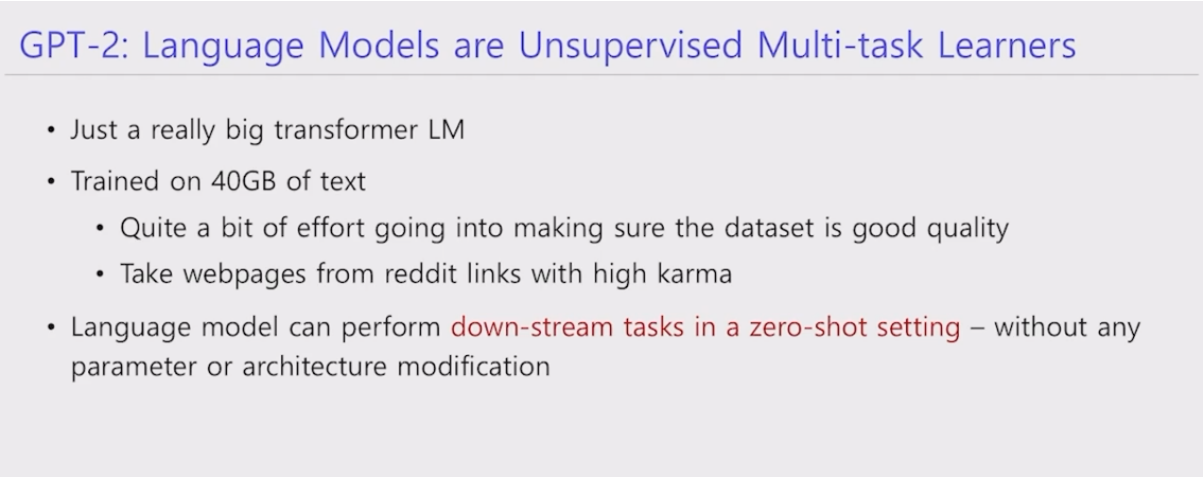
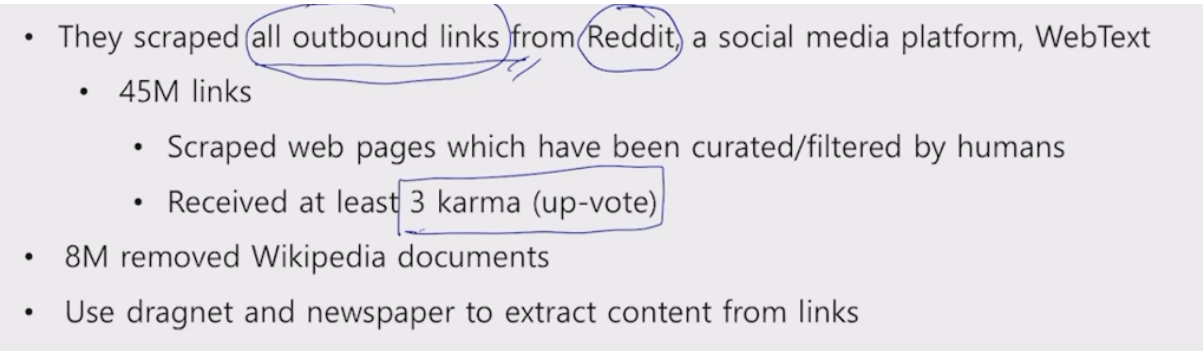


트랜스포머에 레이어를 잔뜩 쌓아서 키운 아키텍쳐로서, 여전히 다음단어를 예측하는 학습을 시켜뒀습니다. 데이터셋을 대규모로 쓰는 과정에서 데이터셋의 퀄리티를 높인 것이 주목된다고 한다.



바이트 페어 인코딩이라는 서브워드 인코딩을 가했습니다만,

세부적인 모델의 차이점이라면 레이어가 위로 가면 갈수록 초기화 값을 작게 만들어서 위쪽의 레이어가 하는 일이 적도록 하였다고 합니다. 제로-샷 러닝의 경우 55f1 스코어가 나왔습니다. 어느정도 충분히 가능성을 보였다고 할 수 있습니다.

GPT-3: Language Models are Few-Shot Learners

Language Models are Few-shot Learners

- Scaling up language models greatly improves task-agnostic, few-shot performance
- An autoregressive language model with 175 billion parameters in the few-shot setting
- 96 Attention layers, Batch size of 3.2M

압도적인 크기의 파라미터 수를 가지도록 더 쌓아낸 것인데, 더 많은 데이터를 더 큰 배치사이즈로 학습시키니 더 쩌는 성능이 나오더라...라는 이야기.

제로샷 세팅에서 성능이 어느정도 나오더라 라는 이야기입니다.

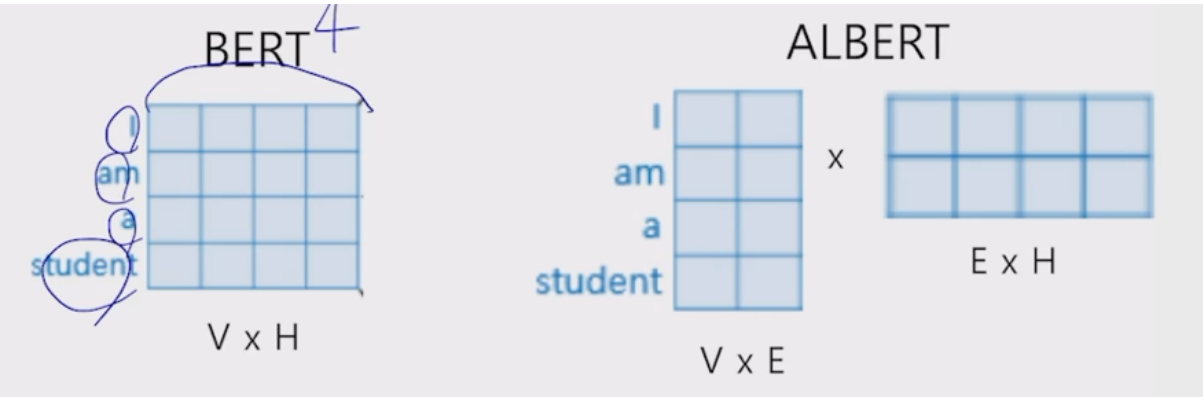
예시를 하나 주고 가르쳐주는 원-샷에서는 역시 성능이 더 올라가고, 퓨-샷에서도 쑥쑥 올라가더라 이 말이야...

그래서 모델이 비대해야 좋은걸까?

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

- Is having better NLP models as easy as having larger models?
  - Obstacles
    - Memory Limitation
    - Training Speed
  - Solutions
    - Factorized Embedding Parameterization
    - Cross-layer Parameter Sharing
    - (For Performance) Sentence Order Prediction

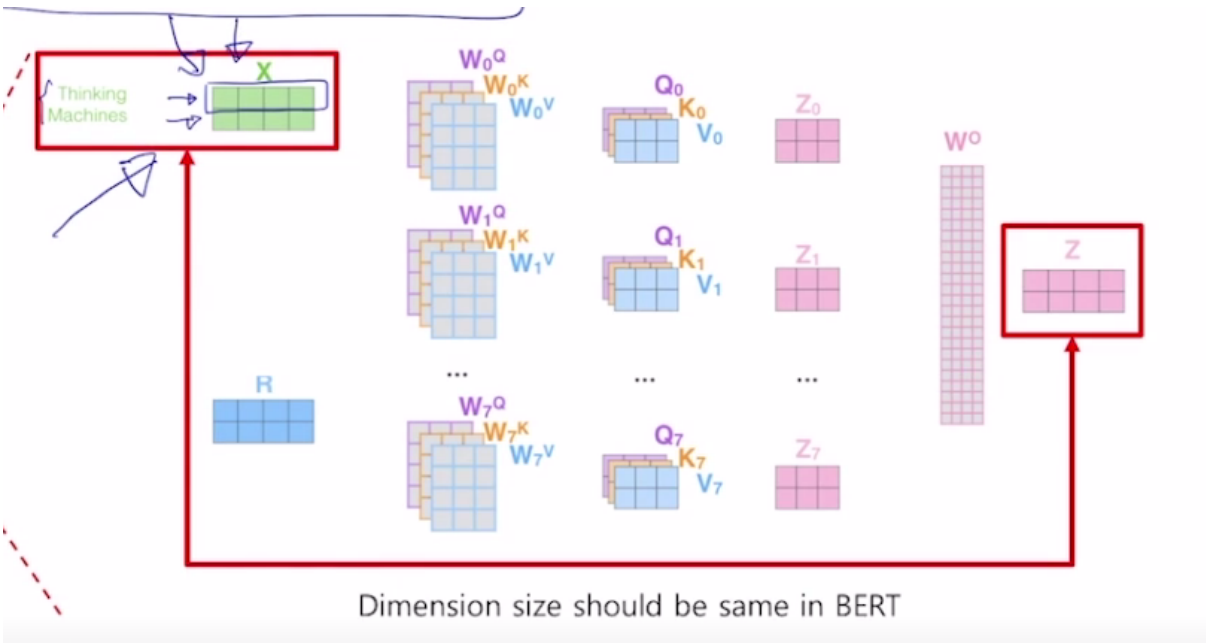
ALBERT : 임베딩 레이어의 디멘션을 줄이는 기법을 가져왔습니다만,



차원을 한번 줄여서 사용하는 것이라고 합니다.

해당 기법을 이전에 한번 본 것 같은데, 어떤 논문이었는지 기억이 안나네요...

하튼 저렇게 파라미터를 줄여서 학습 속도를 올렸던 듯 합니다.



- Cross-layer Parameter Sharing
  - Shared-FFN: Only sharing feed-forward network parameters across layers
  - Shared-attention: Only sharing attention parameters across layers
  - All-shared: Both of them

	Model	Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base E=768	all-shared	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8
	shared-attention	83M	89.9/82.7	80.0/77.2	84.0	91.4	67.7	81.6
	shared-FFN	57M	89.2/82.1	78.2/75.4	81.5	90.8	62.6	79.5
	not-shared	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base E=128	all-shared	12M	89.3/82.3	80.0/77.1	82.0	90.3	64.0	80.1
	shared-attention	64M	89.9/82.8	80.7/77.9	83.4	91.9	67.6	81.7
	shared-FFN	38M	88.9/81.6	78.6/75.6	82.3	91.7	64.4	80.2
	not-shared	89M	89.9/82.8	80.3/77.3	83.2	91.5	67.9	81.6

파라미터를 레이어별로 가를게 아니라 공유시킨다면 어떨까

- 생각보다 성능 하락이 크지 않았다.
- 그럼에도 파라미터 수는 극적으로 줄일 수 있었다.

- Sentence Order Prediction
  - Next Sentence Prediction pretraining task in BERT is too easy
  - Predict the ordering of two consecutive segments of text
    - Negative samples the same two consecutive segments but with their order swapped

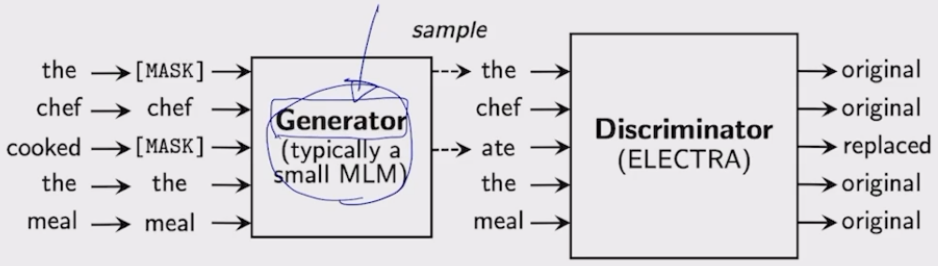
	Intrinsic Tasks			Downstream Tasks					
SP tasks	MLM	NSP	SOP	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
None	54.9	52.4	53.3	88.6/81.5	78.1/75.3	81.5	89.9	61.7	79.0
NSP	54.5	90.5	52.0	88.4/81.5	77.2/74.6	81.6	**91.1**	62.3	79.2
SOP	54.0	78.9	86.5	**89.3/82.3**	**80.0/77.1**	**82.0**	90.3	**64.0**	**80.1**

원래의 다음 문장 추측 훈련은 너무 쉬웠기에 랜덤한 인접문장을 사용하되 단어의 정방향/역방향을뒤틀어 문장의 논리적 요소를 파악해야 맞출 수 있도록 한 것 이다.

ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately

Efficiently Learning an Encoder that Classifies Token Replacements Accurately

- Learn to distinguish real input tokens from plausible but synthetically generated replacements
- Pre-training text encoders as discriminators rather than generators
- **Discriminator** is the main networks for pre-training.

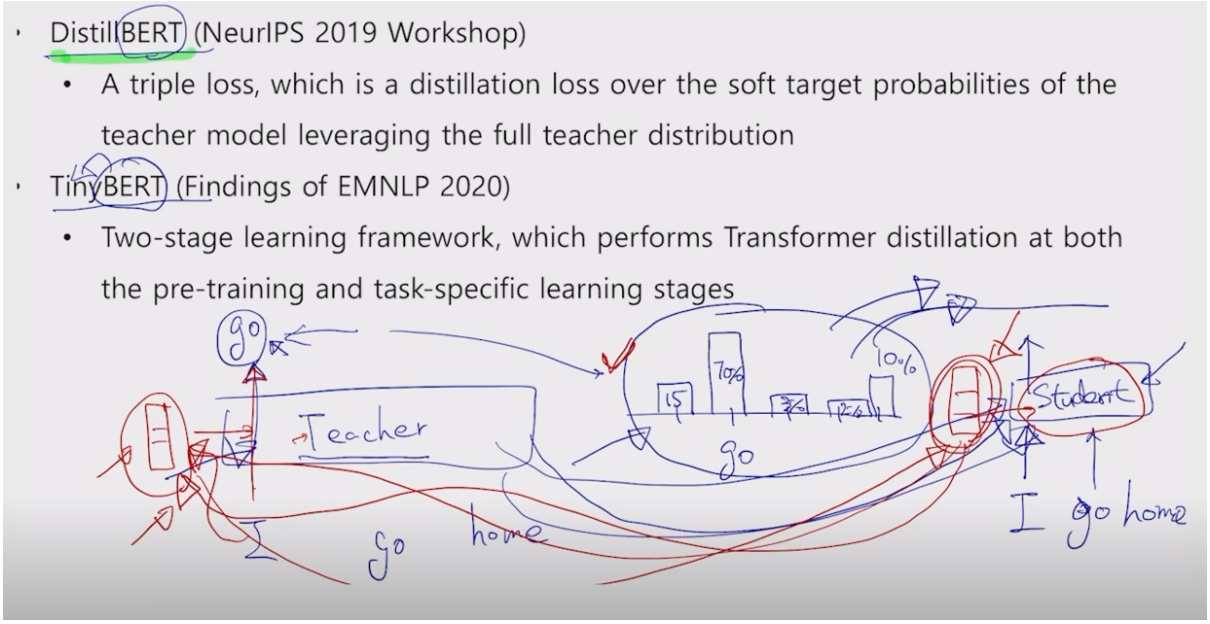


이 단어가 원래 있던 단어다 / 아니다 이진 분류 해주는 불안정성 검출 모델입니다.

적대적 관계로 학습이 됩니다만, GAN과도 연관이 어느정도 있습니다. 갠 모델에서 착안해서 프리트레인드 모델을 찾아내는 것입니다.

Light-weight Models

- DistillBERT (NeurIPS 2019 Workshop)
  - A triple loss, which is a distillation loss over the soft target probabilities of the teacher model leveraging the full teacher distribution
- TinyBERT (Findings of EMNLP 2020)
  - Two-stage learning framework, which performs Transformer distillation at both the pre-training and task-specific learning stages



경량화 모델 예시- 티쳐 모델의 결과를 모사하는 학생 모델을 나타냄.