

Paying More Attention to Self-attention: Improving Pre-trained Language Models via Attention Guiding

Shanshan Wang
Shandong University
Qingdao, China
wangshanshan5678@gmail.com

Zhumin Chen *
Shandong University
Qingdao, China
chenzhumin@sdu.edu.cn

Zhaochun Ren
Shandong University
Qingdao, China
zhaochun.ren@sdu.edu.cn

Huasheng Liang
WeChat, Tencent
Guangzhou, China
watsonliang@tencent.com

Qiang Yan
WeChat, Tencent
Guangzhou, China
rolanyan@tencent.com

Pengjie Ren
Shandong University
Qingdao, China
p.ren@uva.nl

ABSTRACT

Pre-trained language models (PLMs) have demonstrated their effectiveness for a broad range of information retrieval and natural language processing tasks. As the core part of PLMs, multi-head self-attention is appealing for its ability to jointly attend to information from different positions. However, researchers have found that PLMs always exhibit fixed attention patterns regardless of the input (e.g., excessively paying attention to '[CLS]' or '[SEP]'), which we argue might neglect important information in the other positions. In this work, we propose a simple yet effective attention guiding mechanism to improve the performance of PLMs through encouraging the attention towards the established goals. Specifically, we propose two kinds of attention guiding methods, i.e., the attention map discrimination guiding (MDG) and the attention pattern decorrelation guiding (PDG). The former definitely encourages the diversity among multiple self-attention heads to jointly attend to information from different representation subspaces, while the latter encourages self-attention to attend to as many different positions of the input as possible. We conduct experiments with multiple general pre-trained models (i.e., BERT, ALBERT, and Roberta) and domain-specific pre-trained models (i.e., BioBERT, ClinicalBERT, BlueBERT, and SciBERT) on three benchmark datasets (i.e., MultiNLI, MedNLI, and Cross-genre-IR). Extensive experimental results demonstrate that our proposed MDG and PDG bring stable performance improvements on all datasets with high efficiency and low cost.

CCS CONCEPTS

• **Information systems** → **Clustering and classification; Content analysis and feature selection;** • **Computing methodologies** → *Contrastive learning*.

KEYWORDS

Pre-trained models, multi-head self-attention, attention guiding, attention map discrimination, attention pattern decorrelation

ACM Reference Format:

Shanshan Wang, Zhumin Chen *, Zhaochun Ren, Huasheng Liang, Qiang Yan, and Pengjie Ren. 2022. Paying More Attention to Self-attention: Improving Pre-trained Language Models via Attention Guiding. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Pre-trained language models (PLMs) have led to tremendous performance increase in a wide range of downstream tasks, including machine translation [22], text classification [31], document ranking [26], etc. The core component of pre-trained language models (PLMs) is the self-attention mechanism, which allows the model to capture long-range dependency information.

Recently, many studies focus on analyzing the self-attention mechanism, i.e., the weights and connections of attention, to interpret the network or revealing the characteristic of PLMs [4, 10, 35]. These exploration works have found a common phenomenon: despite the success of the self-attention mechanism, these language models exhibit simple attention patterns [28, 36]. For example, Devlin et al. [6] and Kovaleva et al. [10] report the phenomenon that 40% of heads in a pre-trained BERT model simply pay attention to the delimiters, such as '[CLS]' and/or '[SEP]'. Moreover, Michel et al. [20] demonstrate that multi-headed attentions in WMT [34] and BERT [6] are not necessary to obtain competitive performance. Likewise, Raganato et al. [27] also confirm that most attentive connections in the encoder do not need to be learned at all, because most self-attention patterns learned by the transformer architecture merely reflect the positional encoding of contextual information.

The effectiveness of self-attention can be improved by introducing a variety of information. For example, Li et al. [14] demonstrate that the downstream task can be improved by increasing the diversity of attention heads. Besides, many researches focus on modifying self-attention through external information such as syntactic supervision [16, 25, 42] to improve the input representation. Their results suggest that adding additional information does help PLMs improve the effectiveness of downstream tasks. However, since these methods modify the computational process of self-attention,

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Conference'17, July 2017, Washington, DC, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-XXXX-X/18/06.
<https://doi.org/10.1145/1122445.1122456>

they must re-train the PLMs from scratch. As we all know, training a PLM with a large amount of data from scratch will take a lot of computing resources and time. In addition, extracting additional information, such as syntactic structure, will further increase the computational burden.

Therefore, we seek to investigate the following research question in this paper: is it possible to guide self-attention without extra information in the fine-tuning phrase to improve the performance of downstream tasks? As shown in Figure 1, the learned attention heads from PLMs without guiding always present similar patterns, e.g., different heads attend to similar positions. On the contrast, we seek to design an attention guiding mechanism so that comprehensive and diverse information can be taken into account. We expect the attention guiding mechanism acts as auxiliary objective to regularize the fine-tuning of downstream tasks. A similar work is done in [5]. They use several pre-defined attention patterns to guide the training of PLMs. Specifically, they add an auxiliary loss to guide the self-attention heads towards a set of pre-defined patterns (i.e., '[Next]', '[Prev]', '[First]', '[Delim]', and '[Period]'). Since these pre-defined patterns only cover a few fixed patterns and cannot introduce more information, the proposed method has limited ability in improving the diversity of attention heads. Therefore, in this work, we propose to explore the self-attention guiding methods without pre-defining attention patterns or extra knowledge about the input to encourage the diversity among multiple attention heads.

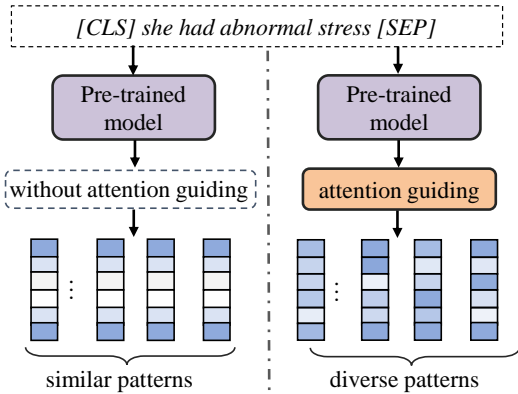


Figure 1: Illustration of attention guiding mechanism. The learned attention heads always exhibits fixed and similar patterns, which might neglect important information. Attention guiding aims to guide the learned attention heads to attend to different parts of the inputs so that more important information can be taken into account.

We propose an attention guiding mechanism to improve the performance of PLMs by regularizing its self-attention by explicitly encourage the diversity among multiple attention heads. Specifically, we propose two kinds of attention guiding methods, i.e., the attention map discrimination guiding (MDG) and the attention pattern decorrelation guiding (PDG). The former is used to encourage self-attention to attend to the information from different aspects of the inputs by diverse attention maps. An attention map is distinctive in its own right, and each could differ significantly from other

attention maps [19] so that it can capture the information of input differently with others. The latter is used to improve the diversity of attention patterns by encouraging self-attention to pay attention to more diverse positions by reducing the correlations with different attention patterns. We validate the effectiveness of the attention guiding mechanism on multiple general and domain-specific PLMs by conducting experiments on three benchmark datasets. Especially, we found that the proposed attention guiding mechanism is still effective on small-scale datasets, demonstrating its significance for low-resource settings.

Our main contributions are as follows:

- We propose two self-attention guiding terms, i.e., MDG and PDG, for guiding self-attention heads which enable PLMs to learn comprehensive and diverse attention patterns.
- We demonstrate the effectiveness of the attention guiding mechanism on seven general and domain-specific PLMs across three different datasets and tasks.

2 METHODOLOGY

2.1 Tasks

In this work, we take the following two different tasks as applications.

Task 1: Natural Language Inference. The goal of this task is to predict whether a given hypothesis can be inferred from a given promise. This task is formulated as a multi-class classification task. In the implementation, we add a classification layer on top of the '[CLS]' representation derived from the output of the last layer of PLMs, like most methods [21, 29, 46]. The PLMs are fine-tuned via minimizing the multi-class cross-entropy loss, which is defined as follows:

$$Loss_t = -\frac{1}{|D|} \sum_i^{|D|} \log(p(y_i|x_i; \theta)), \quad (1)$$

where θ denotes all trainable parameters in the PLM and the classification layer, $|D|$ is the number of training samples, and y_i is the ground truth for the i -th sample x_i .

Task 2: Across Medical Genres Querying. The objective of this task is to find the research publication that supports the primary claim made in a health-related news article. This task is formulated as a binary classification task. Similarly, we add a classification layer on top of the '[CLS]' representation derived from the output of the last layer of PLMs. The PLMs are fine-tuned via minimizing the binary cross-entropy loss as follows:

$$Loss_t = -\frac{1}{|D|} \sum_{i=1}^{|D|} [-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)], \quad (2)$$

where y_i is the ground truth for the i -th sample x_i and $\hat{y}_i = p(y_i|x_i; \theta)$ is the probability that i -th sample belongs to y_i .

2.2 Multi-head self-attention

A *pre-trained language model* (PLM) is normally a large-scale and powerful neural network trained with huge amounts of data samples and computing resources [3, 11]. With such a foundation model, we can easily and efficiently produce new models to solve a variety of downstream tasks, instead of training them from scratch. PLMs

rely on multi-head self-attention to capture dependencies between tokens [6]. Given a hidden state Z , multi-head self-attention first projects it linearly into queries Q_h , keys K_h , and values V_h using parameter matrices W_h^Q, W_h^K, W_h^V , respectively. The formulation is as follows:

$$Q_h, K_h, V_h = ZW_h^Q, ZW_h^K, ZW_h^V. \quad (3)$$

Then, the self-attention distribution A_h is computed via scaled dot-product of query Q_h and key K_h . These weights are assigned to the corresponding value vectors V_h to obtain output states O_h :

$$O_h = A_h V_h \text{ with } A_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right). \quad (4)$$

Here A_h is the attention distribution produced by the h -th attention head. d_k is the hidden size. Finally, the output states O_h of all heads are concatenated to produce the final states Z .

2.3 Attention guiding

Figure 2 shows the workflow of our methodology. Firstly, we use the pre-trained model to encode each sample x_i and obtain its corresponding attention matrix M . Then the task-objective loss is calculated based on sample representation x_i . Meanwhile, the attention matrix M will be guided by two terms, i.e., attention map discrimination guiding (MDG) and attention pattern decorrelation guiding (PDG), which aim to push the diversity of self-attention and further improve the performance of the downstream task. More specifically, for both of downstream tasks, besides the objectives, we also introduce two auxiliary objectives to guide the self-attention. Formally, the training objective of each task is revised as:

$$Loss_{total} = Loss_t + \alpha Loss_{mdg} + \beta Loss_{pdg}, \quad (5)$$

where L_{total} is the ultimate training goal, and it contains three parts of training loss. $Loss_t$ represents the task object, which varies with the task. $Loss_{mdg}$ denotes the attention map discrimination guiding (MDG) term, and $loss_{pdg}$ denotes the attention pattern decorrelation guiding (PDG) term. These two terms can be either used individually or in combination and they are regulated using α and β respectively. Note that the introduced regularization terms work like L_1 and L_2 terms which don't introduce any new parameters and only influence the fine-tuning of the standard model parameters.

2.3.1 Attention Map Discrimination. For a given sample x_i , firstly, we need to get the representation of this sample and its corresponding attention maps from the output of the pre-trained models, such as BERT. The formula can be described as:

$$h_i, \{A_1, A_2, A_i, \dots, A_{N_l * N_h}\} = PLM(x_i | \theta), \quad (6)$$

where h_i denotes the hidden state of '[CLS]' token and we regard it as the representation of sample x_i . $\{A_1, A_2, A_i, \dots, A_{N_l * N_h}\}$ is the set of multi-layer multi-head attention maps. A_i is the i -th attention map, and there are $N_l * N_h$ attention maps produced by the pre-trained model. N_l and N_h denote the layer number, and head number in each layer, respectively.

Then, we add a classification layer on top of the '[CLS]' representation h_i . Formally, the formula of the classification layer is

implemented by:

$$x_i = W \text{Relu}(h_i), \quad (7)$$

where W is the weight matrix, $\text{Relu}(\cdot)$ is the activation function, and x_i is the final representation of the given sample. By Eq.1 or Eq.2 which is determined by the downstream task, we can calculate the task-objective loss L_t with the supervision from the ground-truth label of sample x_i .

To simplify the calculation, each attention map $A_i \in \mathbb{R}^{L \times L}$ is processed as one vector $m_i \in \mathbb{R}^L$ by summing up the attention values that all tokens received. The corresponding formula of transforming the attention map A_i to the attention vector m_i is:

$$m_i = \sum_j A_{i,j}, \quad (8)$$

where i represents the i -th attention map and j is the column index of the attention map A_i .

Since the self-attention mechanism in PLM is multi-layer multi-head architecture, there are multiple attention vectors are produced. we organize all the attention vectors, into a matrix $M \in \mathbb{R}^{(N_l * N_h) \times L}$. Specifically, we concatenate all the attention vector m_i to construct the attention matrix M . Formally, the corresponding formula is as follows:

$$M = m_1 \oplus m_2 \oplus m_i, \dots, \oplus m_{N_l * N_h}, \quad (9)$$

where \oplus denotes the concatenate operation and $M \in \mathbb{R}^{(N_l * N_h) \times L}$ represents the attention matrix.

Inspired by [32, 41], we apply the instance discrimination method to push the diversity of attention maps so that the rich information of the input can be captured. The objective function is formulated based on the softmax criterion. Each attention map is assumed to represent a distinct class. That is, attention map $m_i \in \mathbb{R}^L$, i.e. the i -row of the attention matrix M , is classified into the i -th class. Accordingly, the weight vector for the i -th class can be approximated by a vector m_i . The probability of one attention map m being assigned into the i -th class is:

$$P(i|m) = \frac{\exp(m_i^\top m / \tau)}{\sum_{j=1}^{N_l * N_h} \exp(m_j^\top m / \tau)}, \quad (10)$$

where $m_j^\top m$ measures how well m matches the j -th class because m_j is regarded as the weight of j -th class. τ is a temperature parameter that controls the concentration of the distribution [8], and m is normalized to $\|m\| = 1$. The objective maximizes the joint probability $\prod_{i=1}^{N_l * N_h} P_\theta(i | f_\theta(m_i))$ as

$$\begin{aligned} Loss_{mdg} &= - \sum_{i=1}^{N_l * N_h} \log P(i | f_\theta(m_i)), \\ &= - \sum_i^{N_l * N_h} \log \left(\frac{\exp(m_i^\top m_i / \tau)}{\sum_{j=1}^{N_l * N_h} \exp(m_j^\top m_i / \tau)} \right). \end{aligned} \quad (11)$$

2.3.2 Attention Pattern Decorrelation. We have analyzed that the multi-head attention heads are likely to suffer from the redundancy problem where each attention vector focuses on a very similar region. To encourage each attention head to capture the information from different positions of the inputs, at the same time, we propose another attention guiding term, i.e., the *attention pattern decorrelation guiding* (PDG).

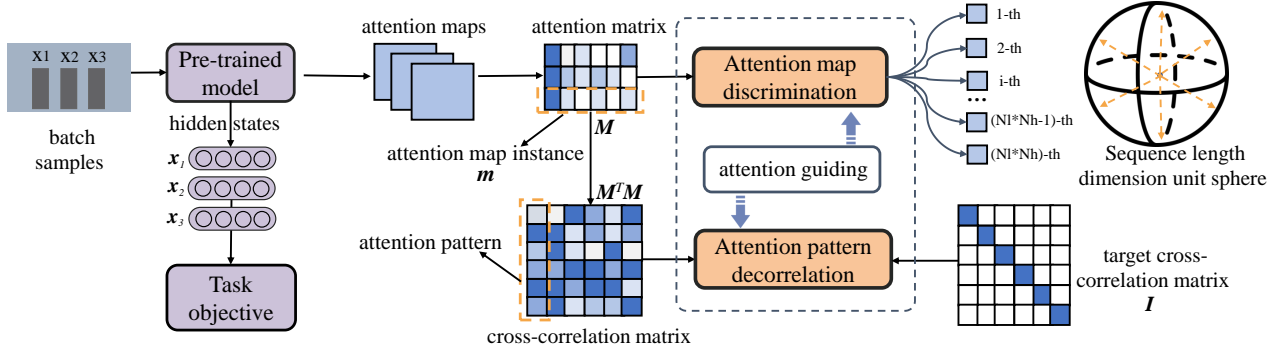


Figure 2: The pipeline of the proposed PLM with attention guiding approach. The attention guiding mechanism contains *attention map discrimination guiding* (MDG) and *attention pattern decorrelation guiding* (PDG), respectively. For a batch sample, we use the pre-trained model to encode each sample as a feature vector x_i , and as well as we obtain the attention matrix M for each sample. The optimal feature embedding is learned via task objective plus the losses produced by two attention guiding methods, i.e., MDG and PDG. Both MDG and PDG try to maximally scatter the attention heads of training samples over the sequence length unit sphere.

Inspired by [15], we regard the l -th column of M , i.e., $M_l^\top \in \mathbb{R}^{N_l \times N_h}$ as the soft representation of the l -th attention pattern. Conventionally, attention patterns should be independent to ensure that redundant information is reduced. The objective function is applied to push the diversity of attention patterns and reduce pattern redundancy, which tries to make the cross-correlation matrix computed from the attention matrix as close to the identity matrix as possible.

The formula of PDG term aims to construct independent attention patterns and is as follows:

$$Loss_{pdg} = \|M^\top M - I\|_F^2, \quad (12)$$

where $M \in \mathbb{R}^{(N_l \times N_h) \times L}$ is the attention matrix, $I \in \mathbb{R}^{L \times L}$ is the identity matrix and $\|\cdot\|_F^2$ denotes the squared Frobenius Norm [47]. $M^\top M \in \mathbb{R}^{L \times L}$ can be seen as the cross-correlation matrix of different attention patterns. Minimizing the difference between the cross-correlation matrix and the identity matrix I is equivalent to making the attention patterns diverse so that they focus on different tokens [23, 48].

3 EXPERIMENTAL SETUP

To evaluate the effectiveness of our proposed attention guiding mechanism, we conduct extensive experiments of a variety of pre-trained models on different downstream tasks. We demonstrate that the attention guiding mechanism can promote the diversity of attention heads and further improve the performance on different downstream tasks.

3.1 Datasets and evaluation

We conduct experiments on the following datasets.

- **MultiNLI**¹ [39] is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information, i.e., entailment, contradiction, and neutral. This dataset is for the natural

language inference task, which is also popular for evaluating various PLMs [5, 37]. Accuracy (ACC for short) is standard metric on this task. At the same time, we also report other metrics commonly used in the classification tasks, such as Precision, Recall and F1.

- **MedNLI**² [30] is for natural language inference in clinical domain, which has the same data structure as MultiNLI. Accuracy is also the standard metric on this dataset. Like on the MultiNLI, we also report the Precision, Recall and F1.
- **Cross-genre-IR**³ [49] is for the across medical genres querying task, where each claim (i.e., the news headline) is associated with at least one peer-reviewed research publication supporting it. For each claim, it needs to re-rank the candidate publications to obtain the correct ones. Following the original authors, we report the Mean Reciprocal Rank (i.e., MRR) and Recall@K (i.e., R@K = 1, 3, 5, 20) metrics.

3.2 PLMs for comparison

We consider seven transformer-based PLMs: three are pre-trained over general language corpora (BERT, ALBERT, and Roberta) and four are pre-trained over biomedical corpora (BioBERT, ClinicalBERT, BlueBERT, and SciBERT).

- **BERT**⁴ [6] is a multi-layer bidirectional Transformer encoder. Since the following versions of the PLMs are often based on the BERT-base-uncased version (12 layers and 768 hidden size with 108M parameters), we use the BERT-base-uncased here for a fair comparison.
- **Roberta**⁵ [18] has the same architecture as BERT, but with a lot of changes on the training mechanism, such as a more random mask mechanism. We use the Roberta-base here for comparison.

² <https://physionet.org/content/mednli/1.0.0/>

³ <https://github.com/chzuo/emnlp2020-cross-genre-IR>

⁴ <https://huggingface.co/bert-base-uncased>

⁵ <https://huggingface.co/roberta-base>

¹ <https://cims.nyu.edu/~sbowman/multinli/>

Table 1: Performance comparison (%) of PLMs with or without attention guiding. Bold face indicates the improved performance with attention guiding (AG) in terms of the corresponding metrics. * means $p < 0.05$ and ** means $p < 0.01$ in t-test, respectively. SOTA represents the best performing methods by Nov. 2021, to the best of our knowledge. The matched test set of MultiNLI for the experiments.

Methods	MultiNLI				MedNLI				Cross-genre-IR				
	ACC	Precision	Recall	F1	ACC	Precision	Recall	F1	MRR	R@1	R@3	R@5	R@20
SOTA	87.90	–	–	–	84.00	–	–	–	69.50	62.50	74.30	78.30	82.80
BERT	83.24	83.19	83.17	83.17	76.02	76.25	76.01	76.11	73.18	62.84	80.38	85.07	94.58
BERT+AG	83.89	83.91	83.82	83.83	77.22*	77.35*	77.24*	77.26*	73.38	63.69	80.74	84.59	94.09
ALBERT	83.09	83.05	83.04	83.02	77.76	78.04	77.78	77.84	71.00	62.06	76.07	80.02	93.21
ALBERT+AG	84.41**	84.44**	84.37**	84.37**	78.76**	78.72**	78.80**	78.74**	71.86*	62.06	78.08*	82.95*	93.42*
Roberta	85.95	85.91	85.94	85.91	80.17	80.12	80.12	80.14	78.67	71.04	83.58	86.85	94.86
Roberta+AG	86.77*	86.67*	86.72*	86.67*	80.31	80.36	80.31	80.32	79.88*	72.05*	84.47*	87.70*	95.68*
BioBERT	81.37	81.35	81.33	81.28	81.86	81.81	81.84	81.83	83.72	76.48	88.83	92.52	97.21
BioBERT+AG	81.54**	81.46**	81.47**	81.45**	82.63	82.68	82.65	82.64	83.76	75.87	90.17	92.88	96.84
ClinicalBERT	79.68	79.81	79.69	79.66	80.73	80.67	80.71	80.66	71.83	61.99	78.30	83.58	92.03
ClinicalBERT+AG	80.07**	80.13**	80.07**	80.09**	82.63*	82.68*	82.67*	82.64*	72.02*	62.24*	78.18	83.70*	91.91
BlueBERT	79.13	79.10	79.10	79.06	83.90	83.94	83.91	83.92	69.68	59.65	75.67	82.08	91.75
BlueBERT+AG	79.21*	79.17*	79.17*	79.13*	84.32**	84.46**	84.33**	84.34**	71.02*	61.47*	77.61*	81.53	91.38
SciBERT	83.61	83.68	83.56	83.58	79.47	79.42	79.47	79.43	80.72	72.80	86.65	90.57	96.48
SciBERT+AG	83.74*	83.79*	83.70*	83.71*	80.03*	80.03*	80.11*	83.01*	81.67*	72.84*	88.07*	91.50*	97.41*

- **ALBERT**⁶ [12] compresses the architecture of BERT by factorized embedding parameterization and cross-layer parameter sharing. We use the ALBERT-base-v2 version.
- **BioBERT**⁷ [13] is the first BERT pre-trained on biomedical corpora. It is initialized with BERT’s pre-trained parameters and then further pre-trained over PubMed abstracts and full-text articles. We use the best version BioBERT V1.1.
- **ClinicalBERT**⁸ [1] is initialized from BioBERT v1.0 and further pre-trained over approximately 2 million notes in the MIMIC-III v1.4 database.
- **BlueBERT**⁹ [24] is firstly initialized from BERT and further pre-trained over biomedical corpus of PubMed and clinical notes.
- **SciBERT**¹⁰ [2] is a BERT-base model pre-trained on 1.4M papers from the semantic scholar, with 18% of papers from the computer science and 82% from the biomedical domain.
- **SOTA.** We also compare with state-of-the-art methods on each dataset, which are based on Roberta, BlueBERT and BERT, to the best of our knowledge [24, 33, 49].

3.3 Implementation details

The proposed attention guiding mechanism acts on all attention heads out from PLMs. We fine-tune all PLM models for 5 epochs, 20 epochs, and 5 epochs on the MultiNLI, MedNLI, and Cross-genre-IR datasets, respectively. The hidden size d_k is 768 and sequence length L is set to 256 of each PLM. We use the Adam optimizer (learning

rate $1e-05$) for all models and the batch size is set as the maximum according to the memory of a GeForce RTX 3090 GPU. Specifically, the batch size of ALBERT is set to 56 and other PLMs is 64 on different datasets. The α and β in Eq. 5 are selected from the set $\{0.1, 0.01, 0.001, 0.0001\}$ according to grid search. The temperature parameter in Eq. 11 is set to 1.0.

3.4 Results on different pre-trained models

The results of all PLMs on different tasks are listed in Table 1. From the results, we have several observations.

First, the proposed AG can improve the performance of the PLMs on all tasks generally. For example, on the MultiNLI and MedNLI datasets, all the selected pre-trained models with attention guiding can promote the performance of downstream tasks on all PLMs in terms of all metrics. Similarly, on the Cross-genre-IR dataset, most of the metrics of the task can be promoted by our attention guiding method. Moreover, the encouraging findings are that simply adding AG to BlueBERT (i.e., BlueBERT+AG) outperforms SOTA on the MedNLI dataset, and BioBERT+AG is far better than SOTA on the Cross-genre-IR dataset. These figures show the effectiveness of the proposed attention guiding. The reason why AG works is that the self-attention after guiding has a better ability to attend to broader and more diverse tokens, which benefit for downstream tasks.

Second, AG plays different roles on different models and datasets. For example, the proposed attention mechanism always improves the performances on MultiNLI and MedNLI datasets, while on the Cross-genre-IR dataset, some metrics drop slightly, e.g., R@20 drops from 94.58% to 94.09% of BERT after attention guiding. Moreover, we also observe that the performances of different PLMs are always improved in terms of MRR, and more PLMs can be improved when

⁶ <https://huggingface.co/albert-base-v2>

⁷ <https://huggingface.co/dmis-lab/biobert-v1.1>

⁸ https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

⁹ <https://huggingface.co/ttuyche/bluebert>

¹⁰ https://huggingface.co/nfliu/scibert_basevocab_uncased

K is small in terms of R@K metrics. This suggests that the attention guiding method may be influenced by some characteristics of the evaluation metrics. For instance, the R@20 is difficult to be promoted by attention guiding. However, although there is a slight decrease in some metrics, in most cases, our attention guiding method can improve the pre-trained model effectively.

Third, according to our results, AG plays a bigger role on small datasets. For example, the biggest improvements of AG reach 3.58% in terms of F1 and 2.95% in terms of R@5 on the MedNLI (11k sentence pairs) and Cross-genre-IR (48k sentence pairs) datasets, respectively, which are greater than these on the MultiNLI (443k sentence pairs) dataset, i.e., 0.98%. To further explore this phenomenon, we vary the training size of MedNLI dataset to evaluate AG's role, and the details refer to §4.3.

3.5 Comparison with different attention guiding methods

We also study how the attention guiding mechanism compared with other related works. Since all the related works need to re-train a new model, rather than our work acts in the fine-tuning phase of the pre-trained model. Therefore, it is difficult to compare them directly. For comparison, we implement different methods to guide self-attention during the fine-tuning phase of pre-trained models. Specifically, we implemented five kinds of attention guiding patterns (i.e., '[Next]', '[Prev]', '[First]', '[Delim]', and '[Period]') proposed by Deshpande and Narasimhan [5]. In addition, we also implemented three kinds of attention guiding methods based on the knowledge extracted from the input. Specifically, the guiding methods based on knowledge are guiding attention heads with co-occurrence relationships, syntactic dependencies, and similar relationships between words/tokens, namely '[PMI]', '[Dependency]' and '[WordSim]' respectively. These prior pieces of knowledge about self-attention we used are proposed by Qi et al. [25] and Xia et al. [42]. The results of these different guiding methods are reported in Table 2, and Table 2 shows that:

First, no matter which kind of guidance method can improve the effectiveness of the pre-trained model, i.e., BERT, in varying degrees. For example, the ACC of BERT on the MultiNLI dataset increases from 83.24% to 83.73% after the guidance of word similarity knowledge (i.e., '[WordSim]'). The results also show that self-attention heads need to be guided, as long as the guiding method is reasonable, such as using some fixed common attention patterns or using the knowledge derived from the input, the performance of mainstream tasks can also be improved.

Second, although three types of attention guiding methods, proposed by ours, Deshpande and Narasimhan [5], Qi et al. [25] and Xia et al. [42] can improve the performances of PLMs in mainstream tasks, different guiding methods play different roles in different datasets. Our AG is superior to other methods on almost all metrics, such as the ACC on MedNLI and MultiNLI datasets. But on the Cross-genre-IR datasets, '[WordSim]' method is better than ours in terms of most metrics. This suggests that the effect of different attention guiding methods may be affected by the dataset, and trying one or more attention guiding methods on a specific dataset may maximize the effectiveness of the pre-trained models. How

to choose appropriate prior knowledge or guidance methods for self-attention may be a problem that needs further exploration.

4 ANALYSIS

4.1 Ablation study

To analyze where the improvements of AG come from, we conduct an ablation study of the best model on each dataset. Obviously, it is much easier to comprehend that models pre-trained by medical corpus are better at handling medical-related tasks (i.e., MedNLI and Cross-genre-IR). Therefore, the best pre-trained models on MultiNLI, MedNLI, and Cross-genre-IR datasets are Roberta, BlueBERT, and BioBERT respectively. The detailed results are shown in Table 3. We consider the following three settings: (1) **w/o-AG** denotes PLMs without AG, i.e., baseline PLMs. (2) **w/o-MDG** denotes PLMs without the MDG term but reserving the PDG term. (3) **w/o-PDG** denotes removing the PDG term but reserving the MDG term.

The results in Table 3 show that MDG and PDG are helpful for PLMs as removing either of them leads to a decrease in performance in almost all the metrics. Besides, on the MedNLI and Cross-genre-IR datasets, the most obvious declines are the variants removing PDG, i.e., w/o-PDG. This illustrates that the attention pattern decorrelation guiding can bring more valuable attention information for PLM because PDG can push each head to attend to different positions of the inputs to capture diversity information.

We also note that the R@1 on the cross-genre-IR dataset declines slightly, e.g., the R@1 drops from 76.48% to 75.87% with AG. Nevertheless, AG is still effective as the other metrics (i.e., MRR, R@3, R@5, and R@20) still get improvements after the AG mechanism.

4.2 Effect of AG on different layers

As the proposed AG (i.e., MDG and PDG) can be applied to any layers of PLMs, we design experiments to see their effect on different layers of BERT. The results of BERT with MDG, BERT with PDG, and BERT with both of MDG and PDG on different layers are summarized in Figure 3. The blue dashed line indicates BERT without AG.

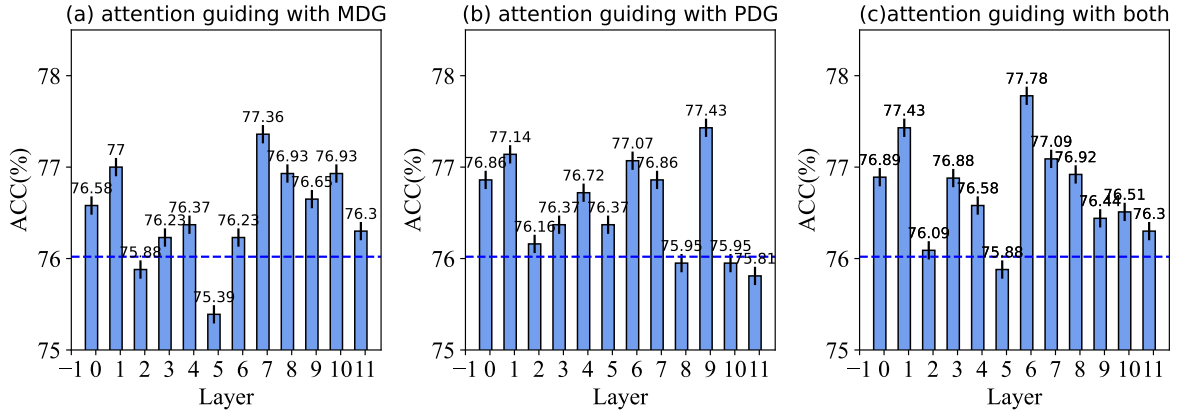
From Figure 3, we can see that most layers can benefit from MDG and PDG obviously, such as the ACC increases from 76.02% to 77.36% at layer 7 after being guided by MDG. And similarly, the ACC is improved from 76.02% to 77.43% at layer 9 after being guided by PDG. Moreover, lower and middle layers can always benefit from PDG, while for top layers, there are some declines occasionally. For example, at layer 11, the ACC drops from 76.02% to 75.81% after being guided by PDG. On the contrary, some declines happen at the lower and middle layers of BERT with MDG, e.g., the ACC of BERT is down 0.63% at layer 5. That is understandable as the functions of MDG and PDG are different. The MDG focuses on distinct attention heads, while PDG pushes each attention head to attend to different tokens/positions of the inputs. So combining MDG and PDG generally leads to an improvement of BERT on almost all layers. For example, the performances increase by 0.21% and 1.51%, after being guided by the MDG and PDG separately at layer 6, while the improvement reaches 1.76% after combining MDG and PDG. Moreover, lower layers were found to perform broad attention across all pairs of tokens [4]. Therefore, lower layers call

Table 2: Performance comparison (%) of PLMs with different guiding methods. Significant improvements over the best baseline results are marked with * (t-test, $p < 0.05$).

Methods	MultiNLI				MedNLI				Cross-genre-IR				
	ACC	Precision	Recall	F1	ACC	Precision	Recall	F1	MRR	R@1	R@3	R@5	R@20
BERT	83.24	83.19	83.17	83.17	76.02	76.25	76.01	76.11	73.18	62.84	80.38	85.07	94.58
BERT+AG	83.89*	83.91*	83.82*	83.83*	77.22*	77.35	77.24*	77.26*	73.38	63.69	80.74	84.59	94.09
+ [First]	83.65	83.63	83.58	83.59	76.02	76.38	76.01	76.13	72.81	62.24	80.24	85.44	94.70
+ [Next]	83.71	83.74	83.64	83.66	76.93	77.23	76.93	77.04	72.84	63.79	77.69	84.47	93.61
+ [Prev]	83.65	83.67	83.57	83.59	77.07	77.52	77.06	77.15	70.98	60.60	77.53	83.39	93.37
+ [Period]	83.70	83.71	83.64	83.65	76.02	76.38	76.01	76.13	73.85	64.40	79.87	85.56	93.73
+ [Delim]	83.65	83.67	83.59	83.60	75.67	75.71	75.66	75.69	70.97	61.69	76.31	82.16	90.86
+ [PMI]	83.68	83.69	83.61	83.62	77.07	77.60	77.05	77.16	72.41	62.61	79.11	84.29	93.12
+ [Dependency]	83.59	83.61	83.53	83.55	76.37	76.61	76.37	76.46	73.10	64.37	78.38	82.76	92.57
+ [WordSim]	83.73	83.74	83.67	83.68	76.72	77.03	76.72	76.81	74.15	64.48	80.91	85.80	94.22

Table 3: Performance comparison (%) of different methods. Bold face indicates the best results in terms of the corresponding metrics. The matched test set of MultiNLI for the experiments. Significant improvements over the best baseline results are marked with * (t-test, $p < 0.05$).

Variants	MultiNLI				MedNLI				Cross-genre-IR				
	Roberta				BlueBERT				BioBERT				
	ACC	Precision	Recall	F1	ACC	Precision	Recall	F1	MRR	R@1	R@3	R@5	R@20
Default	86.77*	86.67*	86.72*	86.67*	84.32*	84.46*	84.33*	84.34*	83.76	75.87	90.17	92.88	96.84
w/o-AG	85.95	85.91	85.94	85.91	83.90	83.94	83.91	83.92	83.72	76.48	88.83	92.52	97.21
w/o-MDG	86.52	86.44	86.49	86.43	84.18	84.21	84.18	84.15	83.48	75.99	88.63	92.52	96.84
w/o-PDG	86.57	86.48	86.52	86.48	82.98	83.15	82.98	83.02	83.28	75.63	88.83	92.76	96.72


Figure 3: Performances of each BERT layer with the proposed AG mechanism. Figure (a), (b) and (c) demonstrate the results of using MDG, PDG, and both of MDG and PDG to guide each layer of BERT, respectively.

for the guidance of PDG to promote the diversity of tokens rather than MDG, compared to the middle and top layers.

4.3 Effect of AG on different training sizes

In Table 1, we found that the proposed AG brings more improvements on MedNLI and Cross-Genre-IR datasets than on the MultiNLI dataset. To explore whether our AG mechanism is affected by the size of the training datasets, we randomly select 20% to 100% data

from the training set of MedNLI for fine-tuning. The detailed results are illustrated in Figure 6. We have the following observations in this experiment.

The PDG and MDG can improve BERT at different training sizes generally, even though the data size is small. Specifically, when only 20% of the training set is used, the MDG increases the Accuracy by 0.84% (i.e., from 68.78% to 69.62%) while PDG also improves the Accuracy by 0.70% (from 68.78% to 69.48%). The reasons of MDG

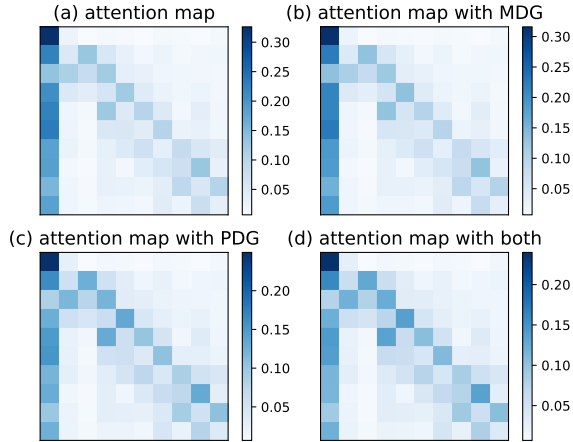


Figure 4: Attention heatmap for a random sample. Figure (a) represents attention without AG, and Figure (b)-(d) represent attention with MDG, with PDG, and with both of MDG and PDG, respectively.

and PDG are effective on small datasets are that when there is not enough training data, it is difficult for PLMs to be adapted to a different task. In other words, the self-attention is not well fine-tuned. Thus, the guidance of self-attention becomes particularly important. Moreover, it is easy for PLMs to over-fit on small datasets. PDG and PDG could help in some ways to alleviate such over-fitting issues.

4.4 Case study with attention heatmap

We visualize the changes of attention heads with/without the guidance of MDG and PDG to see whether there is any diverse attention pattern after guidance. We randomly select one sample from the test set of the MedNLI dataset and visualize the average attention map from all heads at all layers of BERT in Figure 4. To make it easy for observation, Figure 4 only shows the first 15 tokens in the sequence, and the selected token sequence is “[‘[CLS]’, ‘the’, ‘patient’, ‘denied’, ‘any’, ‘headache’, ‘,’ ‘visual’, ‘changes’, ‘,’ ‘chest’, ‘pain’, ‘,’ ‘pl’, ‘##eur’]”.

As shown in Figure 4(d), the attention map with both MDG and PDG could pay attention to more positions compared with Figure 4(a) which is not guided by AG. For example, more attention is paid to the token in the last column (i.e., ‘##eur’), which is overlooked by the attention map without AG in Figure 4(a). In fact, the token ‘##eur’ and the previous token ‘pl’ constitute an important medical concept ‘pleur’, which should be paid more attention. AG can make such kinds of tokens get more attention, which is why PLMs can be improved by AG.

4.5 Case study with attention principal component analysis

In order to explore whether our AG mechanism promotes the richness and diversity of self-attention, we randomly select some samples from the test set of MedNLI dataset and perform dimensionality reduction through Principal Component Analysis (PCA) [9] on BERT’s all attention heads from all layers. Figure 5 shows the

Table 4: Per-epoch training time (in seconds) on the three datasets using different PLMs with or without the proposed AG.

	MultiNLI	MedNLI	Cross-genre-IR
BERT	2190	95	339
BERT+AG	2334	112	361
ALBERT	2639	96	401
ALBERT+AG	2770	101	423
Roberta	2190	82	340
Roberta+AG	2344	87	354
BioBERT	2225	82	341
BioBERT+AG	2318	87	361
ClinicalBERT	2228	82	339
ClinicalBERT+AG	2372	85	355
BlueBERT	2210	100	340
BlueBERT+AG	2324	102	359
SciBERT	2215	81	338
SciBERT+AG	2345	86	360

spatial distribution of each attention head with/without the AG mechanism. From Figure 5, we can see that the attention distributions with MDG, PDG and both of them ((b)-(d) of Figure 5) are more dispersed than the distribution without AG (see Figure 5(a)). This suggests that the proposed AG mechanism (MDG and PDG included) is effective, and AG does encourage self-attention to pay attention to wider positions of the sequence.

Moreover, the distribution of multi-headed attention in Figure 5(b) (i.e., being guided by MDG) is more scattered than in Figure 5(c) (i.e., being guided by PDG). Obviously, the reason is that MDG is designed to push the diversity of different attention maps which will lead to scattered attention heads.

4.6 Time cost analysis

Most of the previous studies directly modify the computation process of self-attention, e.g. [25, 42], which means that they need to re-train the PLMs. On the contrast, our method works in the fine-tuning phrase, and does not need to re-train the PLMs. Thus, our AG also has merits in terms of time cost.

Nevertheless, the calculation process of AG will take more time than directly fine-tuning the pre-trained models on specific datasets. Table 4 shows the per-epoch training time of different PLMs with or without our AG on three datasets.

As can be seen from Table 4, the increased time cost is minor by adding AG. Specifically, the extra time cost by AG per-epoch training is about 130 seconds, 7 seconds and 19 seconds on MultiNLI, MedNLI and Cross-genre-IR datasets, respectively. We consider the time cost acceptable, since AG can improve different pre-trained models significantly.

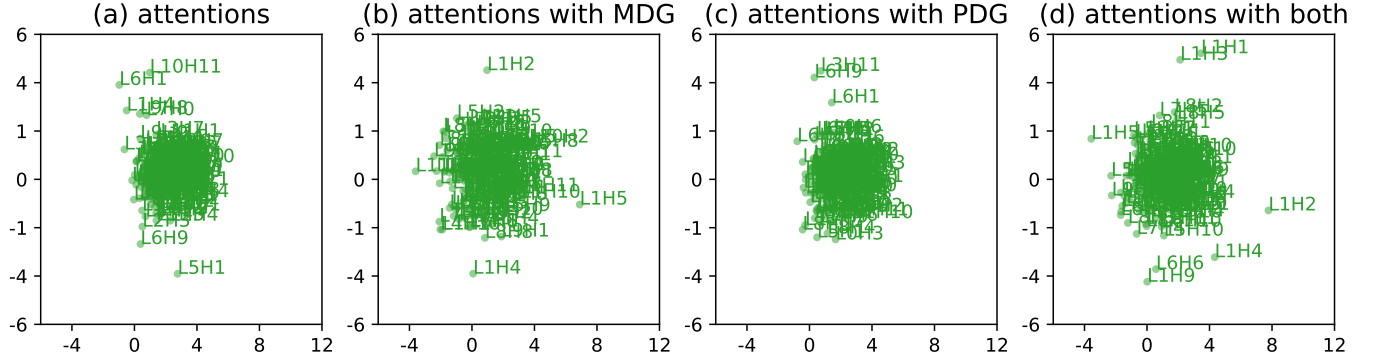


Figure 5: Attention principal component analysis for a random sample. Figure (a) denotes the spatial distribution of multi-layer multi-head attentions, and Figure (b)-(d) represent the attentions with MDG, with PDG, and with both of MDG and PDG respectively.

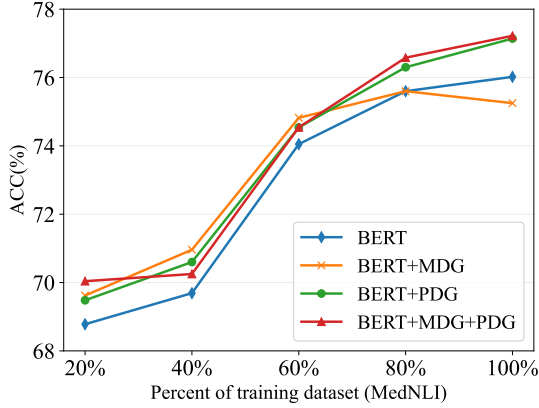


Figure 6: Performance of BERT with AG on different amounts of training data.

5 RELATED WORK

Existing studies on self-attention can be roughly classified into three groups: self-attention probing, self-attention revising and self-attention guiding.

5.1 Self-attention probing

This line of research focuses on the pattern probing of self-attention, i.e., the analysis of the interpretability of the weights and connections in particular. For example, Vig and Belinkov [35] visualize attentions and analyze the interaction between attention and syntax over a large corpus. They find that different attentions target different parts of speech at different layers of the model, and that the attentions align with the dependency relations closely, especially in the middle layers. Similarly, Clark et al. [4] demonstrate through visual analysis and statistical analysis that the substantial syntactic information is captured in BERT’s attentions. Kovaleva et al. [10] summarize 5 kinds of frequent attention patterns, called vertical, diagonal, vertical+diagonal, block, and heterogeneous, respectively. Voita et al. [36] identify the most important heads in each encoder layer using layer-wise relevance propagation, and then attempt to characterize the roles they perform. Wang et al.

[38] present DODRIO, an open-source interactive visualization tool to help researchers and practitioners analyze attention mechanisms to meet the needs of self-attention visualization.

5.2 Self-attention revising

This line of research modifies the attention formula to bias the attention weights towards local areas [7, 43, 45]. For example, Wu et al. [40] and Yang et al. [44] use convolutional modules to replace self-attentions in some parts, making the networks computationally more efficient. Raganato et al. [27] design seven predefined patterns, each of which takes the place of an attention head to train Neural Machine Translation models without the need of learning them. The advantage of this method is that it can reduce the parameter footprint without loss of translation quality. Qi et al. [25] consider direction mask, word distance mask, and dependency distance mask simultaneously, and add them into the attention calculation to obtain the structural priors. Similarly, Li et al. [17] map each token into a tree node, and calculate the distance of any two nodes, after which the distance is added to the attention calculation. Xia et al. [42] inject word similarity knowledge into the attention calculation to make the BERT model aware of the word pair similarity.

5.3 Self-attention guiding

Different from the above two research lines, our work belongs to self-attention guiding, which guides the learning of self-attention without introducing any new parameters or modifying the attention calculation formulas. Deshpande and Narasimhan [5]’s work belongs to this category. In their work, five fixed patterns are predefined based on the analyses of attentions, based on which a regularization term is added to force the attentions to approach the predefined attention patterns in the training phase of PLMs. There are at least two differences compared with our work. First, we do not need to predefine attention patterns. Instead, the attention is guided adaptively through the MDG and PDG parts. Second, we do not need to train PLMs from scratch. Our attention guiding method works in the fine-tuning phase of the PLMs. We compared with Deshpande and Narasimhan [5]’s work in Table 2, demonstrating

that our method achieves comparable or better performance without introducing new knowledge or predefining attention patterns beforehand.

6 CONCLUSION AND FUTURE WORK

In this work, we have proposed two kinds of attention guiding methods, i.e., the attention map discrimination guiding (MDG) and the attention pattern decorrelation guiding (PDG), to improve the performance of PLMs by encouraging the learned attentions to derive more information from the inputs and to be more diverse. Experimental results of seven PLMs on three datasets have validated the effectiveness of our proposed methods. Especially, we have found that the proposed attention guiding mechanism works on small datasets and large datasets, which is attractive as building large labeled dataset is time consuming and labor intensive.

As to future work, we plan to explore how to incorporate more domain-specific knowledge to guide self-attention learning in low-resource domains, e.g., the relations of diseases, drugs, and symptoms in medical domains.

REPRODUCIBILITY

This work uses publicly available data. To facilitate the reproducibility of the reported results, we release the code at <https://anonymous.4open.science/r/attentionGuiding-F6C0>.

REFERENCES

- [1] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. *CoRR* abs/1904.03323 (2019).
- [2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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*. 3613–3618.
- [3] Monojit Choudhury and Amit Deshpande. 2021. How Linguistically Fair Are Multilingual Pre-Trained Language Models?. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*. 12710–12718.
- [4] 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, BlackboxNLP@ACL*. 276–286.
- [5] Ameesh Deshpande and Karthik Narasimhan. 2020. Guiding Attention for Self-Supervised Learning with Transformers. In *Findings of the Association for Computational Linguistics, EMNLP*. 4676–4686.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 4171–4186.
- [7] José A. R. Fonollosa, Noe Casas, and Marta R. Costa-jussà. 2019. Joint Source-Target Self Attention with Locality Constraints. *CoRR* abs/1905.06596 (2019).
- [8] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015).
- [9] Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2065 (2016), 20150202.
- [10] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. 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*. 4364–4373.
- [11] Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! Rewarding Pre-Trained Models Improves Formality Style Transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*. 484–494.
- [12] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR*.
- [13] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.* 36, 4 (2020), 1234–1240.
- [14] Jian Li, Zhaopeng Tu, Baosong Yang, Michael R. Lyu, and Tong Zhang. 2018. Multi-Head Attention with Disagreement Regularization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 2897–2903.
- [15] Yunfan Li, Peng Hu, Jerry Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive Clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*. 8547–8555.
- [16] Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2020. Improving BERT with syntax-aware local attention. *arXiv preprint arXiv:2012.15150* (2020).
- [17] Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2021. Improving BERT with Syntax-aware Local Attention. In *Findings of the Association for Computational Linguistics, ACL/IJCNLP*. 645–653.
- [18] Yinhan Liu, Mylène 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. *CoRR* abs/1907.11692 (2019).
- [19] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. 2011. Ensemble of exemplar-SVMs for object detection and beyond. In *IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, 89–96.
- [20] Paul Michel, Omer Levy, and Graham Neubig. 2019. Are Sixteen Heads Really Better than One?. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*. 14014–14024.
- [21] Elias Moons, Aditya Khanna, Abbas Akkasi, and Marie-Francine Moens. 2020. A comparison of deep learning methods for ICD coding of clinical records. *Applied Sciences* 10, 15 (2020), 5262.
- [22] Mylène Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT*. 1–9.
- [23] Geondo Park, Chihye Han, Daeshik Kim, and Wonjun Yoon. 2020. MHSAN: Multi-Head Self-Attention Network for Visual Semantic Embedding. In *IEEE Winter Conference on Applications of Computer Vision, WACV*. IEEE, 1507–1515.
- [24] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL*. 58–65.
- [25] Le Qi, Yu Zhang, Qingyu Yin, and Ting Liu. 2020. Multiple Structural Priors Guided Self Attention Network for Language Understanding. *CoRR* abs/2012.14642 (2020).
- [26] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. *CoRR* abs/1904.07531 (2019).
- [27] Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. Fixed Encoder Self-Attention Patterns in Transformer-Based Machine Translation. In *Findings of the Association for Computational Linguistics, EMNLP*. 556–568.
- [28] Alessandro Raganato, Jörg Tiedemann, et al. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP*.
- [29] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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*. 3980–3990.
- [30] Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 1586–1596.
- [31] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification?. In *Chinese Computational Linguistics - 18th China National Conference, CCL*, Vol. 11856. 194–206.
- [32] Yaling Tao, Kentaro Takagi, and Kouta Nakata. 2021. Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation. In *9th International Conference on Learning Representations, ICLR 2021*.
- [33] Ehsan Tavan, Ali Rahmati, Maryam Najafi, Saeed Bibak, and Zahed Rahmati. 2021. BERT-DRE: BERT with Deep Recursive Encoder for Natural Language Sentence Matching. *CoRR* abs/2111.02188 (2021).
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*. 5998–6008.
- [35] 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, BlackboxNLP@ACL*. 63–76.
- [36] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*. 5797–5808.

- [37] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pre-trained Transformers. In *Findings of the Association for Computational Linguistics, ACL/IJCNLP (Findings of ACL, Vol. ACL/IJCNLP 2021)*. 2140–2151.
- [38] Zijie J. Wang, Robert Turko, and Duen Horng Chau. 2021. Dodrio: Exploring Transformer Models with Interactive Visualization. *CoRR* abs/2103.14625 (2021).
- [39] Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 1112–1122.
- [40] Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay Less Attention with Lightweight and Dynamic Convolutions. In *7th International Conference on Learning Representations, ICLR*.
- [41] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*. 3733–3742.
- [42] Tingyu Xia, Yue Wang, Yuan Tian, and Yi Chang. 2021. Using Prior Knowledge to Guide BERT's Attention in Semantic Textual Matching Tasks. In *WWW '21: The Web Conference 2021*. 2466–2475.
- [43] Mingzhou Xu, Derek F. Wong, Baosong Yang, Yue Zhang, and Lidia S. Chao. 2019. Leveraging Local and Global Patterns for Self-Attention Networks. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*. 3069–3075.
- [44] Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019. Convolutional Self-Attention Networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 4040–4045.
- [45] Shan Yang, Heng Lu, Shiyin Kang, Liumeng Xue, Jinba Xiao, Dan Su, Lei Xie, and Dong Yu. 2020. On the localness modeling for the self-attention based end-to-end speech synthesis. *Neural Networks* 125 (2020), 121–130.
- [46] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. In *SIGIR '21: The 44th International ACM Conference on Research and Development in Information Retrieval, SIGIR*. 2666–2668.
- [47] Shi-Fang Yuan, Yi-Bin Yu, Ming-Zhao Li, and Hua Jiang. 2020. A direct method to Frobenius norm-based matrix regression. *Int. J. Comput. Math.* 97, 9 (2020), 1767–1780.
- [48] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning, ICML, Vol. 139*. 12310–12320.
- [49] Chaoyuan Zuo, Narayan Acharya, and Ritwik Banerjee. 2020. Querying across genres for medical claims in news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*. 1783–1789.