Improving Language Understanding by Generative Pre-Training

Alec Radford, et al

OpenAI

2018

Abstract

- Large unlabeled text corpora
  - It could not be used for training

- Solution to unlabeled data
  - *Generative pre-training* (Unsupervised)
  - *Discriminative fine-tuning* (Supervised)

- Experiment with benchmarks
  - We show effectiveness of our approach

Introduction

- Need for unsupervised NLP
  - Manually labeling data is time-consuming
- Challenge: What type of objectives are effective?
  - Translation[1], Predicting next word[2], etc.
  - Each works better for some tasks than others
- Challenge: How to Transfer the Learned Knowledge?
  - Changing architecture for each task
  - Complicated fine-tuning for each task
  - Adding auxiliary objectives

---

[1] Bryan McCann et al. "Learned in translation: Contextualized word vectors". In: *Advances in neural information processing systems* 30 (2017).

[2] Matthew E Peters et al. "Dissecting contextual word embeddings: Architecture and representation". In: *arXiv preprint arXiv:1808.08949* (2018).

Introduction

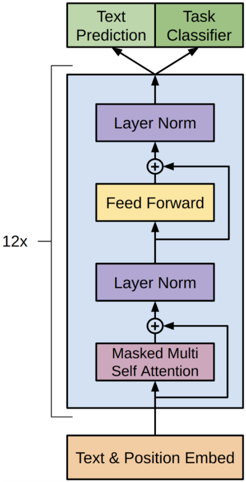
- We suggest semi-supervised approach
  - Unsupervised Pre-training by predicting next word
  - Supervised Fine-tuning for specific task

- For our architecture, we use the Transformer
  - It outperforms others (RNNs, LSTMs)

- Experiment - Four types of tasks
  - NLI, QA, Semantic Similarity, Text Classification
  - SOTA on 9 out of 12 benchmarks

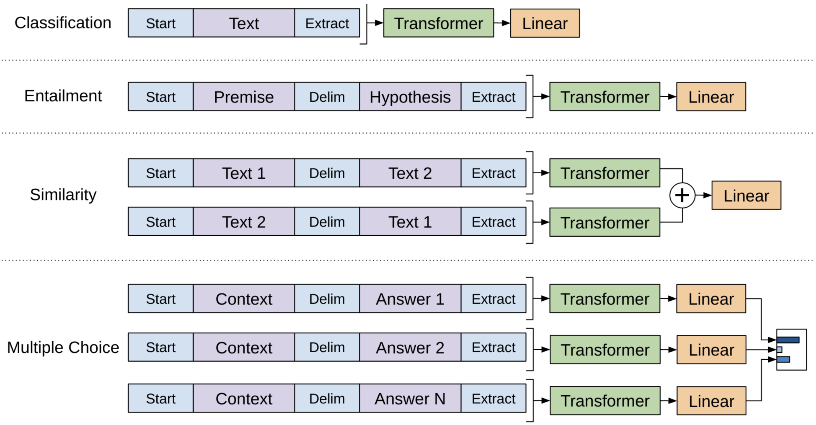
Framework

- Unsupervised pre-training
  - We uses decoder-only transformer [1]
  - $L_1(U) = \sum_i \log P(u_i|u_{i-k}, \cdots, u_{i-1}; \Theta)$
- Supervised fine-tuning
  - $L_2(U) = \sum_{(x,y)} \log P(y|x^1, \cdots, x^m)$
  - $L_3(U) = L_2(U) + \lambda * L_1(U)$
  - Auxiliary objective improves generalization



[1] Peter J Liu et al. "Generating wikipedia by summarizing long sequences". In: *arXiv preprint arXiv:1801.10198* (2018).

Framework



- Task-specific input transformations

Experiment

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]

- Pre-training
  - BooksCorpus dataset containing 7,000 books [1]
- Model specifications
  - Decoder only transformer with 12 layers (unlike BERT)
  - Learned positional embedding instead of sinusoidal
  - Attention - 12 heads and 768 dimension

[1] Yukun Zhu. "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books". In: *arXiv preprint arXiv:1506.06724* (2015).

Experiment

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

- Natural Language Inference
  - Given two sentences, determine their relationship
  - Entailment, Contradiction, Neutral
  - GPT outperformed previous SOTA on 4 out of 5
  - RTE is small dataset, making it harder to adapt

Experiment

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**

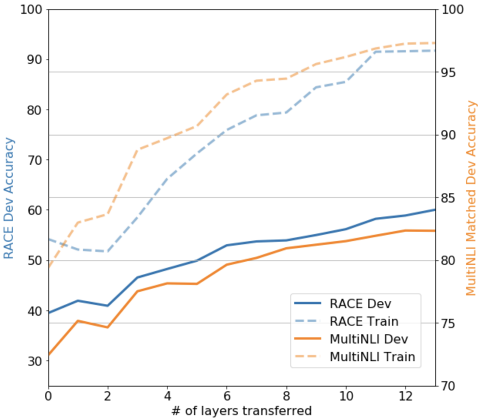
- QA and commonsense reasoning
  - Story Cloze: Complete a multi-sentence story
  - RACE: Middle/High School Exams
  - GPT outperforms prior SOTA models
  - Transformer allows it to capture long-range dependencies

Experiment

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]	-	-	-	<u>81.0</u>	-	-
Single-task BiLSTM + ELMo + Attn [64]	<u>35.0</u>	90.2	80.2	55.5	<u>66.1</u>	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	<u>68.9</u>
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

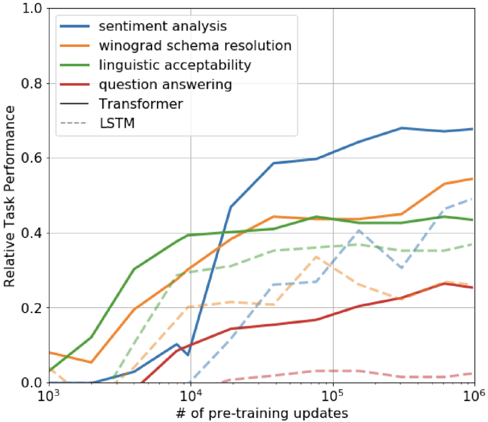
- Semantic Similarity
  - Determining if two sentences express the same idea
  - SOTA results on 2 out of 3 datasets
- Classification
  - CoLA: Is a sentence grammatical?
  - SST: Is a review positive or negative?

Analysis



- Impact of number of layers transferred
  - Fine-tuning all layers may cause overfitting
  - However, full model transfer leads to best results

Analysis



- Zero-shot Behaviors (No supervised fine-tuning)
  - Performance improves throughout training
  - It means pre-training develops general reasoning abilities

Analysis

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

- Ablation studies
  - Without pre-training, there is massive performance drop
  - Without auxiliary objective, performance drop in larger dataset
  - Using LSTM, model struggles with long-term dependency

Conclusion

- Task-agnostic model
  - Previous models required task-specific architectures
  - GPT uses one model for multiple NLP tasks
- Unsupervised learning
  - NLP models relied heavily on supervised learning
  - GPT showed that pre-training on raw text boosts performance
- Trained on long-form contiguous text
  - Helps capture long-range dependencies in language
  - Provides world knowledge for solving downstream NLP tasks