DeBERTa

DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION

Published as a conference paper at ICLR 2021

Pengcheng He[1], **Xiaodong Liu**[2], **Jianfeng Gao**[2], **Weizhu Chen**[1]
[1] Microsoft Dynamics 365 AI [2] Microsoft Research
{penhe,xiaodl,jfgao,wzchen}@microsoft.com

채형주

목차

- intro
- disentangled attention mechanism
- EMD
- performance
- analysis

intro

- disentangled attention
- EMD
- RoBERTa-Large와 비교했을때 절반의 training data로 더 나은 성능을 보여줌
- ensemble DeBERTa로 SuperGLUE에서 사람을 이김(90.3 vs 89.8)

intro

- Disentangled attention
  - BERT는 하나의 word가 word_embedding + position_embedding으로 표현
  - DeBERTa에서는 2개의 vector(content, position)으로 표현
  - attention weight을 disentangled matrix로 구함

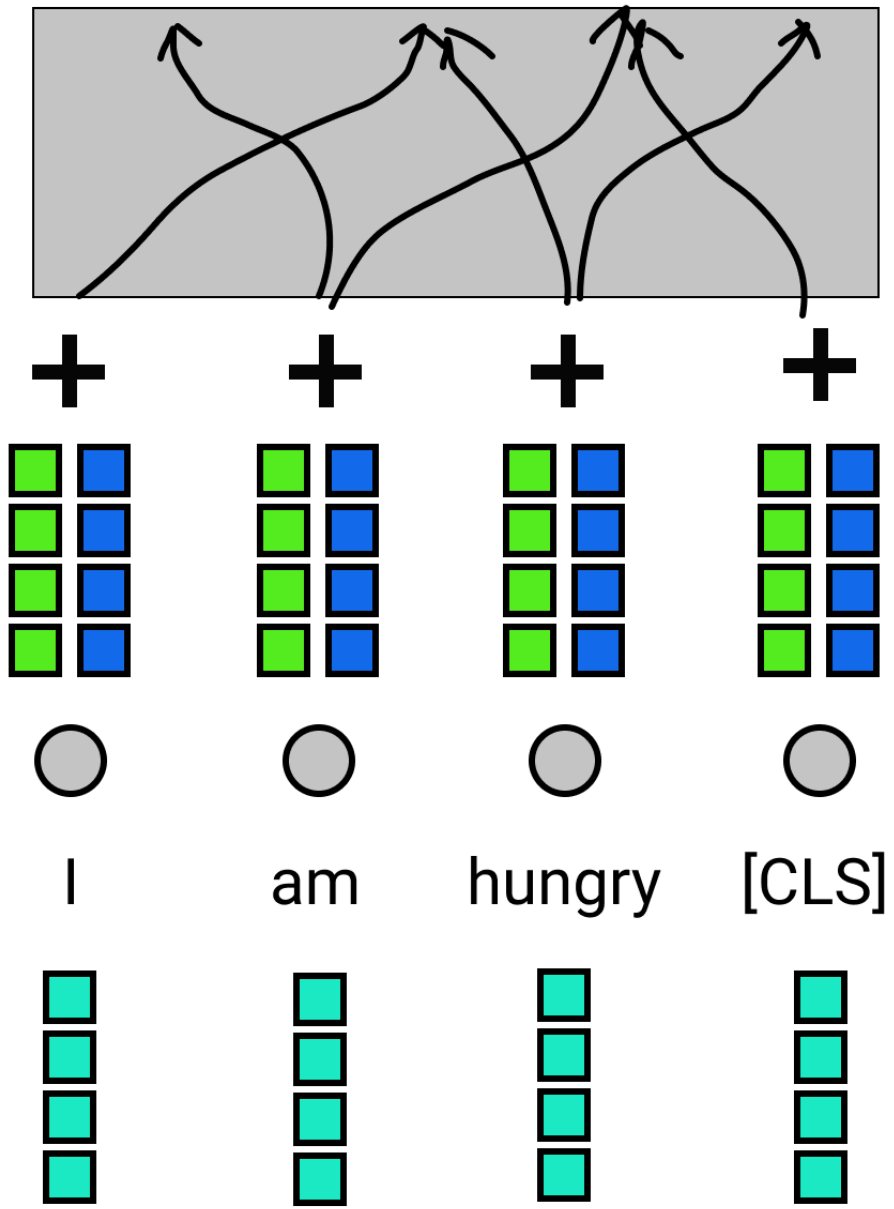
- Enhanced mask decoder
  - BERT처럼 MLM이다.
  - 하지만 relative positio만을 고려하고 absolute position은 고려되지 않는다.

Disentangled attention

- classic attention
  - entangled information
  - let's separate them
  - Content+Position attention

$$Q = HW_q, K = HW_k, V = HW_v, A = \frac{QK^\intercal}{\sqrt{d}}$$

$$H_o = \text{softmax}(A)V$$



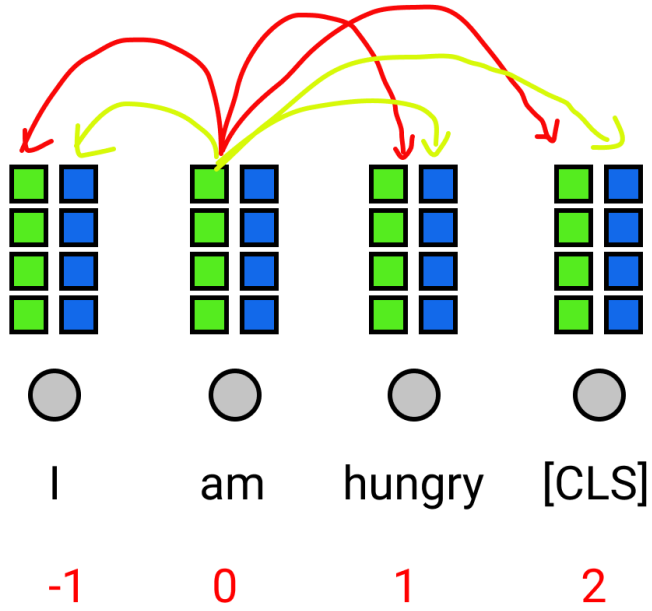
Disentangled attention

Ai,j : cross attention score between tokens I and token j
Hi : content vector
Pi|j : relative position with the token at position k

$$A_{i,j} = \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^\intercal$$
$$= H_i H_j^\intercal + H_i P_{j|i}^\intercal + P_{i|j} H_j^\intercal + P_{i|j} P_{j|i}^\intercal$$

	Content j	**Position j**
Content i	Content i – Content j	Content i – Position j
Position i	Position i – Content j	Position i – Position j

Disentangled attention



$$\boldsymbol{Q} = \boldsymbol{H}\boldsymbol{W}_q, \boldsymbol{K} = \boldsymbol{H}\boldsymbol{W}_k, \boldsymbol{V} = \boldsymbol{H}\boldsymbol{W}_v, \boldsymbol{A} = \frac{\boldsymbol{Q}\boldsymbol{K}^\intercal}{\sqrt{d}}$$

$$\boldsymbol{H}_o = \mathrm{softmax}(\boldsymbol{A})\boldsymbol{V}$$

why relative position???

$$\boldsymbol{Q}_c = \boldsymbol{H}\boldsymbol{W}_{q,c}, \boldsymbol{K}_c = \boldsymbol{H}\boldsymbol{W}_{k,c}, \boldsymbol{V}_c = \boldsymbol{H}\boldsymbol{W}_{v,c}, \boldsymbol{Q}_r = \boldsymbol{P}\boldsymbol{W}_{q,r}, \boldsymbol{K}_r = \boldsymbol{P}\boldsymbol{W}_{k,r}$$

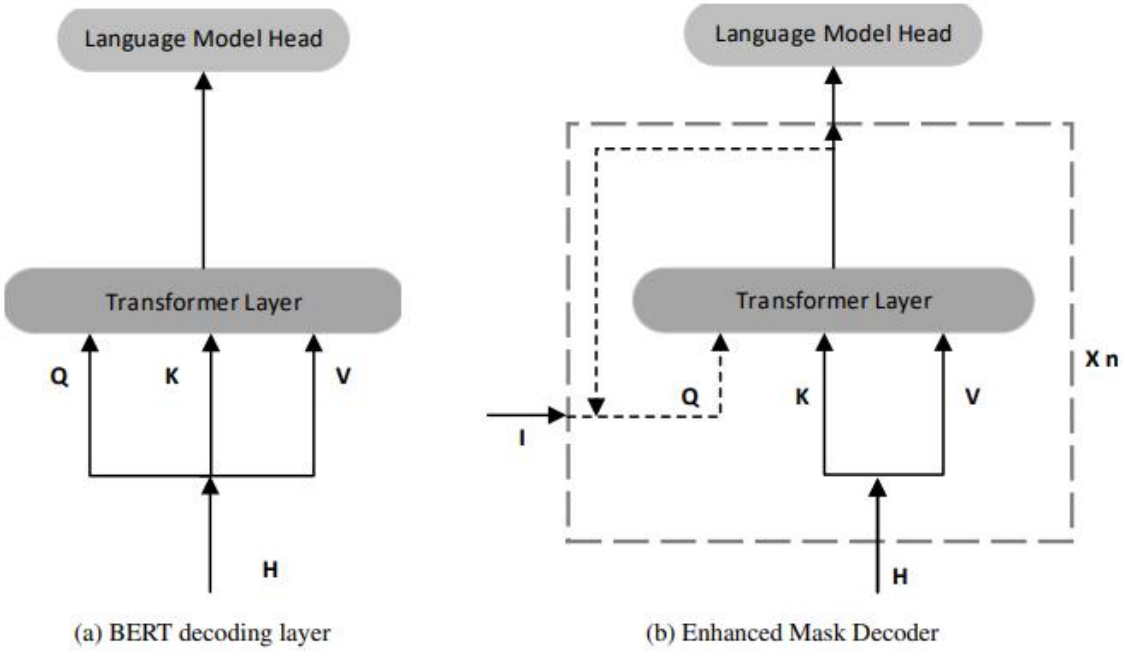
$$\tilde{A}_{i,j} = \underbrace{\boldsymbol{Q}_i^c \boldsymbol{K}_j^{c\intercal}}_{\text{(a) content-to-content}} + \underbrace{\boldsymbol{Q}_i^c \boldsymbol{K}_{\delta(i,j)}^{r\intercal}}_{\text{(b) content-to-position}} + \underbrace{\boldsymbol{K}_j^c \boldsymbol{Q}_{\delta(j,i)}^{r\intercal}}_{\text{(c) position-to-content}}$$

$$\boldsymbol{H}_o = \mathrm{softmax}(\frac{\tilde{\boldsymbol{A}}}{\sqrt{3d}})\boldsymbol{V}_c$$

$$\delta(i,j) = \begin{cases} 0 & \text{for} \quad i-j \leqslant -k \\ 2k-1 & \text{for} \quad i-j \geqslant k \\ i-j+k & \text{others.} \end{cases}$$

Enhanced Mask Decoder

- "a new [redacted] opened beside the new [redacted] [mall / store]
- absolute position embedding is added to the hidden state of content as query vector

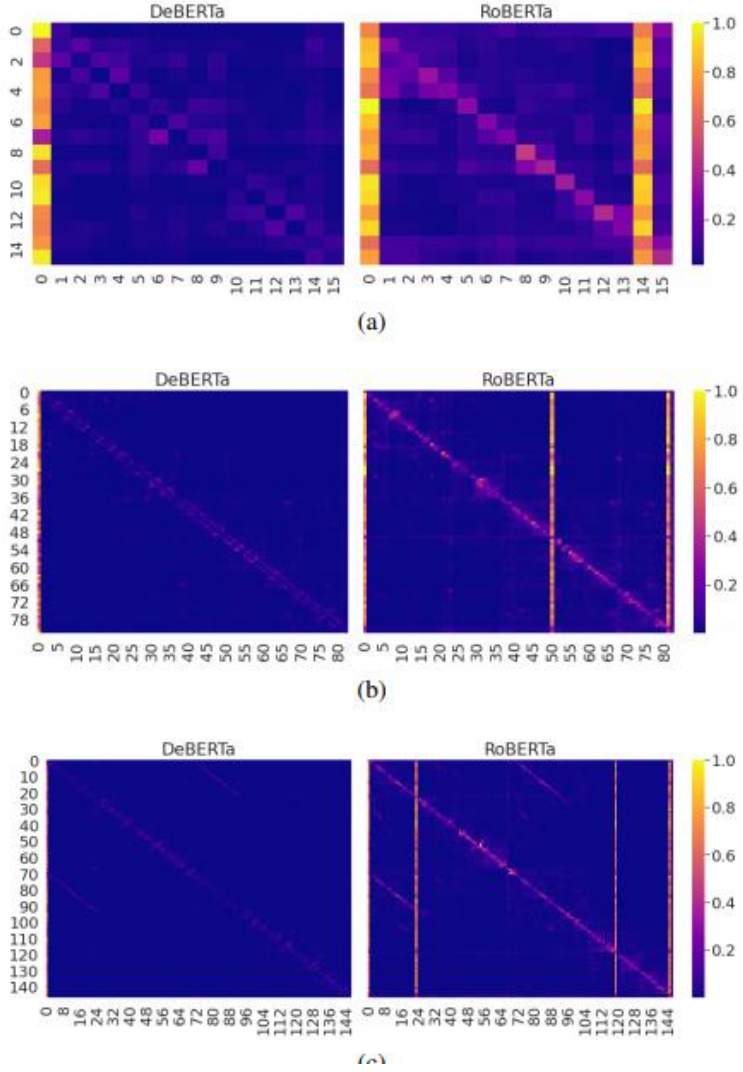


(a) BERT decoding layer

(b) Enhanced Mask Decoder

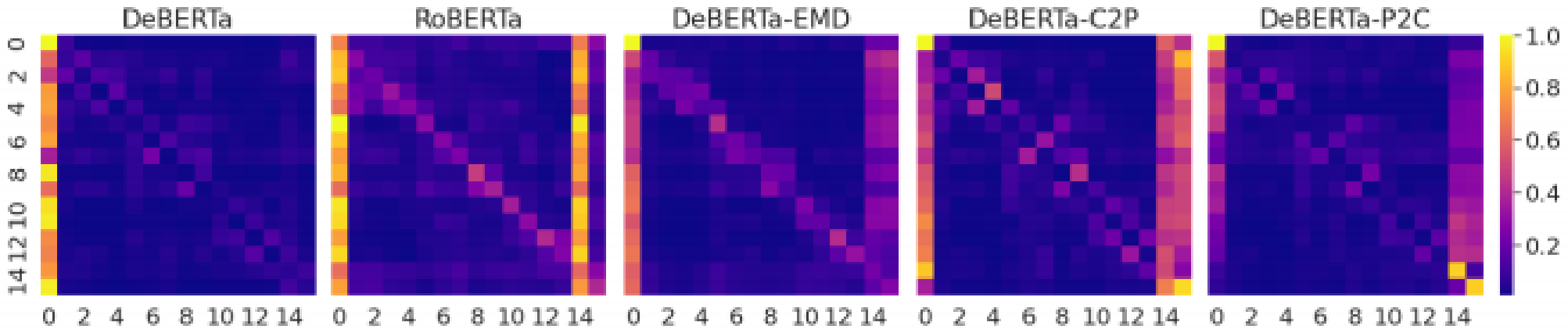
Enhanced Mask Decoder

- vertical strips in RoBERTa (a, the, .)
- DeBERTa 는 [CLS] 에만 strip



Enhanced Mask Decoder

- RoBERTa 에는 뚜렷한 대각선이 나타남
- DeBERTa에는 EMD때문에 나타나지 않음



Performance

- RoBERTa and XLNet : 500k steps, 8K samples
- DeBERTa : one million steps, 2K samples(half of RoBERTa and XLNet)
- 2K batch size, 1M steps takes 20days(96 V100 GPUs)

Model	CoLA Mcc	QQP Acc	MNLI-m/mm Acc	SST-2 Acc	STS-B Corr	QNLI Acc	RTE Acc	MRPC Acc	Avg.
BERT_{large}	60.6	91.3	86.6/-	93.2	90.0	92.3	70.4	88.0	84.05
RoBERTa_{large}	68.0	92.2	90.2/90.2	96.4	92.4	93.9	86.6	90.9	88.82
XLNet_{large}	69.0	92.3	90.8/90.8	97.0	92.5	94.9	85.9	90.8	89.15
$\text{ELECTRA}_{\text{large}}$	69.1	92.4	90.9/-	96.9	92.6	95.0	88.0	90.8	89.46
DeBERTa_{large}	70.5	92.3	91.1/91.1	96.8	92.8	95.3	88.3	91.9	90.00

Table 1: Comparison results on the GLUE development set.

Performance

- Megatron1.3B 이 3배 크지만 3, 4개의 벤치마크에서 성능을 능가함
- SOTA 달성

Model	MNLI-m/mm Acc	SQuAD v1.1 F1/EM	SQuAD v2.0 F1/EM	RACE Acc	ReCoRD F1/EM	SWAG Acc	NER F1
BERT_{large}	86.6/-	90.9/84.1	81.8/79.0	72.0	-	86.6	92.8
ALBERT_{large}	86.5/-	91.8/85.2	84.9/81.8	75.2	-	-	-
RoBERTa_{large}	90.2/90.2	94.6/88.9	89.4/86.5	83.2	90.6/90.0	89.9	93.4
XLNet_{large}	90.8/90.8	95.1/89.7	90.6/87.9	85.4	-	-	-
Megatron_{336M}	89.7/90.0	94.2/88.0	88.1/84.8	83.0	-	-	-
DeBERTa_{large}	**91.1/91.1**	**95.5/90.1**	**90.7/88.0**	**86.8**	**91.4/91.0**	**90.8**	**93.8**
$\text{ALBERT}_{xxlarge}$	90.8/-	94.8/89.3	90.2/87.4	86.5	-	-	-
$\text{Megatron}_{1.3B}$	90.9/91.0	94.9/89.1	90.2/87.1	87.3	-	-	-
$\text{Megatron}_{3.9B}$	91.4/91.4	95.5/90.0	91.2/88.5	89.5	-	-	-

Table 2: Results on MNLI in/out-domain, SQuAD v1.1, SQuAD v2.0, RACE, ReCoRD, SWAG, CoNLL 2003 NER development set. Note that missing results in literature are signified by "-".

Analysis

- EMD를 삭제했을 경우 1.4%(RACE), 0.3%(SQuAD v1.1), 1.2%(SQuAD v2.0) loss

- -EMD is the DeBERTa base model without EMD.
- -C2P is the DeBERTa base model without the content-to-position term ((c) in Eq. 4).
- -P2C is the DeBERTa base model without the position-to-content term ((b) in Eq. 4). As XLNet also uses the relative position bias, this model is close to XLNet plus EMD.

Model	MNLI-m/mm Acc	SQuAD v1.1 F1/EM	SQuAD v2.0 F1/EM	RACE Acc
BERT_{base} Devlin et al. (2019)	84.3/84.7	88.5/81.0	76.3/73.7	65.0
RoBERTa_{base} Liu et al. (2019c)	84.7/-	90.6/-	79.7/-	65.6
XLNet_{base} Yang et al. (2019)	85.8/85.4	-/-	81.3/78.5	66.7
$\text{RoBERTa-ReImp}_{base}$	84.9/85.1	91.1/84.8	79.5/76.0	66.8
DeBERTa_{base}	**86.3/86.2**	**92.1/86.1**	**82.5/79.3**	**71.7**
-EMD	86.1/86.1	91.8/85.8	81.3/78.0	70.3
-C2P	85.9/85.7	91.6/85.8	81.3/78.3	69.3
-P2C	86.0/85.8	91.7/85.7	80.8/77.6	69.6
-(EMD+C2P)	85.8/85.9	91.5/85.3	80.3/77.2	68.1
-(EMD+P2C)	85.8/85.8	91.3/85.1	80.2/77.1	68.5

Table 4: Ablation study of the DeBERTa base model.

Analysis

Model	BoolQ Acc	CB F1/Acc	COPA Acc	MultiRC F1a/EM	ReCoRD F1/EM	RTE Acc	WiC Acc	WSC Acc	Average Score
RoBERTa$_{\text{large}}$	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	84.6
NEXHA-Plus	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	86.7
T5$_{11B}$	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	89.3
T5$_{11B}$+Meena	**91.3**	**95.8/97.6**	97.4	88.3/63.0	94.2/93.5	92.7	**77.9**	95.9	90.2
Human	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	89.8
DeBERTa$_{1.5B}$+SiFT	90.4	94.9/97.2	96.8	**88.2/63.7**	**94.5/94.1**	**93.2**	76.4	**95.9**	89.9
DeBERTa$_{Ensemble}$	90.4	95.7/97.6	**98.4**	88.2/63.7	94.5/94.1	93.2	77.5	95.9	**90.3**

Table 5: SuperGLUE test set results scored using the SuperGLUE evaluation server. All the results are obtained from https://super.gluebenchmark.com on January 6, 2021.