DEBERTAV3 : IMPROVING DEBERTA USING ELECTRA-STYLE PRE-TRAINING WITH GRADIENT - DISENTANGLED EMBEDDING SHARING

https://arxiv.org/pdf/2111.09543

0. Introduction

- 사전학습 언어 모델은 자연어 이해 성능을 크게 향상시켜 왔음
- DeBERTa는 disentangled attention을 통해 BERT 계열 대비 성능을 개선
- 기존 사전학습 방식은 학습 효율과 성능 사이의 트레이드오프 존재
- ELECTRA 방식은 샘플 효율이 높지만 embedding 공유로 인한 gradient 충돌 문제가 있음
- 본 논문은 embedding gradient 충돌을 해결하여 사전학습 효율과 성능을 동시에 개선하는 것을 목표로 함

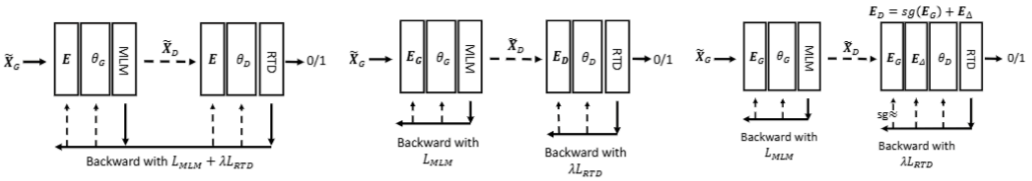
1. Overview

- DeBERTaV3는 DeBERTa 구조를 기반으로 ELECTRA 스타일 사전학습을 적용
- MLM 대신 replaced token detection 방식을 사용
- generator와 discriminator가 embedding을 공유하되 gradient 흐름을 분리
- 동일한 계산 자원 대비 더 높은 downstream 성능을 달성하는 것이 목적
- 멀티링구얼 확장 모델에서도 성능 향상을 확인

2. Challenges

- generator와 discriminator는 서로 다른 학습 목표를 가짐
- embedding을 단순 공유할 경우 gradient 간섭이 발생
- 이로 인해 embedding 품질 저하 및 학습 비효율 발생
- 대규모 사전학습 환경에서는 작은 비효율도 큰 비용 증가로 이어짐

3. Method



(a) (b) (c)

Figure 1: Illustration of different embedding sharing methods. (a) ES: $\boldsymbol{E}$, θ_G and θ_D will be jointly updated in a single backward pass with regards to $L_{MLM} + \lambda L_{RTD}$. (b) NES: $\boldsymbol{E}_G$ and θ_G will first be updated via the backward pass with regards to L_{MLM}, then $\boldsymbol{E}_D$ and θ_D will be updated via the backward pass with regards to λL_{RTD} . (c) GDES: $\boldsymbol{E}_G$ and θ_G will first be updated in the backward pass with regards to L_{MLM}, then $\boldsymbol{E}_\Delta$ and θ_D will be updated via the backward pass with regards to λL_{RTD} and $\boldsymbol{E}_G$. sg is the stop gradient operator that prevents the discriminator from updating $\boldsymbol{E}_G$

Table 1: Average cosine similarity of word embeddings of the generator and the discriminator with different embedding sharing methods.

Word Embedding Sharing	$\boldsymbol{E}_G$	$\boldsymbol{E}_D$	$\boldsymbol{E}_\Delta$
① ES	0.02	0.02	-
② NES	0.45	0.02	-
③ GDES	0.45	0.29	0.02

Table 2: Fine-tuning results on MNLI and SQuAD v2.0 tasks of base models trained with different embedding sharing methods.

Model	MNLI-m/mm Acc	SQuAD v2.0 F1/EM
$\text{BERT}_{\text{base}}$	84.3/84.7	76.3/73.7
$\text{ELECTRA}_{\text{base}}$	85.8/-	-/-
$\text{DeBERTa}_{\text{base}}$	86.3/86.2	82.5/79.3
$\text{DeBERTa+RTD}_{\text{base}}$		
① ES	88.8/88.4	86.3/83.5
② NES	88.3/87.9	85.3/82.7
③ GDES	**89.3/89.0**	**87.2/84.5**

- 핵심 설계는 gradient-disentangled embedding sharing
- generator와 discriminator가 동일한 embedding 값을 사용하되 gradient는 독립적으로 흐르도록 구조를 분리
- 사전학습은 replaced token detection 기반으로 진행
- generator는 토큰을 생성하고 discriminator는 토큰의 진위를 판별
- 기존 DeBERTa 아키텍처는 대부분 유지됨

4. Experiments

- 자연어 이해 벤치마크 태스크를 통해 성능 평가
- 기존 DeBERTa 및 ELECTRA 계열 모델과 비교
- 단일 언어 및 멀티링구얼 환경에서 실험 수행
- downstream 태스크 평균 성능을 주요 평가 기준으로 사용

5. Results

Table 3: Comparison results on the GLUE development set.

Model #Train	CoLA Mcc 8.5k	QQP Acc 364k	MNLI-m/mm Acc 393k	SST-2 Acc 67k	STS-B Corr 7k	QNLI Acc 108k	RTE Acc 2.5k	MRPC Acc 3.7k	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
$ELECTRA_{large}$	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
$DeBERTaV3_{large}$	**75.3**	**93.0**	**91.8/91.9**	96.9	**93.0**	**96.0**	**92.7**	**92.2**	**91.37**

Table 4: Results on MNLI in/out-domain, SQuAD v2.0, RACE, ReCoRD, SWAG, CoNLL 2003 NER development set. Note that missing results in literature are signified by "-".

Model	MNLI-m/mm Acc	SQuAD v2.0 F1/EM	RACE Acc	ReCoRD F1/EM	SWAG Acc	NER F1
$BERT_{large}$	86.6/-	81.8/79.0	72.0	-	86.6	92.8
$ALBERT_{large}$	86.5/-	84.9/81.8	75.2	-	-	-
$RoBERTa_{large}$	90.2/90.2	89.4/86.5	83.2	90.6/90.0	89.9	93.4
$XLNet_{large}$	90.8/90.8	90.6/87.9	85.4	-	-	-
$ELECTRA_{large}$	90.9/-	-/88.1	-	-	-	-
$Megatron_{336M}$	89.7/90.0	88.1/84.8	83.0	-	-	-
$DeBERTa_{large}$	91.1/91.1	90.7/88.0	86.8	91.4/91.0	90.8	93.8
$DeBERTaV3_{large}$	**91.8/91.9**	**91.5/89.0**	**89.2**	**92.3/91.8**	**93.4**	**93.9**
$ALBERT_{xxlarge}$	90.8/-	90.2/87.4	86.5	-	-	-
$Megatron_{1.3B}$	90.9/91.0	90.2/87.1	87.3	-	-	-
$Megatron_{3.9B}$	91.4/91.4	91.2/88.5	89.5	-	-	-
$DeBERTa_{1.5B}$	91.7/91.9	92.2/89.7	90.8	94.5/94.0	92.3	-

Table 5: Results on MNLI in/out-domain (m/mm) and SQuAD v2.0 development set. TinyBERT$_{small}$(Jiao et al., 2019), MiniLMv2$_{small}$ and MiniLMv2$_{xsmall}$ models are pre-trained with knowledge distillation while BERT$_{small}$, DeBERTaV3$_{small}$ and DeBERTaV3$_{xsmall}$ are trained from scratch with MLM and RTD objective, respectively.

Model	Vocabulary Size(K)	Backbone #Params(M)	MNLI-m/mm ACC	SQuAD v2.0 F1/EM
Base models:12 layers,768 hidden size,12 heads				
BERT$_{base}$	30	86	84.3/84.7	76.3/73.7
RoBERTa$_{base}$	50	86	87.6/-	83.7/80.5
XLNet$_{base}$	32	92	86.8/-	-/80.2
ELECTRA$_{base}$	30	86	88.8/-	-/80.5
DeBERTa$_{base}$	50	100	88.8/88.5	86.2/83.1
DeBERTaV3$_{base}$	128	86	**90.6/90.7**	**88.4/85.4**
Small models:6 layers,768 hidden size,12 heads				
TinyBERT$_{small}$	30	44	84.5/-	77.7/-
MiniLMv2$_{small}$	30	44	87.0/-	81.6/
BERT$_{small}$	30	44	81.8/-	73.2/-
DeBERTaV3$_{small}$	128	44	**88.2/87.9**	**82.9/80.4**
XSmall models:12 layers,384 hidden size,6 heads				
MiniLMv2$_{xsmall}$	30	22	86.9/-	82.3/
DeBERTaV3$_{xsmall}$	128	22	**88.1/88.3**	**84.8/82.0**

Table 6: Results on XNLI test set under the cross-lingual transfer and the translate-train-all settings.

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Cross-lingual transfer																
XLM	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
mT5$_{base}$	84.7	79.1	80.3	77.4	77.1	78.6	77.1	72.8	73.3	74.2	73.2	74.1	70.8	69.4	68.3	75.4
XLM-R$_{base}$	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
mDeBERTaV3$_{base}$	**88.2**	**82.6**	**84.4**	**82.7**	**82.3**	**82.4**	**80.8**	**79.5**	**78.5**	**78.1**	**76.4**	**79.5**	**75.9**	**73.9**	**72.4**	**79.8**
Translate train all																
XLM	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
mT5$_{base}$	82.0	77.9	79.1	77.7	78.1	78.5	76.5	74.8	74.4	74.5	75.0	76.0	72.2	71.5	70.4	75.9
XLM-R$_{base}$	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
mDeBERTaV3$_{base}$	**88.9**	**84.4**	**85.3**	**84.8**	**84.0**	**84.5**	**83.2**	**82.0**	**81.6**	**82.0**	**79.8**	**82.6**	**79.3**	**77.3**	**73.6**	**82.2**

- DeBERTaV3는 기존 DeBERTa 대비 일관된 성능 향상
- ELECTRA 기반 모델 대비도 평균 성능 우위
- 동일하거나 더 적은 사전학습 비용으로 높은 성능 달성
- 멀티링구얼 환경에서도 zero-shot 성능 개선 확인
- 사전학습 안정성과 수렴 속도 모두 개선됨

6. Insight

- 이 논문의 핵심은 새로운 모델 구조가 아니라 학습 신호 제어
- embedding 공유 자체보다 gradient 충돌이 성능 저하의 원인이었음을 명확히 지적
- 사전학습 효율 개선은 단순한 속도 문제가 아니라 표현 품질 문제와 직결됨
- ELECTRA 방식의 한계를 구조적으로 보완한 점이 실질적 기여