auxilary atomic-task supervision for improved faithfulness are specific to the context of NMNs. Pruthi et al. (2020) demonstrate that it is possible to train a model that produces a deceptive attention mask, questioning the use of attention weights as explanation from the fairness and accountability perspective. Alvarez-Melis and Jaakkola (2018) investigate the interpretability methods from the robustness perspective. They attempt to quantify robustness and show that current interpretability methods cannot be considered as robust.

Sparsity For Improved Interpretability This line of work suggests making attention sparser so that the most contributing input word is more distinguishable over other input words. Martins and Astudillo (2016); Malaviya et al. (2018) propose sparse but differential alternatives to *softmax* function for calculating attention weights, while Zhang et al. (2018) propose sparsity regularization terms such as entropy regularization to promote sparsity in the attention.

Regularizing Explanations Ross et al. (2017) augment the loss function of their classification model with an explanation objective to constrain input gradient explanations. Rieger et al. (2019) follow a similar spirit but they use *contextual decomposition* (Murdoch et al., 2018) to extract explanations offered by the model. Aligning attention (as explanation) with prior knowledge has also been extensively studied. This prior knowledge can include alignment data (Mi et al., 2016; Liu et al., 2016), human rationales (Zhong et al., 2019), or even structural biases (Cohn et al., 2016).

Inherently Interpretable Neural Models Contrary to post-hoc explanation methods for interpreting a neural model, Stahlberg et al. (2018) show that the NMT model can be made self-explanatory by training it to produce the discrete decisions made by the model (from which the translations can be extracted later). In another work, (Lei et al., 2016; Bastings et al., 2019) propose models in which first a rationale is selected from the input and then is further used for prediction.

7 Conclusion

We proposed a method for quantifying faithfulness of NMT models. To optimize faithfulness we have defined a novel objective function that rewards faithful behavior through probability divergence. We also show that the additional constraint in the training objective for NMT does not harm translation quality and in some cases we see some better translations presumably due to the regularization effect of our faithfulness objective.

Future Work We aim to investigate and improve faithfulness of attention-based explanations in more sophisticated attention models such as Transformers (Vaswani et al., 2017). We can generalize our approach by designing explanatory modules in NMT through functionality separation (alignment, reordering, etc.) instead of relying only on attention. We also plan to investigate if faithful models can also be more useful for copy models and other applications of attention heatmaps in NMT.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. The research was partially supported by the Natural Sciences and Engineering Research Council of Canada grants NSERC RGPIN-2018-06437 and RGPAS-2018-522574 and a Department of National Defence (DND) and NSERC grant DGDND-2018-00025.

References

David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8).

Joost Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 142–151, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, Vancouver, Canada. Association for Computational Linguistics.

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)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, Brussels, Belgium. Association for Computational Linguistics.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*, page 8.

Alon Jacovi and Yoav Goldberg. 2020a. Aligning faithful interpretations with their social attribution. *arXiv preprint arXiv:2006.01067*.

Alon Jacovi and Yoav Goldberg. 2020b. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

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.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Zachary C Lipton. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3093–3102, Osaka, Japan. The COLING 2016 Organizing Committee.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Chaitanya Malaviya, Pedro Ferreira, and André F. T. Martins. 2018. Sparse and constrained attention for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–376, Melbourne, Australia. Association for Computational Linguistics.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2283–2288, Austin, Texas. Association for Computational Linguistics.

Sina Mohseni and Eric D Ragan. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2019. Interrogating the explanatory power of attention in neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 221–230, Hong Kong. Association for Computational Linguistics.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *International Conference on Learning Representations*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 340–350, Melbourne, Australia. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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.

Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. 2019. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *arXiv preprint arXiv:1909.13584*.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

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.

Felix Stahlberg, Danielle Saunders, and Bill Byrne. 2018. An operation sequence model for explainable neural machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 175–186, Brussels, Belgium. Association for Computational Linguistics.

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. 2020. Obtaining faithful interpretations from compositional neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5594–5608, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Jiajun Zhang, Yang Zhao, Haoran Li, and Chengqing Zong. 2018. Attention with sparsity regularization for neural machine translation and summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):507–518.

Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*.