Implement " What does BERT Look At ? An Analysis of BERT's Attention." on XLNet or RoBERTa. ...

Intent Detection

Nutapol Thungpao st122148
Arnajak tungchoksongchai st122458
Praewphan Tocharoenkul st122497

Already done those paper

Nutapol

- Attention is all you need
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Emergent linguistic structure in artificial neural networks trained by self-supervision

Arnajak

- Attention is all you need
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning

Praewphan

- Attention is all you need
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- What Does BERT Look At? An Analysis of BERT's Attention

Emergent linguistic structure in artificial neural networks trained by self-supervision

This paper explores the knowledge of linguistic structure learned by large artificial neural networks, trained via self-supervision, whereby the model simply tries to predict a masked word in a given context. They develop methods for identifying linguistic hierarchical structure emergent in artificial neural networks and demonstrate that components in these models focus on syntactic grammatical relationships and anaphoric coreference.

Emergent linguistic structure in artificial neural networks trained by self-supervision

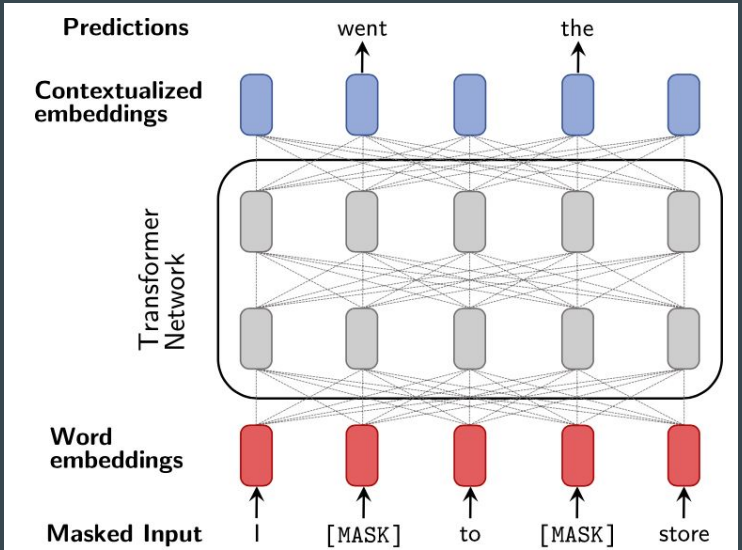


Fig. 3. A high-level illustration of BERT. Words in the input sequence are randomly masked out and then all words are embedded as vectors in $\mathbb{R}^d$. A Transformer network applies multiple layers of multiheaded attention to the representations. The final representations are used to predict the identities of the masked-out input words.

Emergent linguistic structure in artificial neural networks trained by self-supervision

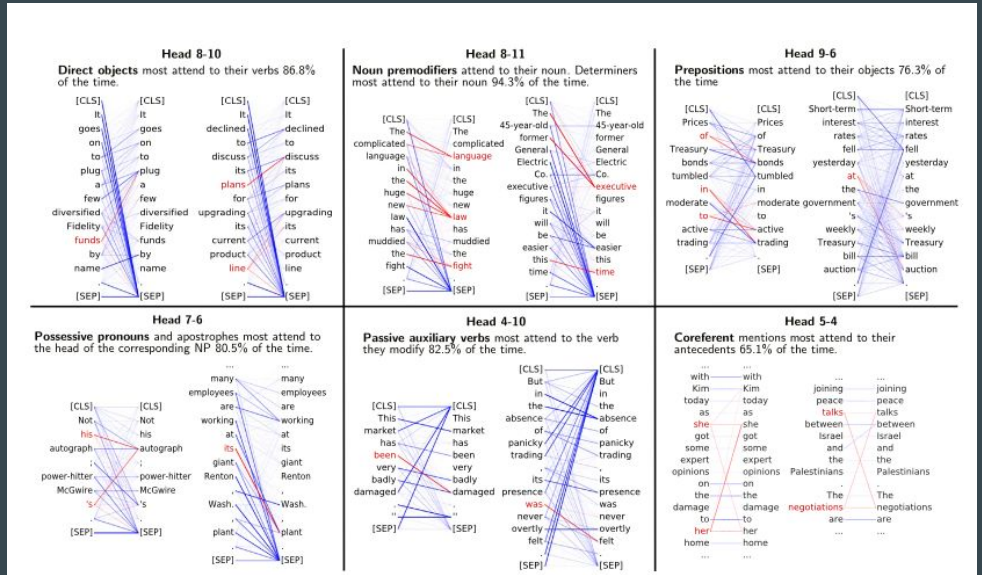


Fig. 6. Some BERT attention heads that appear sensitive to linguistic phenomena, despite not being explicitly trained on linguistic annotations. In the example attention maps, the darkness of a line indicates the size of the attention weight. All attention to/from red words is colored red; these words are chosen to highlight certain of the attention heads' behaviors. [CLS] (classification) and [SEP] (separator) are special tokens BERT adds to the input during preprocessing. Attention heads are numbered by their layer and index in BERT. Reprinted with permission from ref. 59, which is licensed under CC BY 4.0.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- Two steps in this framework: **pre-training and fine-tuning.**
- **Pre-training** : The model is trained on unlabeled data over different pre-training tasks
    - Task 1: Masked LM
    - Task 2: Next Sentence Prediction (NSP)
- **Fine-tuning** : The BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks.
    - For each task, we simply plug in the task specific inputs and outputs into BERT and finetune all the parameters end-to-end.

Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning

- This paper specifically tackle the few-shot intent detection task rather than the general full-shot learning
- This paper design a schema and employ contrastive learning in both self-supervised pre-training and supervised fine-tuning stages.
- **CPFT Methodology :** consider a few-shot intent detection task that handles C user intents, where the task is to classify a user utterance u into one of the C classes.
    - Self-supervised Pre-training : the model efficiently utilizes many unlabeled user utterances
    - Supervised Fine-tuning : treat two utterances from the same class as a positive pair and the two utterances across different classes as a negative pair for contrastive learning

Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning

Model	CLINC150		BANKING77		HWU64	
	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
RoBERTa+Classifier (Zhang et al., 2020a)	87.99	91.55	74.04	84.27	75.56	82.90
USE (Casanueva et al., 2020)	87.82	90.85	76.29	84.23	77.79	83.75
CONVERT (Casanueva et al., 2020)	89.22	92.62	75.32	83.32	76.95	82.65
USE+CONVERT (Casanueva et al., 2020)	90.49	93.26	77.75	85.19	80.01	85.83
CONVBERT (Mehri et al., 2020a)	-	92.10	-	83.63	-	83.77
CONVBERT + MLM (Mehri et al., 2020a)	-	92.75	-	83.99	-	84.52
CONVBERT + Combined (Mehri et al., 2020b)	-	93.97	-	85.95	-	86.28
DNNC (Zhang et al., 2020a)	91.02	93.76	80.40	86.71	80.46	84.72
CPFT	**92.34**	**94.18**	**80.86**	**87.20**	**82.03**	**87.13**

Table 2: Testing accuracy (×100%) on three datasets under 5-shot and 10-shot settings.

Model	CLINC150		BANKING77		HWU64	
	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
CPFT	92.34	94.18	80.86	87.20	82.03	87.13
w/o Contrastive pre-training	-4.15	-2.63	-4.11	-2.37	-6.01	-4.17
w/o Supervised contrastive learning	-0.56	-0.32	-2.06	-0.88	-1.14	-0.27
w/o Contrastive pre-training + w/o Supervised contrastive learning	-4.35	-2.69	-6.82	-2.93	-6.47	-4.23

Attention is all you need

- This paper proposes the Transformer based on attention mechanisms, does not use recurrence and convolutions entirely.
- Transformer is superior in quality while being more parallelizable
- Transformer is requiring significantly less time to train.

- Attention mechanisms allow modeling of dependencies without regard to their distance in the input or output sequences
- The Transformer using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder

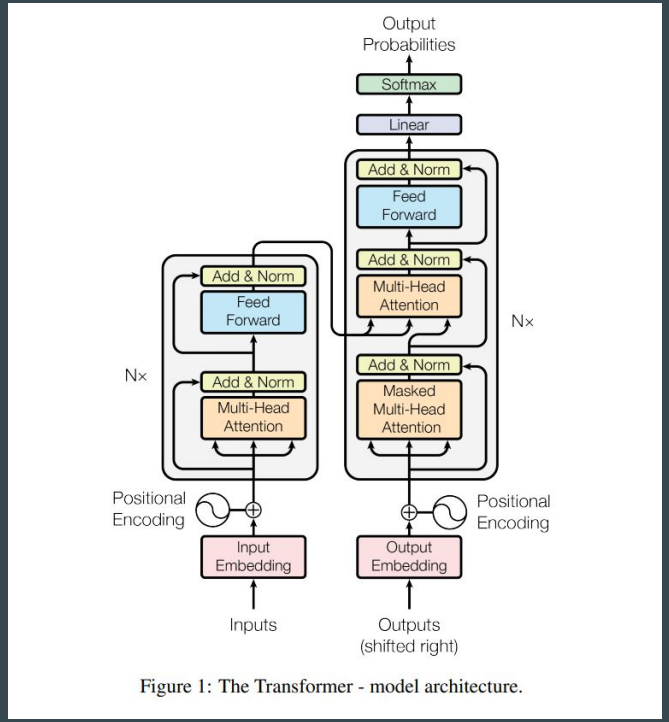


Figure 1: The Transformer - model architecture.

Why Self-Attention

This paper compare various aspects of self-attention layers to the recurrent and convolutional layers.

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

- Enable parallelizability hence more efficient
- Learn long-range dependencies better
- Provides interpretability

What Does BERT Look At? An Analysis of BERT's Attention

For example, Surface-Level Patterns in Attention

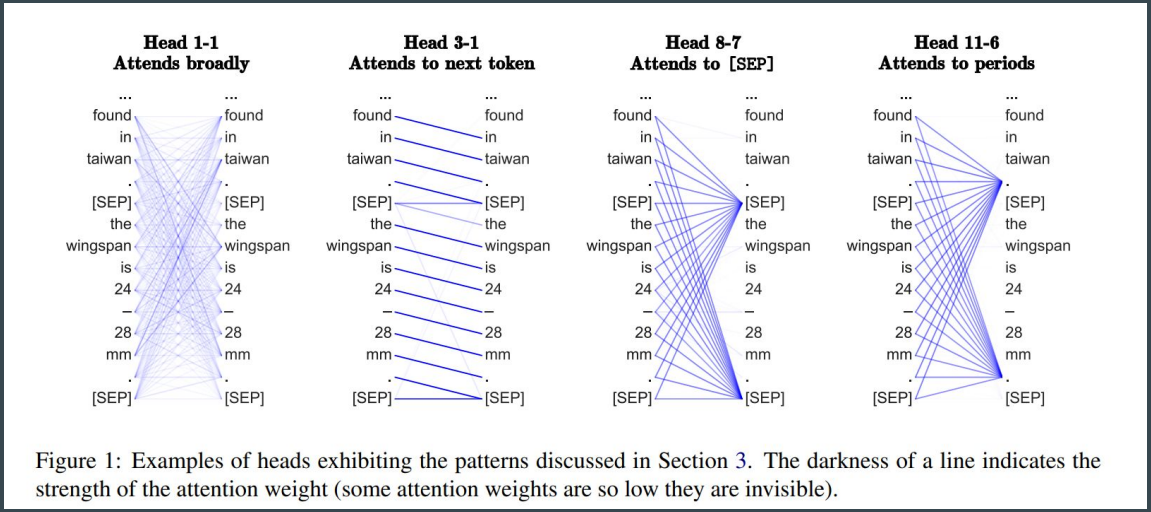


Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

BERT's attention heads exhibit patterns such as attending to delimiter tokens, specific positional offsets, or broadly attending over the whole sentence, with heads in the same layer often exhibiting similar behaviors.

Surface-Level Patterns in Attention: Relative Position

They compute how often BERT's attention heads attend to the current token, the previous token, or the next token.

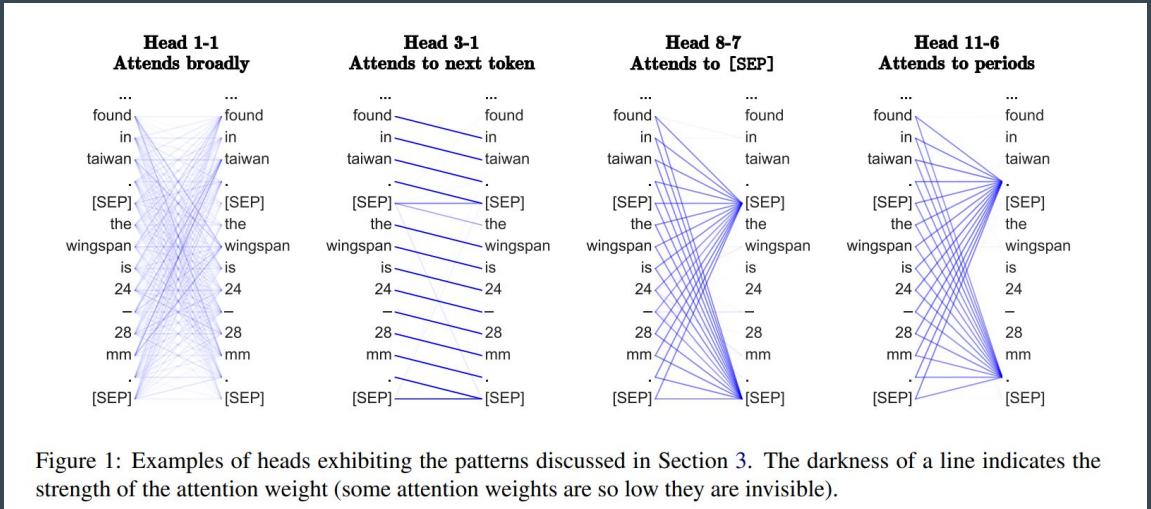
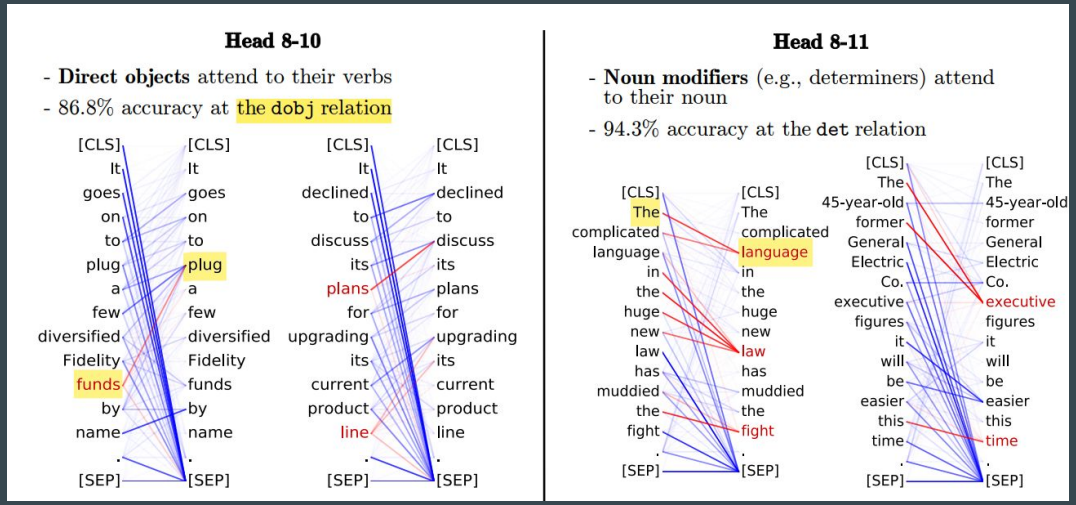


Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

Result: The most heads put little attention on the current token. However, there are heads that specialize to attending heavily on the next or previous token, especially in earlier layers of the network.

Surface-Level Patterns in Attention: Attending to Separator Tokens



They show that certain attention heads correspond well to linguistic notions of syntax and coreference. For example, they find heads that attend to the direct objects of verbs, determiners of nouns.

Furthermore, qualitative analysis show that heads with specific functions attend to [SEP] when the function is not called for.

Project Topic:

An Analysis of Transformer Model's Attention

Objective :

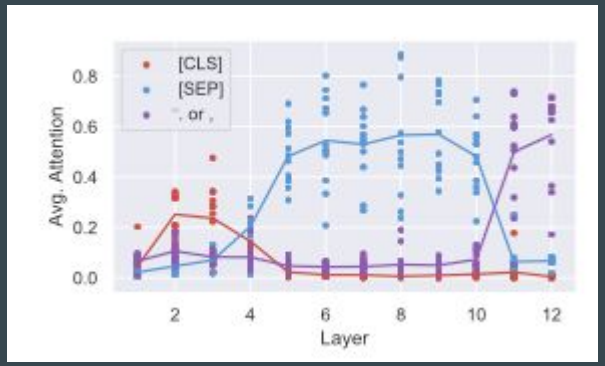
This project will analyze Transformer Model's Attention by comparison with BERT, XLNet and RoBERTa which are which are state-of-the-art for NLU tasks.

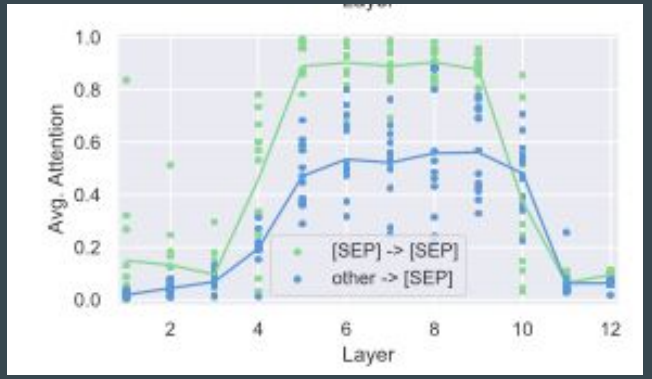
Method :

To Implement " What does BERT Look At ? An Analysis of BERT's Attention." on XLNet or RoBERTa.

First Step : Surface-Level Patterns in Attention.

We are planned to start with perform an analysis of surface-level patterns in how models attention heads behave.





Pic : The average attention a particular BERT attention head puts toward a token type.