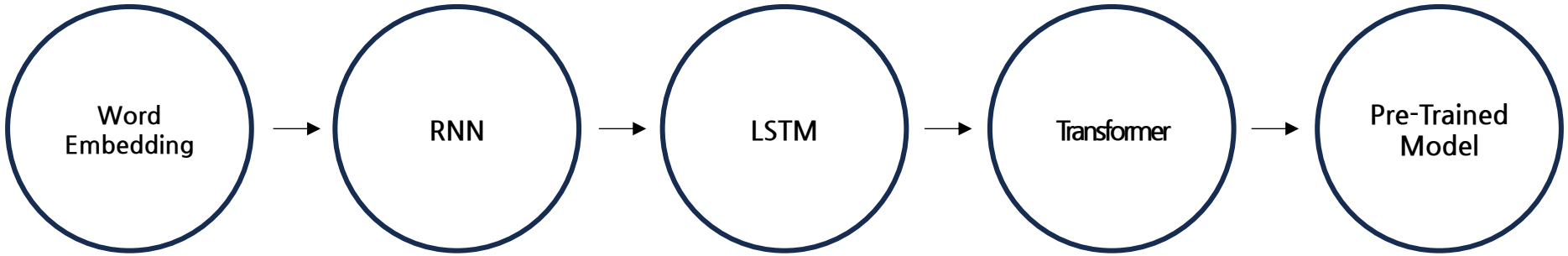
improving language understanding by generative pre-training

Background



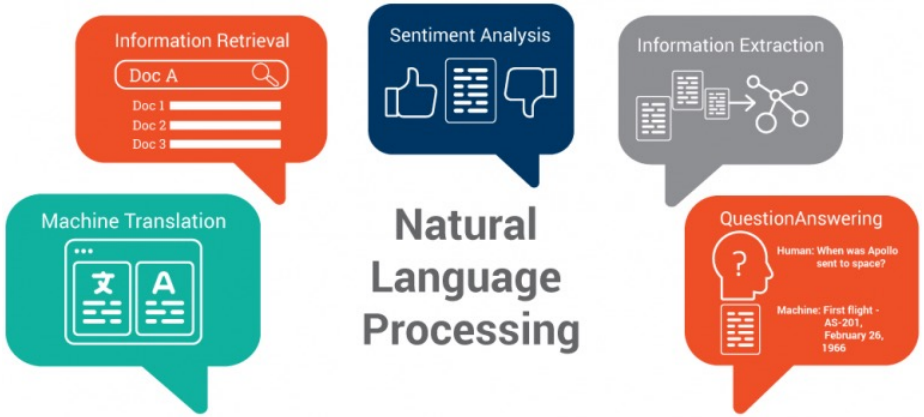
✓ 지금까지의 LM은 대용량 labeled data를 기반으로 학습됐으며 이는 다양한 분야에서 활용되기 어려움

✓ unlabeled data는 시간, 비용적인 측면에서 이점이 있으며 훌륭한 representation을 보임

✓ 데이터의 양 또한 풍부하고 다양한 특성을 학습할 수 있음

Background

unlabeled data의 어려움



✓ 어떤 optimization object가 representation 학습에 효과적인지 불분명

✓ Representation을 target task에 어떻게 효과적으로 전이할 것인지에 대한 합의가 없음

Background

Unsupervised pre-training **supervised fine-tuning**

✓ 다양한 task에 사용될 수 있도록 약간의 조정만으로 범용 representation 학습

✓ 대용량 unlabeled-corpus data와 task 학습에 쓰일 훈련 샘플들이 존재한다고 가정하며 corpus data는 한 domain에 국한되지 않음

✓ 모델은 long-term dependency에 강건한 transformer 사용

Related Work

✓ **Semi-supervised learning for NLP**

1. Unlabeled data를 사용하여 단어, 문맥 수준을 계산하고 이를 supervised model에 feature로 사용

2. 이는, 최근에 word embedding을 pre-train하는 방식으로 사용해왔으며 성능 개선에 큰 강점을 보임

3. 그러나 이는 word-level information에만 그치며 이 연구에서는 고수준의 의미를 학습

✓ **Unsupervised Pre-training**

1. 비지도 학습의 목적은 좋은 시작점을 찾는 것으로 초기 연구에서는 이미지 분류, 회귀 등에 쓰임

2. 최근 연구에서는 이러한 학습이 우수한 정규화, 즉 뉴럴 네트워크에 범용성을 더해준다는 것을 확인

3. 사전 연구로 이 논문과 비슷하게 언어 모델 objective를 사용해 사전학습과 미세조정을 하는 연구가 있었지만 이는 LSTM을 사용

Framework

1-stage

Unsupervised pre-training

✓ generative pre-training language model
✓ unlabeled Large corpus 학습

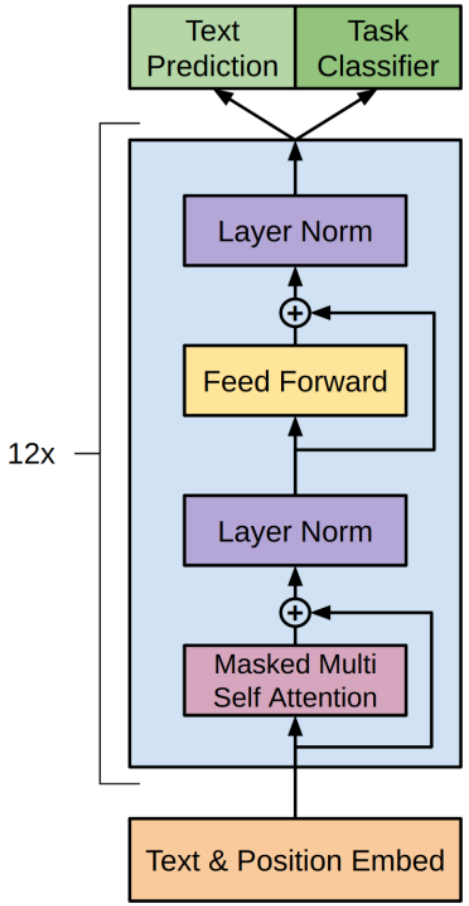
2-stage

supervised fine-tuning

✓ Discriminative fine tuning
✓ 특정 task에 맞춘 Labeled data 활용

Framework

Unsupervised pre-training



[모델 구조]

- 12개의transformer block으로 이루어짐
- transformer의 decoder 부분만 활용

$$\mathcal{U} = \{u_1, \ldots, u_n\}, \qquad L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

[학습과정]

1. Unlabeled된 corpus token U 준비
2. 다음 token에 대해 likelihood 최대화 하도록 학습(L1(U))
3. 이 때, K는 context window의 사이즈, Θ는 parameter를 의미

Framework

supervised fine-tuning

$$P(y|x^1,\ldots,x^m) = \texttt{softmax}(h_l^m W_y).$$

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1,\ldots,x^m).$$

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

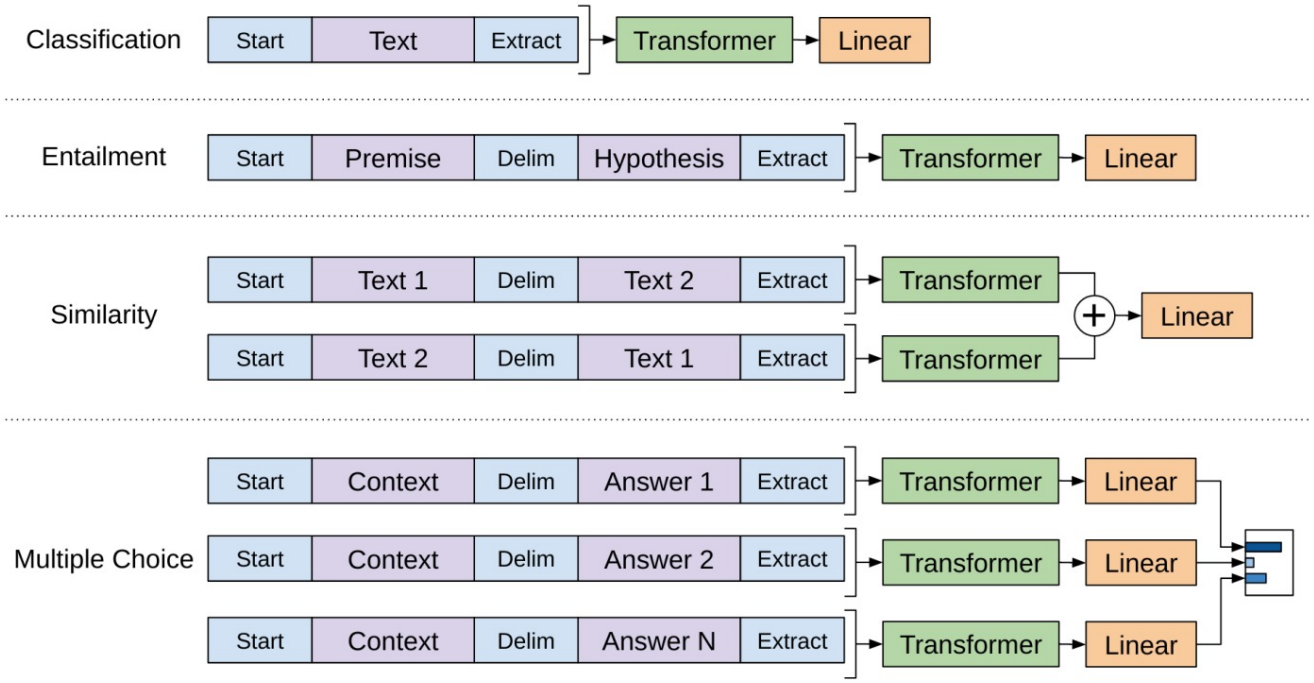
- 일반적인 언어 이해 능력 향상
- 수렴이 빨라짐

✓ 특정 task에 맞는 라벨 데이터 C가 있다고 가정하며 이는 x1, x2, x3…의 시퀀스와 라벨 y로 구성

✓ Input을 pre-trained model에 통과시켜 최종 출력물을 얻게 되고 선형 레이어에 입력으로 들어옴

✓ 마지막으로 최대 우도가 되는 L2(c) 출력

Task-specific input transformations



전제, 가설이 구분자 토큰 사이에 위치

유사도는 순서가 없으므로 text1과 text2를 위치를 바꾸어 적용

✓ 다양한 분야에 맞게 입력을 수정할 필요가 있으며 traversal-style approach 사용

Experiment

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

✓ 7,000권 이상의 다양한 도서로 구성된 BookCorpus Dataset 사용

✓ 특정 task를 위한 다양한 Dataset 사용

Experiment

Natural Language Inference

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

✓ 짝지어진 문장이 중립, 모순인지 판단하는 추론

✓ RTE를 제외하고 GPT에서 모두 SOTA를 달성한 것을 볼 수 있음

Experiment

Question answering and commonsense reasoning

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

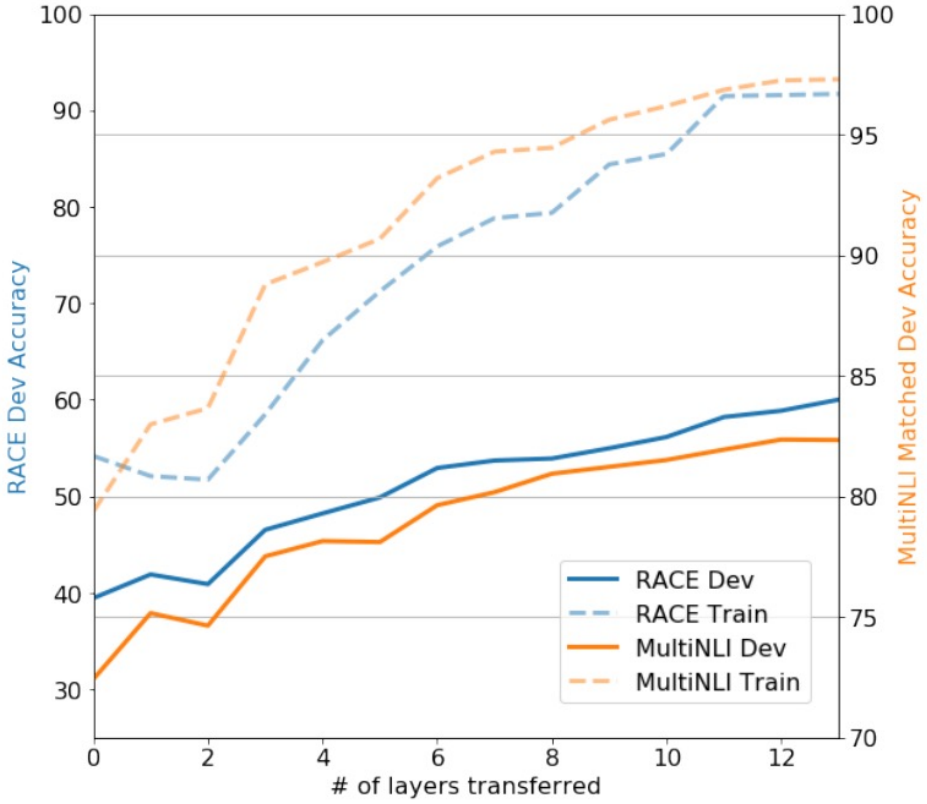
✓ 역시 유의미한 성능 향상을 보임

✓ 전반적으로 12개의 Dataset에서 9개의 SOTA 달성

✓ 이는 작은 Dataset부터 Large Dataset까지 범용적으로 적용할 수 있음을 뜻함

Analysis

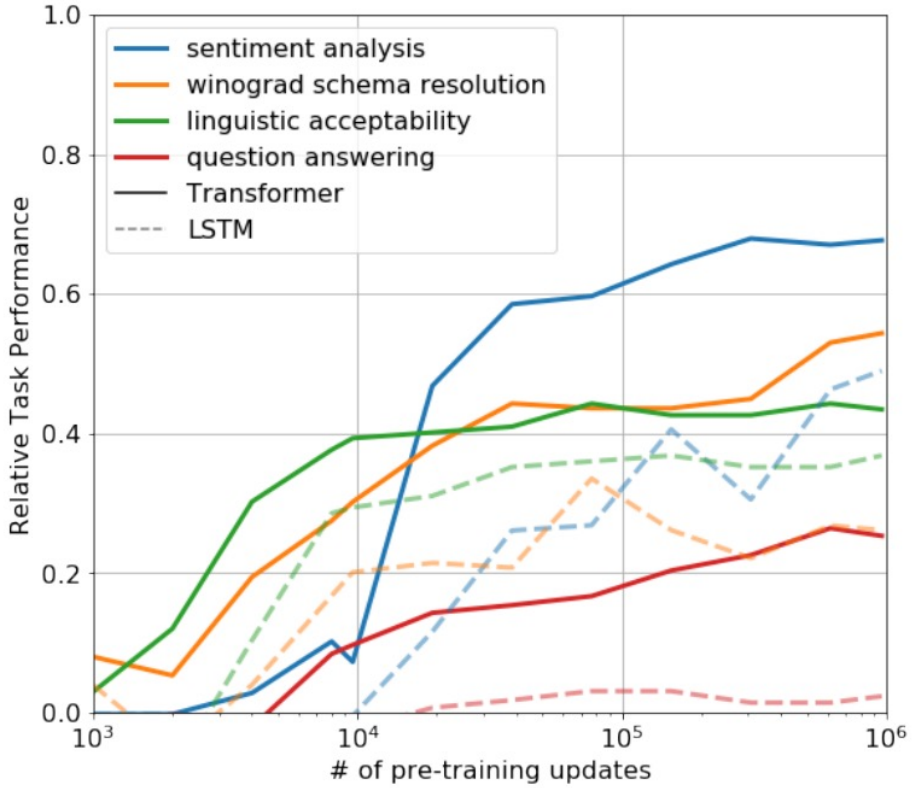
Impact of number of layers transferred



- ✓ Race 데이터와 MutilNLI 데이터 사용
- ✓ X축은 transferred된 레이어의 개수
- ✓ 최대 9%의 성능 향상을 이루어냄으로 사전학습된 layer가 긍정적인 영향을 미친다는 것을 알 수 있음

Analysis

Zero-shot Behaviros



- ✓ 사전학습이 왜 도움이 되는지에 대한 실험 진행
- ✓ Fine tuning 없는 zero-shot generative model 그대로 사용
- ✓ LSTM과 비교시 안정적으로 update 되는 것을 볼 수 있음
- ✓ 한마디로 일반화의 성능이 더 높다는 것을 증명

Conclusion

- ✓ GPT는 labeld data가 부족한 상황에서 해법을 제시함
- ✓ 모델의 간단한 구조 변형과 데이터셋의 활용 방법만으로도 상당한 성능 향상 개선을 할 수 있음을 증명
- ✓ 특히, NLP의 발전을 위해서 꼭 필요한 Unsupervised Learning에 대한 방법론을 제시

감사합니다 ☺