Causal Transformers: Improving the Robustness on Spurious Correlations

Anonymous ACL submission

Abstract

The fully-connected dependencies in self-attention over-fit spurious correlations and limit the generalization on out-of-distribution data. Pre-trained language models (PLMs) alleviate this problem benefitted from the appreciable counterexamples in large-scale pre-training corpora. However, there is no study to resolve this problem by improving the model structure. We enforced the *causal independence mechanism* in the self-attention network, which constrains attention mapping topologies (AMGs) as causal structures. To implement it, we defined a smooth loss on the Markov boundary constrained directed acyclic graph (DAG) with the Lagrange duality, and used it to optimize the AMGs towards causal structures. Further, this causal attention network was applied on Transformer (Causal Transformer). The empirical results on two spurious correlation challenging (SCC) datasets, neural machine translation (NMT) and natural language inference (NLI) tasks demonstrated that our Causal Transformer outperforms the state-of-the-art model and improves the out-of-distribution prediction.

1 Introduction

Recently, Transformer (Vaswani et al., 2017) and the pre-trained language models (PLMs) on top of it, such as the BERT (Devlin et al., 2019), GPT (Radford et al., 2019), ViLBERT (Lu et al., 2019), ViT (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021) have achieved state-of-the-art performance across a wide range of natural language processing (NLP) (Lin et al., 2021), computer vision (CV) (Han et al., 2020), multimodal (Yao and Wan, 2020) and other tasks. As a vital component in Transformer, the self-attention is optimized to learn the fully-connected dependencies of a single token (or batch) with respect to all other tokens (or batches) in the same sequence. The self-attention outperforms traditional deep learning models (like recurrent neural networks (RNNs), convolutional neural networks (CNNs) and et al.) as it can capture long-term and non-consecutive dependencies. This superiority helps the models to overcome the catastrophic forgetting in previous deep learning models and contributes to the popularity of PLMs on top of Transformers.

Yaghoobzadeh et al. (2021) conducted an experiment that trains the BoW, BiLSTM and BERT with Transfer learning on MNLI, QQP and FEVER sequentially. The number of forgotten examples in the training process and the final prediction on test data were recorded in Figure 1.

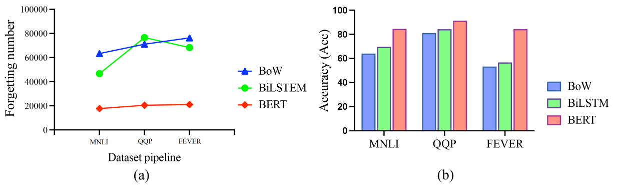

Figure 1: The experiment conducted on BoW, BiLSTM, BERT with multitask learning on MNLI, QQP and FEVER, where (a) is the numbers of forgetting examples in the Transfer learning process on three datasets and (b) is the test accuracies of different models.

This experiment demonstrated that BERT can remember most examples that it has encountered. This superiority improves the robustness of PLMs on spurious correlations because the models encounter enough counterexamples in large-scale pre-training corpora (Tu et al., 2020). But, Tu et al. (2020) demonstrated with experimental results that BERT does not outperform traditional models for spurious correlation challenging datasets when it is trained from scratch. The fully-connection in self-attention networks is one of the important causes of this phenomenon (Wang and Culotta, 2020), and which limits the generalization on out-of-distribution test data.

Recent studies (Jain and Wallace, 2019; Lin et al., 2021) proposed the sparse attentions to im-

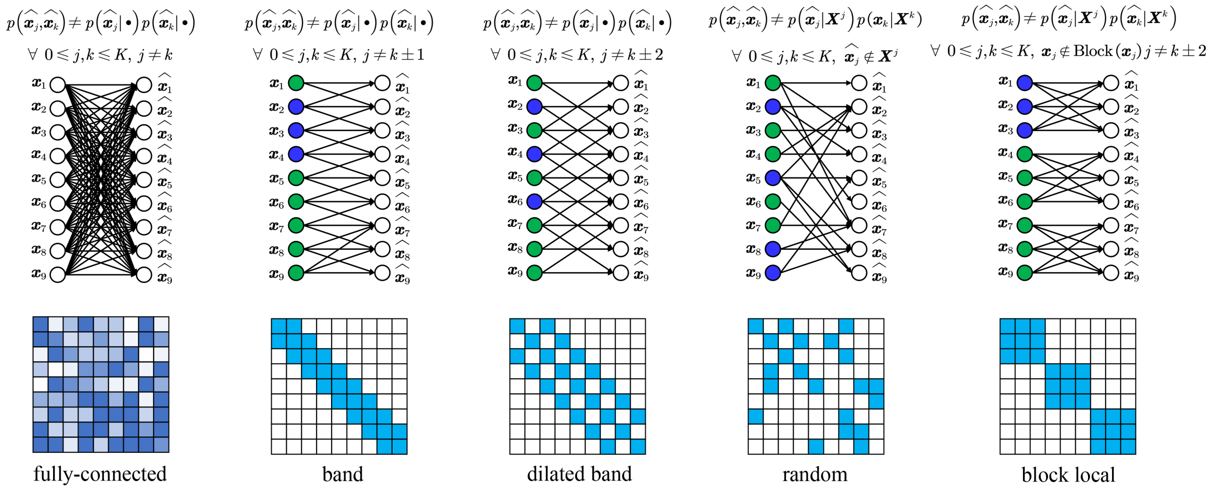
$p(\widehat{x}_j,\widehat{x}_k) \neq p(\widehat{x}_j|\bullet)p(\widehat{x}_k|\bullet)$
$\forall\ 0 \leqslant j,k \leqslant K,\ j \neq k$
$p(\widehat{x}_j,\widehat{x}_k) \neq p(\widehat{x}_j|\bullet)p(\widehat{x}_k|\bullet)$
$\forall\ 0 \leqslant j,k \leqslant K,\ j \neq k \pm 1$
$p(\widehat{x}_j,\widehat{x}_k) \neq p(\widehat{x}_j|\bullet)p(\widehat{x}_k|\bullet)$
$\forall\ 0 \leqslant j,k \leqslant K,\ j \neq k \pm 2$
$p(\widehat{x}_j,\widehat{x}_k) \neq p(\widehat{x}_j|X^j)p(x_k|X^k)$
$\forall\ 0 \leqslant j,k \leqslant K,\ \widehat{x}_j \notin X^j$
$p(\widehat{x}_j,\widehat{x}_k) \neq p(\widehat{x}_j|X^j)p(\widehat{x}_k|X^k)$
$\forall\ 0 \leqslant j,k \leqslant K,\ x_j \notin \mathrm{Block}(x_j)\, j \neq k \pm 2$
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9
$\widehat{x}_1$ $\widehat{x}_2$ $\widehat{x}_3$ $\widehat{x}_4$ $\widehat{x}_5$ $\widehat{x}_6$ $\widehat{x}_7$ $\widehat{x}_8$ $\widehat{x}_9$
fully-connected
band
dilated band
random
block local

Figure 2: The fully-connected self-attention and different sparse attentions (band, dilated band, random and local block band), where x, $\widehat{x}$ represent the input token and output token respectively. In fully-connected attention, the great majority tokens have not conditional independent distributions with each other. In the band, dilated band and local block band attentions, the blue tokens have conditional independencies with the green tokens.

prove the test generalization, and these existing methods can be categorized into the ***band***, ***random*** as shown in Figure 2. The ***band attentions*** explore the localness modelling for the self-attention network, and it has been branched in to *local consecution*, *dilatation* and *local block* (Yang et al., 2018; Guo et al., 2019a). The ***random sparsity*** randomly activates attention connections taking the attention scores as the probabilities, or samples the attention values from prior distribution only (Wang et al., 2020; He et al., 2020; Ying et al., 2021; You et al., 2020). Band sparsities improve the conditional independence on the attention mapping graphs (AMGs), which reduces great spurious correlations. But the localness limits the reasoning on long-term and non-consecutive relations. The random attentions can learn more non-consecutive relations than the ***local band*** in each layer, but which impede the limitation on spurious correlations.

There is no research to explore how to eliminate the useless attention mapping dependencies in Transformers. To tackle this problem, we introduced the *causal independence mechanism* (ICM) (Peters et al., 2017) on the attention mapping graph (AMGs). This mechanism means that the feature updater on a token, giving its causes, does not inform or influence the feature updaters on other tokens. Which aims to alleviate the over-fitting problem on useless correlations (spurious correlations). Our method implements this ICM by enforcing a directed acyclic graph (DAG) constraint in the self-attention mappings (Luo et al., 2020).

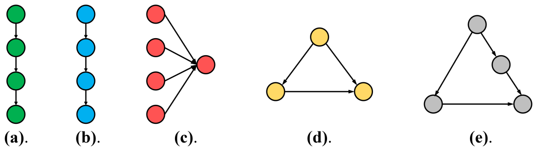

Figure 3: The examples of DAGs that satisfy the no-loop constraint. The DAGs (a), (b) and (c) have simple relations that cannot capture long-range and non-consecutive dependencies and DAGs (d), (e) cannot distinguish the confusions.

The DAGs learning has become practicability in deep learning to improve the out-of-distribution prediction (Krueger et al., 2021) after the smooth losses on DAGs were proposed by Zheng et al. (2018a, 2020). This loss optimizes random graphs towards DAGs by eliminating all loops. However, *the no-loop is a necessary but insufficient criterion for the ICM* in causal Bayesian networks, as shown in Figure 3.

In this paper, we first analyzed the over-fitting problem on the spurious correlations in the self-attention networks. Further, we chose a stable DAG constraint - Markov Blanket (also called Markov boundary) (Liu and Liu, 2018) as the prior sparsity on the attention mapping. Next, we defined the smooth loss with the Lagrange duality to turn this Markov Blanket constraint into a pure deep learning optimization. The Markov blanket on each node in the attention mapping graphs (AMGs) is a three-layer Bayesian network, which needs two-step inference to traverse through all relevant nodes. Consequently, we took a two-layer graph convolutional network (GCN) as the node feature updater in causal attention network (CAN). Finally, we applied this CAN in the Transformers (Causal Transformers) and evaluated it on spurious correlation challenging (SCC) datasets, natural language inference (NLI), neural machine translation (NMT).

Our contributions are summarized as follows.

- We analyzed the over-fitting problem on spurious correlations in the self-attention networks.
- We introduced a DAGs topology - Markov blanket as the prior sparsity on the self-attention networks, and turned the Markov blanket constraint into a pure deep learning problem by defining its smooth loss with the Lagrange duality.
- To reason over the Markov Blanket, we used a two-layer GCN as the feature updater.
- We enforced this causal attention network in language Transformers.
- We empirically showed that our Causal Transformer outperforms the newest Transformer variants on out-of-distribution generalization and other NLP tasks.

2 Background

2.1 Self-Attention Over-Fits on Spurious Correlations

As the most pivotal component in the Transformer, the self-attention captures fully-connected dependencies in a sequence with a parallel mode.

$$\text{Attention}\left(\boldsymbol{X}\right) = \text{softmax}\left(\frac{\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_k\boldsymbol{X}^T}{\sqrt{d}}\right)\boldsymbol{X}\boldsymbol{W}_v$$

where $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ is the input sequence, $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the attention score matrix and $\boldsymbol{W}_q \in \mathbb{R}^{d\times d_k}$, $\boldsymbol{W}_k \in \mathbb{R}^{d_k\times d}$, $\boldsymbol{W}_v \in \mathbb{R}^{d\times d_v}$.

As fully-connected attention has a dense score matrix, as shown in the Figure 2, few causal independencies exist in the self-attention mapping as any token j's ($1 \leq j \leq K$) feature is updated depended on all tokens. The fully-connected dependencies easily capture the spurious correlations occurring in the training data, which is the reason why Transformers do not outperform the traditional deep learning models when it is trained from scratch (Tu et al., 2020).

In the band (includes dilatation, local block) attentions, each token k, $1 \leq k \leq K$ is conditional independent with token j, $j \notin [k-b, k+b]$ lying out the band. That reduces the spurious correlations significantly. But the band attentions cannot capture long-term dependencies, which is why sparse attentions achieve better performance only when they are used together.

Another popular sparsity is random attention, like the ***dropout*** trick we often use, which alleviates the over-fitting on spurious correlations by randomly activating part of attention mapping connections with probability.

2.2 Causal Inference with DAGs Learning

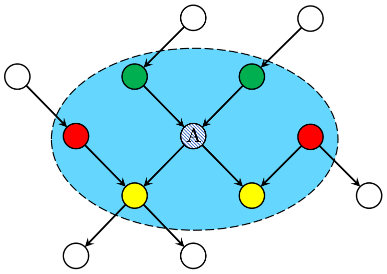

Figure 4: A graph constrained with Markov blanket (boundary). The Markov blanket on node A is marked with blue roundel where the green nodes are A's parents, yellow nodes are A's childrens and red nodes are A's spouses.

Zheng et al. (2018b) firstly proposed a continuous optimization on DAGs, which turns the no-loop constraint on DAGs topology into a pure deep learning problem. Afterwards, the causal inference has shown a big step forward in the deep learning community (Luo et al., 2020).

$$\begin{aligned} &\min_{\boldsymbol{W}\in\mathbb{R}^{d\times d}} F\left(\boldsymbol{W}\right) \\ &\text{subject to} \quad h\left(\boldsymbol{W}\right) = \text{tr}\left(e^{\boldsymbol{W}\circ\boldsymbol{W}}\right) - d = 0 \end{aligned} \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ represents the weighted adjacency matrix, $F : \mathbb{R}^{d\times d} \to \mathbb{R}$ is the regression function on the DAGs, $h\left(\boldsymbol{W}\right)$ is the smooth loss on graph $g(\boldsymbol{W})$ and $\circ$ is the Hadamard product.

Minimizing the smooth loss $h\left(\boldsymbol{W}\right) \to 0$ will remove all loops in the graph $g(\boldsymbol{W})$, which optimizes the $g(\boldsymbol{W})$ as a DAG. But the no-loop constraint, as shown in Figure 3, (i) is *a necessary but insufficient condition* for DAGs learning and cannot distinguish confusions; (ii) easily optimizes the graphs towards simple structures. These cases are incapable for reasoning in many realistic tasks (like the NLP, CV and et al.).

In causal Bayesian networks, Pellet and Elisseeff (2008) proved that a Markov blanket (or Markov boundary) on a random variable is a minimum set with all the useful information to infer it. For a random variable $\boldsymbol{x}_i$ in the random variable set $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_n\}$, its Markov Blanket is a subset $\boldsymbol{S} \subset \boldsymbol{X}$, and $\boldsymbol{x}_i$ is conditional independent with all random variables $\boldsymbol{x}_j \in \{\boldsymbol{X} - \boldsymbol{S}\}$. In general, the Markov blanket of a node $\boldsymbol{x}$ in Bayesian networks is a set of $\boldsymbol{x}$'s parent nodes, children nodes and spouse nodes (i.e., children's other parents) as shown in Figure 4.

3 Causal Transformer

In this section, we describe the causal independence constraint on the self-attention network and use it to implement the Causal Transformer. Our Causal Transformer has the similar Encoder-Decoder framework as the one in vanilla Transformer. The model schematics are shown in Figure 6.

3.1 Continuous Optimization on Causal Self-Attention

As shown in the Figure 6, in the self-attention $\mathbf{A}$ mapping graph, each token is constrained with the Markov Blanket. We propose a smooth loss $h(\mathbf{A})$ to optimize the attention $\mathbf{A}$ mapping graph toward the Markov Blanket constrained DAG.

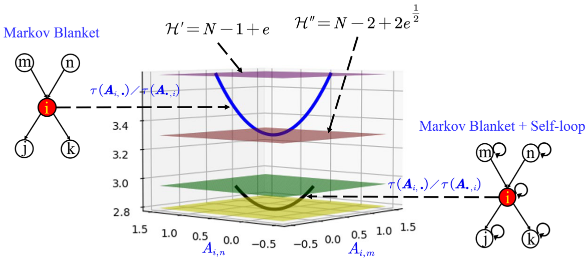

Figure 5: The space for the measurement $\tau(\mathbf{A}_{i,\cdot})$ on the Markov blanket of node i with two parents m, n and two childrens j, k. The blue line is the constraint on $\tau(\mathbf{A}_{i,\cdot})$ and $\tau(\mathbf{A}_{\cdot,i})$ for the Markov Blanket without self-loop and the black line is the constraint for the Markov Blanket with self-loop.

Definition 1. *The Markov blanket of a node has $1 \sim 2$ parents and $1 \sim 2$ childrens, this constraint can be defined as an mathematical inequality on the normalized attention scores $\mathbf{A} \in \mathbb{R}^{N \times N}$, $\sum_{j=1}^{N} A_{i,j} = 1 \wedge \sum_{i=1}^{N} A_{i,j} = 1$:*

$$
\begin{aligned}
N - 2 + 2e^{\frac{1}{2}} \leq \tau\left(\mathbf{A}_{i,\cdot}\right), \tau\left(\mathbf{A}_{\cdot,i}\right) \leq N - 1 + e,& \\
\text{for} \quad i = 1, \cdots, N& \\
\text{where} \quad \tau\left(\mathbf{A}_{i,\cdot}\right) = \sum_{j=1}^{N} e^{A_{ij}},& \\
\tau\left(\mathbf{A}_{\cdot,i}\right) = \sum_{j=1}^{N} e^{A_{ji}}&
\end{aligned}
$$

Definition 2. *To adapted for wide NLP tasks, the DAG constraint on the attention mapping graph must has slack variables to tolerate the self-loops:*

As each node contains important information about itself in many tasks, we enforce the self-loop in attention map $\mathbf{A}$ by slack variables $\boldsymbol{\varepsilon}^{'}, \boldsymbol{\varepsilon}^{''} \in \mathbb{R}^N$. The optimization objective can be defined with the Lagrange duality as follows,

$$
\begin{aligned}
\min : & \sum_i \left(\boldsymbol{\varepsilon}_i^{'} + \boldsymbol{\varepsilon}_i^{''}\right) \\
\text{s.t.} \quad & 2\mathcal{H}^{'} - \tau\left(\mathbf{A}_{i,\cdot}\right) - \tau\left(\mathbf{A}_{\cdot,i}\right) + \varepsilon_i^{'} \leqslant 0 \\
& \tau\left(\mathbf{A}_{\cdot,i}\right) + \tau\left(\mathbf{A}_{i,\cdot}\right) - 2\mathcal{H}^{''} + \varepsilon_i^{''} \leqslant 0 \\
& 0 \leqslant \varepsilon_i^{'} \leqslant \varepsilon_{max}^{'} \\
& 0 \leqslant \varepsilon_i^{''} \leqslant \varepsilon_{max}^{''}
\end{aligned}
\tag{2}
$$

where $\mathcal{H}^{'} = N - 2 + 2e^{\frac{1}{2}}$, $\mathcal{H}^{''} = N - 1 + e$.

As shown in Figure 5, the self-loop is set with a $1/3$ weight when the attention $\mathbf{A}$ is in the ***Non-Autoregressive Encoding block***, and $\varepsilon_{max}^{'} = 0.7045$, $\varepsilon_{max}^{''} = 0.5062$ correspondingly. When the attention $\mathbf{A}$ is in the ***Autoregressive Decoding block***, there is no self-loop and $\varepsilon_{max}^{'} = 0.$, $\varepsilon_{max}^{''} = 0.$.

Definition 3. *The loss $h(\mathbf{A})$ must be differentiable and can be optimized with the deep learning models together:*

Given the Lagrange multipliers $\boldsymbol{\gamma}$, the smooth loss function $h(\mathbf{A})$ is defined as follows,

$$
\begin{aligned}
h(\mathbf{A}) = & \sum_i \left[\gamma_1 \left(\varepsilon_i^{'} - \varepsilon_{max}^{'}\right) + \gamma_2 \left(\varepsilon_i^{''} - \varepsilon_{max}^{''}\right)\right] \\
& + \sum_i \gamma_3 \left(2\mathcal{H}^{'} - \tau\left(\mathbf{A}_{i,\cdot}\right) - \tau\left(\mathbf{A}_{\cdot,i}\right) + 2\varepsilon_i^{'}\right) \\
& + \sum_i \gamma_4 \left(\tau\left(\mathbf{A}_{i,\cdot}\right) + \tau\left(\mathbf{A}_{\cdot,i}\right) - 2\mathcal{H}^{''} + 2\varepsilon_i^{''}\right)
\end{aligned}
\tag{3}
$$

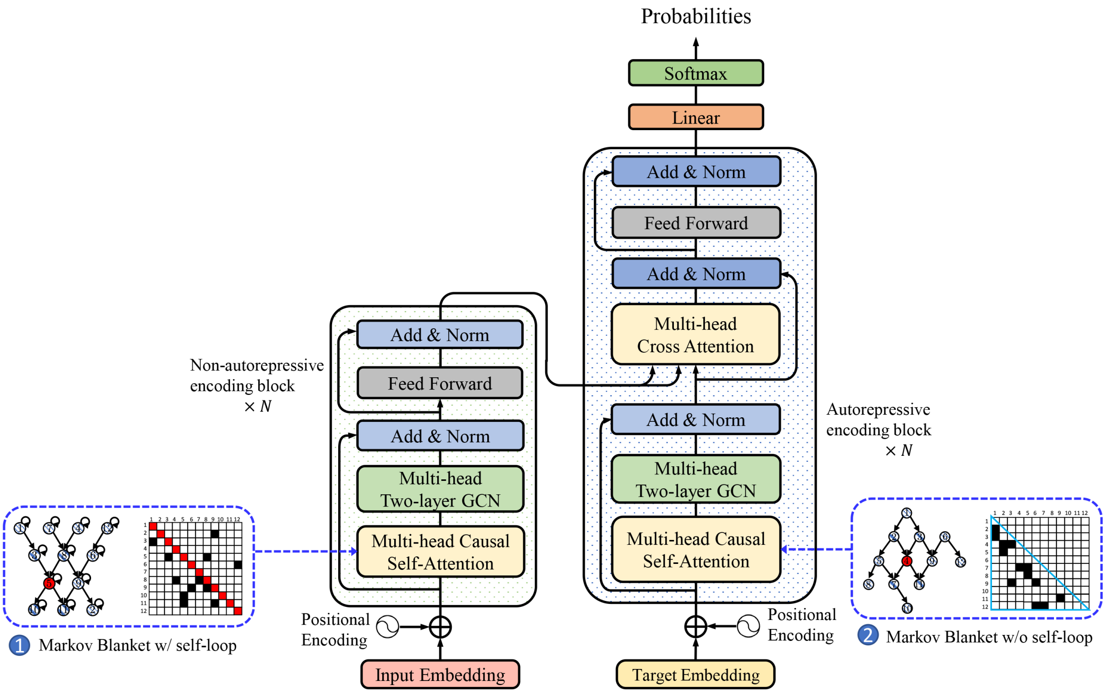

Figure 6: The schematics of Causal Transformer where the causal attention in left encoding block is enforced with self-loop and the causal attention in decoding block masked the upper triangular for autoregressive prediction.

3.2 Implementation

As shown in the Figure 6, the encoding and decoding blocks in our Causal Transformer are composed of the multi-head causal self-attention networks, two-layer GCN feature updaters, the multi-head cross-attentions, normalization (Norm) layers and position-wise feed-forward (FFN) layer. The causal self-attention networks in the encoding block are added the self-loops.

Causal Self-Attention Network Given the input sequence $\mathbf{X} \in \mathbb{R}^{N \times d}$, the feature updater in the causal self-attention network is defined as follows,

$$\begin{aligned} \text{Self-Attention}\left(\boldsymbol{X}\right) &= \underbrace{\sigma\left(\mathbf{A}\sigma\left(\mathbf{A}\boldsymbol{X}\boldsymbol{W}_v^{(0)}\right)\boldsymbol{W}_v^{(1)}\right)}_{\text{2-layer GCN}} \\ \mathbf{A} &= \operatorname{softmax}\left(\frac{\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_k\boldsymbol{X}}{\sqrt{d_k}}\right) \end{aligned} \tag{4}$$

where $\boldsymbol{W}_q \in \mathbb{R}^{d \times d_k}$, $\boldsymbol{W}_k \in \mathbb{R}^{d_k \times d}$, $\boldsymbol{W}_v^{(0)} \in \mathbb{R}^{d \times d_v}$ and $\boldsymbol{W}_v^{(1)} \in \mathbb{R}^{d_v \times d_v}$.

As each node in the Markov Blanket mapping graph needs a two-step inference path to reach all relevant nodes, we replace the Query-Key-Value updater with the two-layer GCN. The attention score matrix $\mathbf{A}$ in Eq 4 works as an adjacency matrix in the 2-layer GCN.

Decoding Block The decoding block predicts the targets $\boldsymbol{Y} \in \mathbb{R}^{M \times d}$ in an auto-regressive way,

$$\begin{aligned} \widehat{y}_m =& p\left(\boldsymbol{y}_m|\boldsymbol{y}_{0:m-1}, \boldsymbol{Z}_{\text{enc}}\right) \\ & m = 1, \cdots, M \end{aligned} \tag{5}$$

where $\boldsymbol{Z}_{\text{enc}}$ is the output of the encoding block.

The model is optimized by minimizing the sum of the causal loss function $h\left(\right)$ on all self-attentions and the cross-entropy between the predicted target $\widehat{y}$ and true label y.

$$\begin{aligned} \mathcal{L} =& -\sum_m y_m \log\left(\widehat{y}_m\right) \\ &+ \sum_k \left[h\left(\mathbf{A}_{\text{enc}}^{(l)}\right) + h\left(\mathbf{A}_{\text{dec}}^{(l)}\right)\right] \end{aligned} \tag{6}$$

where l represents the layer number, **enc** and **dec** represent the self-attentions in encoding and decoding blocks respectively.

4 Experiments and Analysis

To experimentally demonstrate the robustness of our Causal Transformer on spurious correlations, we conducted an experiment on two spurious correlation challenging (SCC) datasets. Besides, we also conducted the natural language inference (NLI), neural machine translation (NMT) tasks to evaluate the performance of our causal attention network.

Task	Dataset	Spurious correlations	Input	Label
natural language inference (NLI)	MNLI (In-distribution training)	high word overlap ⇒ entailment	P: What happened in actual fact to a young man who ran incessantly and persistently through the London streets? H: What happened to the guy running through the streets? P: You'd think this sort of thing would show up in more educational reform plans. H: You would think that would be in reform plans.	entailment
	HANS (Out-of-distribution test)	high word overlap ⇒ non-entailment	P: The authors introduced the students. H: The students introduced the authors. P: The managers were introduced by the senator. H: The managers introduced the senator.	non-entailment
paraphrase identification (PI)	QQP (In-distribution training)	same bag-of-words ⇒ paraphrase	S1: Bangkok vs Shanghai? S2: Shanghai vs Bangkok? S1: What is the funniest joke you know? S2: What is the funniest joke of all time?	paraphrase
	PAWS-QQP (Out-of-distribution test)	same bag-of-words ⇏ paraphrase	S1: What is it like to be poor after being rich? S2: What is it like to be rich after being poor? S1: Are all dogs smart or can some be dumb? S2: Are all dogs dumb or can some be smart?	non-paraphrase

Figure 7: Some examples in the in-distribution training datasets (MNLI, QQP) and spurious correlation challenging datasets (HANS, PAWS-QQP). The high word overlaps are highlighted with blue text. The MNLI dataset has high word overlaps on the entailment examples, and the challenging dataset HANS has high word overlaps on the non-entailment examples. Likewise, the in-distribution dataset QQP and the out-distribution challenging dataset PAWS-QQP have high word overlaps on paraphrase and non-paraphrase respectively.

- **SCC** is a task to evaluate whether the models trained on the in-distribution dataset (high word overlap, i.e., the spurious correlations) have robust performance on the out-of-distribution challenging test set (Tu et al., 2020). Our experiment was conducted on the MNLI (Nangia et al., 2017), HANS (McCoy et al., 2019), QQP (Chen et al., 2018) , PAWS-QQP datasets (Zhang et al., 2019). As shown in Figure 7, the MNLI and QQP have high spurious correlations (word overlaps) in the entailment and paraphrase sentence pairs. In contrast, the HANS and PAWS-QQP have out-distributions with high spurious correlations on non-entailment and non-paraphrase sentence pairs.

- **Machine Translation** task was conducted on WMT2014 English-German (EN-DE) with 4.5M translation pairs and WMT2016 English-Romanian (EN-RO) with 610k translation pairs. These datasets are tokenized and segmented into subword units with BPE encodings (Sennrich et al., 2016). The WMT2014 EN-DE was preprocessed with the scripts provided by (Vaswani et al., 2017) and the WMT2016 EN-RO was preprocessed data provided by (Lee et al., 2018).

- **NLI** task is to predict whether the given premise sentence and hypothesis sentence is "entailed by", "neutral with" or "contradicts". Our experiment was conducted on the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015), a collection of 570k human-written English sentence pairs manually labeled for balanced classification.

4.1 Robustness on Spurious Correlations

The first experiment was conducted on MNLI, HANS, QQP, PAWS-QQP datasets to evaluate the robustness of our Causal Transformer on the spurious correlations. Our model was trained from scratch on the in-distribution datasets - HANS and QQP respectively and evaluated on challenging datasets - HANS, PAWS-QQP without any fine-tuning. The model was trained 200 epochs on the base hyper-parameters as shown in Table, and the learning scheduler uses warm-up with 12800 warm steps. The dropout $p = 0.1$ was used apart from the causal attention networks, and the word embeddings were initialized from the 300-d GloVe (Pennington et al., 2014). Besides, we also trained our Causal Transformer in the single-task learning (STL) way. In the STL mode, our Causal Transformer was trained only on the challenging datasets from scratch.

As shown in the Figure 7 that HANS dataset has an out-distribution with the MNLI dataset on

Training mode	Models	MNLI		QQP	
		In-distribution MNLI-m	Challenging HANS	In-distribution QQP	Challenging PAWS-QQP
scratch	BERT-scratch (Devlin et al., 2019)	67.9	49.9	83.0	40.6
	ESIM (Chen et al., 2017)	78.1	49.1	85.3	38.9
pre-trained	BERT-base	84.0	53.8	90.5	33.5
	BERT-base (STL) (Devlin et al., 2019)	84.5	62.5	90.8	36.1
	BERT-base (Tu et al., 2020)	83.7	68.2	91.3	45.9
	BERT-large	86.2	71.4	91.3	40.1
	RoBERTa-base (STL) (Liu et al., 2019)	87.4	74.1	91.5	42.6
	RoBERTa-base	86.4	72.8	**91.7**	51.7
	RoBERTa-large	**89.1**	77.1	89.0	39.5
ours	Causal Transformer (STL)	-	81.7	-	67.1
	Causal Transformer	72.8	**85.2**	87.8	**78.1**

Table 1: The test accuracies on the in-distribution datasets (MNLI, QQP) and out-distribution challenging datasets (HANS, PAWS-QQP).

entailment examples. The challenging examples in HANS can verify whether models can improve the robustness on spurious correlations. The results are shown in Table 1. Our model achieved the best results on both challenging datasets with 85.2%, 78.0%, and these accuracies are far better than the pre-trained language models. On the in-distribution datasets, even though our model had not achieved the best results, it outperformed the BERT-scratch, ESIM and DA. Our Causal Transformer also achieved better results than BERT and RoBERTa when trained just on the challenging datasets.

We chose a case example "premise: *no, sir i'm afraid i didn't*; hypothesis: *yes, of course i did, sir*, entailment" to visualize the attention heatmaps on the vanilla Transformer and our Causal Transformer in Figure 8. The causal attention dependencies in our model on this example were plotted in Figure 9.

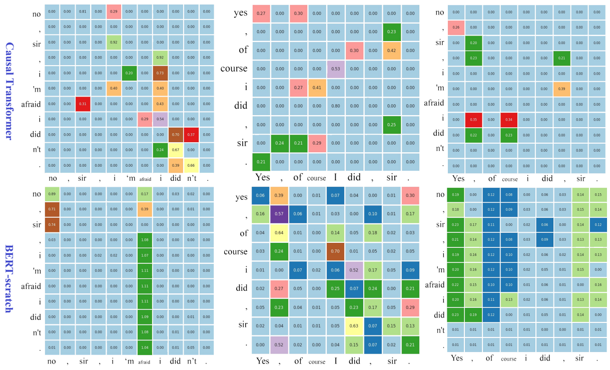

Figure 8: The self-attention heatmaps in Causal Transformer and BERT when they are both trained from scratch, this figure just exhibits the values in the 1-th head of 1-th layer in encoding, decoding blocks.

In these results, (i) our Causal Transformer achieved the new state-of-the-art test accuracies on HANS, PAWS-QQP respectively. (ii) The comparison on different pre-trained models (BERT, RoBERTa) shown that the size of the pre-training corpus contributes to the prediction on out-of-distribution challenging data. (iii) Our Causal Transformer also outperforms all compared models when it is trained only on the challenging dataset. (iv) Combined the results shown in Table 1, Figure 8 9 together to analyze, the causal attention networks in our Causal Transformer capture sparse dependencies than the BERT. These sparse dependencies maintain the most meaningful and interpretable word relations compared with the ones in BERT.

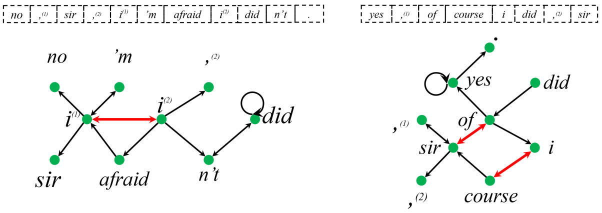

Figure 9: The causal attention mapping dependencies in MNLI dataset where there is just little confusions in these graphs.

4.2 Machine Translation

Our second experiment was conducted on the WMT2014 English-to-German (En-De), En-De and WMT2016 English-to-Romanian (En-Ro), Ro-En neural machine translation (NMT) tasks. We fixed the Causal Transformer with architecture $d_{model} = 512$, $n_{layer} = 6$, $d_k, d_v = 64$ and $n_{head} = 8$. We chose the preprocessed dataset in (Vaswani et al. (2017)) for the WMT2014 EN-DE and the dataset preprocessed in (Lee et al. (2018))

Models	WMT14		WMT16	
	EN-DE	DE-EN	EN-RO	RO-EN
Transformer (Vaswani et al., 2017)	27.30	-	-	-
Transformer_Rep (Takase and Kiyono, 2021)	**32.35**	-	-	
Glancing Transformer (Qian et al., 2021)	27.48	31.27	33.70	34.05
LevT (Gu et al., 2019)	27.27	-	-	33.26
Mask-Predict (Ghazvininejad et al., 2019)	27.03	30.53	33.08	33.31
JM-NAT (Guo et al., 2020)	27.31	31.02	-	-
imit-NAT (Wei et al., 2019)	22.44	25.67	28.61	28.90
Flowseq (Ma et al., 2019)	23.72	28.39	29.73	30.72
NAT-DCRF (Sun et al., 2019)	23.44	27.22	-	-
Imputer (Saharia et al., 2020)	25.80	28.40	32.30	31.70
Causal Transformer	31.30	**32.90**	**34.85**	**34.73**

Table 2: The BLEU scores on WMT2014 EN-DE, DE-EN and WMT2016 EN-RO, RO-EN.

for the WMT2016 EN-RO. Both datasets were tokenized and segmented into subword units with the BPE encodings (Sennrich et al., 2016). The word embeddings in this experiment were initialized randomly and trained with the model together. All results on the WMT2014 EN-DE, DE-EN and WMT2016 EN-RO, RO-EN are shown in Table 2. In which, our CausalTransformer, vanilla Transformer, Transformer_Rep use the autoregressive decoding block, and the other compared models use the non-autoregressive decoding blocks.

Table 2 shows that our Causal Transformer achieved new state-of-the-art scores on the WMT2014 DE-EN and WMT2016 EN-RO, RO-EN tasks respectively. Moreover, our Causal Transformer was also the one closest to the best result achieved by the Transformer_Rep with external knowledge on WMT2014 EN-DE task. Our Causal Transformer outperforms the Glancing Transformer with random sparsity and all state-of-the-art non-autoregressive Transformers.

4.3 Natural Language Inference

The third experiment was conducted NLI task to demonstrate the effectiveness of our Causal Transformer by comparing with other sparse attentions used in the newest Transformer variants. The compared models include vanilla Transformer, Transformer-RPR, Transformer-XL, Adapted Transformer, Star-Transformer and DA-Transformer. Here we use the Stanford Natural Language Inference (SNLI) dataset to evaluate the performance of CausalTransformer and all compared models. Our Causal Transformer in this experiment was trained with the hyper-parameters same with the BERT-base, and the word embeddings were initialized from 300-d GloVe. The results are shown in Table 3, and our CausalTransformer achieved new state-of-the-art accuracy 86.7% and Macro-F 85.2% respectively.

Models	SNLI	
	Acc	Macro-F
Transformer (Vaswani et al., 2017)	81.45	81.42
Transformer-RPR (Shaw et al., 2018)	82.20	82.18
Transformer-XL (Dai et al., 2019)	83.19	83.15
Adapted Transformer (Yan et al., 2019)	82.35	82.31
Star-Transformer (Guo et al., 2019b)	86.00	-
DA-Transformer (Wu et al., 2021)	84.18	84.16
CausalTransformer (Ours)	**86.78**	**85.56**

Table 3: Test accuracy and Macro-F on SNLI dataset.

5 Conclusion

This study implemented a Causal Transformer to improve the robustness for spurious correlations. In which, we proposed the causal attention network to elimate the useless dependencies by the Markov blanket constrained DAGs structure. To this end, (i) we defined the smooth optimization with Lagrange duality on the Markov blanket constraint; (ii) we taken the two-layer GCN as the feature updater. We further enforced this causal attention network in Transformers to implement our Causal Transformer on language and vision tasks. Our experimental results on cpurious correlation challenging datasets, neural machine translation (NMT), natural language inference (NLI) and image classification demonstrated that our Causal Transformer outperformes compared SOTA models.

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs. *URL https://www. kaggle. com/c/quora-question-pairs*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. 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 6112–6121.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. *Advances in Neural Information Processing Systems*, 32:11181–11191.

Junliang Guo, Linli Xu, and Enhong Chen. 2020. Jointly masked sequence-to-sequence model for non-autoregressive neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 376–385.

Maosheng Guo, Yu Zhang, and Ting Liu. 2019a. Gaussian transformer: a lightweight approach for natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6489–6496.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019b. Star-transformer. In *NAACL-HLT (1)*.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2020. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*.

Ruining He, Anirudh Ravula, Bhargav Kanagal, and Joshua Ainslie. 2020. Realformer: Transformer likes residual attention. *arXiv preprint arXiv:2012.11747*.

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.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. A survey of transformers. *arXiv preprint arXiv:2106.04554*.

Xu-Qing Liu and Xin-Sheng Liu. 2018. Markov blanket and markov boundary of multiple variables. *The Journal of Machine Learning Research*, 19(1):1658–1707.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Yunan Luo, Jian Peng, and Jianzhu Ma. 2020. When causal inference meets deep learning. *Nature Machine Intelligence*, 2(8):426–427.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. Flowseq: Non-autoregressive conditional sequence generation with generative flow. 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 4282–4292.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255.

Jean-Philippe Pellet and André Elisseeff. 2008. Using markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9(7).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2021. Glancing transformer for non-autoregressive neural machine translation. *ACL 2021*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1098–1108.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. 2019. Fast structured decoding for sequence models. *Advances in Neural Information Processing Systems*, 32:3016–3026.

Sho Takase and Shun Kiyono. 2021. Rethinking perturbations in encoder-decoders for fast training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5767–5780.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

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.

Yujing Wang, Yaming Yang, Jiangang Bai, Mingliang Zhang, Jing Bai, Jing Yu, Ce Zhang, and Yunhai Tong. 2020. Predictive attention transformer: Improving transformer with attention map prediction. *Eighth International Conference on Learning Representations.*

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3431–3440.

Bingzhen Wei, Mingxuan Wang, Hao Zhou, Junyang Lin, and Xu Sun. 2019. Imitation learning for non-autoregressive neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1304–1312.

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2021. Da-transformer: Distance-aware transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2059–2068.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, Timothy J Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Baosong Yang, Zhaopeng Tu, Derek F Wong, Fandong Meng, Lidia S Chao, and Tong Zhang. 2018. Modeling localness for self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4449–4458.

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

Chengxuan Ying, Guolin Ke, Di He, and Tie-Yan Liu. 2021. Lazyformer: Self attention with lazy update. *arXiv preprint arXiv:2102.12702*.

Weiqiu You, Simeng Sun, and Mohit Iyyer. 2020. Hard-coded gaussian attention for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7689–7700.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. 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 1298–1308.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018a. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018b. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. 2020. Learning sparse nonparametric dags. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425. PMLR.