Improving Language Understanding by Generative Pre-Training

발표자: 박채원

0. Abstract

- Unlabeled data는 풍부하지만 labeled data는 풍부하지 않음.
  -> 판별식으로 훈련된 모델이 적절하게 수행하기 어렵다.

- Unlabeled 데이터에 대한 언어 모델의 generative pre-training과 각 특정 작업에 대한 fine-tuning을 통해 이러한 작업에 대한 큰 이득을 실현할 수 있음.

- 이전 접근 방식과 달리 모델 아키텍처에 대한 변경을 최소화하면서 효과적인 transfer을 위해 파인 튜닝 중에 입력 변환을 사용.

- 12개의 작업 중 9개 task에서 SOTA를 달성함.

1. Introduction

- 머신러닝에서 지도학습을 위해 라벨링 된 데이터가 필요되지만, 많은 데이터는 라벨링 돼있지 않음
- 본 논문에서는 레이블링이 되어있지 않은 데이터로 모델을 학습시켜 레이블링 데이터를 이용했을 때의 단점을 극복하고 사람이 알지 못하는 데이터의 특성까지 모델이 학습하게 하고, 이 후 작은 수정만으로 효과적인 transfer를 하게 함으로써 높은 성능을 달성할 수 있다는 것을 입증했다.

- 레이블링 되지 않은 데이터로의 학습의 한계점
  - 1) it is unclear what type of optimization objectives are most effective at learning text representations that are useful for transfer.
  - 2) there is no consensus on the most effective way to transfer these learned representations to the target task.

이 논문에서 제시하고자 하는 것

semi-supervised approach for language understanding tasks using a combination of unsupervised pre-training and supervised fine-tuning.

1. Introduction

- semi-supervised approach
  - 다량의 unlabeled data와 task에 맞는 labeled data가 있다고 가정했을 때, unlabeled data로 모델의 초기 파라미터를 학습하고, 최적화된 파라미터를 원하는 목적에 맞게 labeled data를 이용해 추가학습
  - utilize task-specific input adaptations, which process structured text input as a single contiguous sequence of tokens.

- 사전학습된 모델의 구조에 최소한의 변화를 주고 파인튜닝을 효과적으로 하는 것을 가능하게 함.

2. Related work

- NLP에서의 준지도학습

초기의 연구들은 unlabeled data로 모델이 단어 또는 구문 수준의 통계값들을 연산하고 이를 이후 지도 학습의 특성으로 사용하는 방식을 사용했다. 지난 몇 년간 이루어진 연구들은 word2vec, GloVe와 같은 unlabeled 코퍼스를 이용한 훈련을 통해 단어 임베딩을 사용하는 방식들이 높은 성능을 낸다는 것을 입증했다.

- Unsupervised pre-training

Unsupervised pre-training의 목적은 이 후 수행될 supervised learning에 좋은 초기화 포인트를 제공하는 것이다. 이전에는 이미지 분류, 회귀 문제 등에 이 방법이 사용됐었다. 후속 연구에서 Pre-training 기법은 정규화 작용을 하여 딥러닝 모델을 더 잘 일반화하는 것이 밝혀짐

- Auxiliary training objectives

보조적으로 비지도 학습 목적함수를 추가하는 것 또한 준지도학습의 형태 중 하나이다. 하지만 본 연구에서 시도한 결과 unsupervised pre-training에서 이미 target task와 관련있는 언어적 특성을 충분히 학습했음을 확인함.

3. Framework

1) high-capacity 언어 모델을 학습하고, 2) 특정 task에 맞는 labeled data로 파인튜닝

- unsupervised learning
  - 다음의 likelihood를 최대화 하기 위해 standard language modeling objective를 사용

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

  - 토큰으로 이뤄진 코퍼스 v={u1,u2 … un}, k는 context window의 크기, 세타는 신경망 모델의 파라미터(SGD로 학습됨)

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

- 학습을 위한 언어 모델로 트랜스포머의 디코더를 사용.
- U는 토큰의 문맥 벡터
- n은 레이어의 수
- We는 토큰 임베딩 행렬
- Wp는 포지션 임베딩 행렬

3. Framework

- Supervised fine-tuning
  - 언어모델의 objective에 대해 pre-training 후 labeled dataset으로 target task에 맞게 학습하며 파라미터 조정.
  - 이때, 예측값을 얻기 위해 transformer의 마지막 블락의 activation을 input으로 하는 linear층 추가

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

  - 해당 층은 다음 목적함수를 최대화 하는 방향으로 학습됨

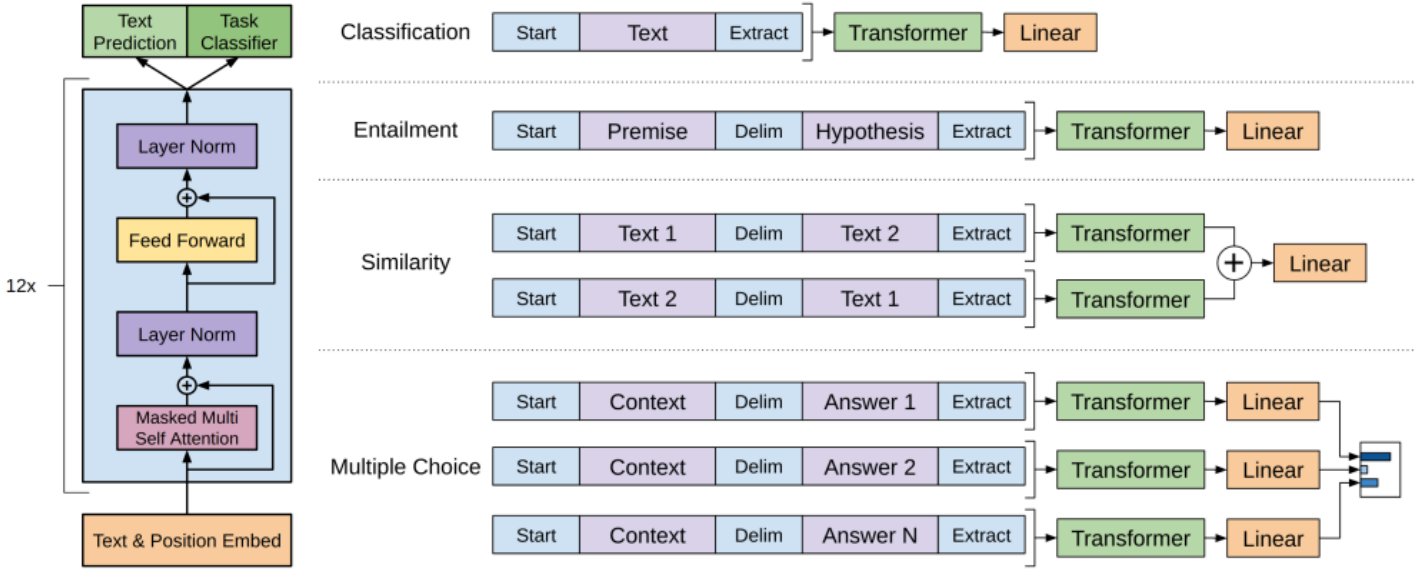
$$L_2(C) = \sum_{(x,y)} logP(y|x^1,\ldots,x^m)$$

  - 또한, 보조 목적으로 언어 모델링을 파인튜닝에 추가하는 것이 1)지도학습 모델의 일반화를 향상시키고, 2)모델이 빠르게 수렴하는데 도움을 줌.
  - 다음 목적 함수의 가중치 λ를 최적화함.

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

3. Framework

- Task specific input transformations



- 이전 연구와 달리 GPT-1은 순회 접근법을 사용해 구조화된 입력을 pre-trained모델이 사용할 수 있는 순서화된 시퀀스로 변환함으로써 약간의 변형만으로 여러가지 task에 적용할 수 있게 됨
- All transformations include adding randomly initialized start and end tokens ([s], [e]).

3. Framework

- Textual entailment task
  - Premise p와 hypothesis h 토큰 시퀀스를 구분 문자(delimiter token) ($)로 concat

- Similarity task
  - 유사도 측정 task에선 비교되는 두 문장에 순서가 없기때문에 고려될 수 있는 순서를 모두 입력을 모두 포함하도록 입력 시퀀스를 구분문자와 함께 수정하고, 각각을 독립적으로 처리한 후 linear 층에 들어가기 전에 더한다.

- Question Answering and Commonsense Reasoning
  - context document z와 question q, 가능한 답변 set인 a_k가 주어지고 z와 q를 구분 문자를 사용해 각 답변과 concat한다.
  - 각 시퀀스별로 독립적으로 처리됨. 최종 출력된 가능한 답변의 분포를 생성해내기 위해 softmax층에 전달됨.

4. Experiments

- Model specifications
  - Pre-training 데이터로 책 7000여권 이상이 포함돼있는 BookCorpus 데이터셋 사용
  - 모델 구조는 transformer의 디코더만 가져옴. 12개의 레이어 사용
  - Optimization – learning rate가 조정된 Adam 사용
  - Batch size – 64 / Epoch – 100
  - Encoding 방식 – BPE
  - 전처리 – fifty liabrary를 이용, 구두점, white space를 표준화, spaCy tokenizer 사용

  - Fine-tuning시에는 pre-trainin에서 사용한 하이퍼파라미터는 그대로 하용하고, lr을 6.25e-5로 수정, 분류 task의 경우 dropout 추가
  - Epoch는 3, batch size는 32

4. Experiments

- 각 task의 Fine-tuning 단계에서 사용된 데이터셋

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]

- Natural language inference task (자연어 추론)
  - 두 문장이 주어졌을때 두 문장의 관계를 수반하는 사이인지, 모순되는 사이인지, 중립 관계인지 판단

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

4. Experiments

- Question answering and commonsense reasoning
  - 긴 텍스트가 포함된 데이터를 평가에 사용하여 long range context도 효과적으로 처리함을 보여줌

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**

- Semantic Similarity
  - 입력된 두개의 문장이 의미론적으로 동일한지 예측
  - 평가에 사용된 세 개의 데이터셋 중 두 개의 데이터셋에서 최고성능을 달성함.

4. Experiments

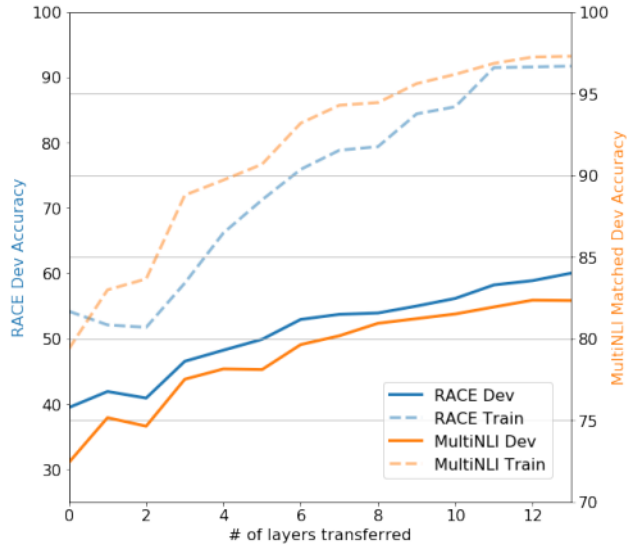
- Classification
  - 데이터셋으로 입력 문장이 문법적으로 옳은지 이진으로 분류하는 CoLA와, 영화 코멘트 데이터를 이진으로 분류하는 SST-2가 사용됨

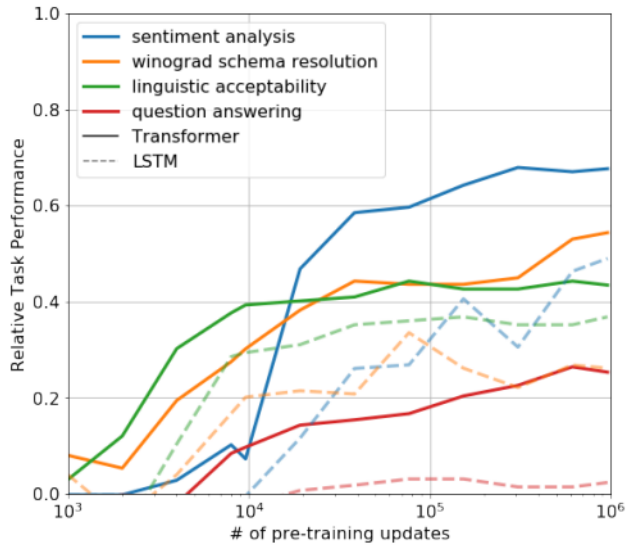
Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

  - 전체적으로 GPT-1 모델은 12개의 dataset 중 9개의 dataset에서 SOTA를 달성함
  - 작은 데이터셋부터 큰 데이터셋까지 좋은 성능을 냄.

5. Analysis

- Impact of number of layers transferred
  - MultiNLI Train에서 full transfer했을때 transformer layer가 9%까지 추가 향상을 이끌어냄
- Zero-shot Behaviors
  - 본 연구는 기본 generative model이 language model의 capability를 향상시키기 위해 많은 task를 수행하는 법을 배우고 transformer의 구조화된 attentional memory가 LSTM에 비해 transfer에 도움이 된다는 것을 가정
  - transformer의 language model pre-training이 효과적인지 이해하기 위해 기본 generative model을 supervised fine-tuning을 하지 않았을 때 task에 대한 성능이 어떤지 확인함.





5. Analysis

- Ablation studies
  (모델이나 알고리즘의 “feature”들을 제거해 나가면서 그 행위가 성능에 얼마나 영향을 미치는지를 확인)
  - 실험을 통해 auxiliary object가 큰 데이터 셋에서는 성능 향상에 도움이 되지만 작은 데이터 셋에서는 아님을 알 수 있었음.

Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	MNLI (acc)	QNLI (acc)	RTE (acc)
Transformer w/ aux LM (full)	74.7	45.4	91.3	82.3	82.0	**70.3**	**81.8**	**88.1**	**56.0**
Transformer w/o pre-training	59.9	18.9	84.0	79.4	30.9	65.5	75.7	71.2	53.8
Transformer w/o aux LM	**75.0**	**47.9**	**92.0**	**84.9**	**83.2**	69.8	81.1	86.9	54.4
LSTM w/ aux LM	69.1	30.3	90.5	83.2	71.8	68.1	73.7	81.1	54.6

6. Conclusion

- 본 논문은 생성적 사전학습과 특정과제에 특화된 파인튜닝을 통해 학습된, task에 대해 별다른 지식이 없으며 자연어이해 능력이 뛰어난 단일 모델(framework)를 소개.

- 넓은 분야의 다양한 말뭉치에 대해 사전학습을 진행하여 중요한 일반지식과 질답, 의미론적 유사성 평가, 함의 확인, 문서분류 등의 task에서 성공적으로 transfer되는 장거리 의존성을 처리하는 능력을 학습하여 12개 중 9개의 과제에 대해 state-of-the-art를 달성함.