부스트캠프 AI Tech 2기 LEVEL2 NLP-2 논문리뷰

boostcamp ai tech

GPT-1

Improving Language Understanding by Generative Pre-Training

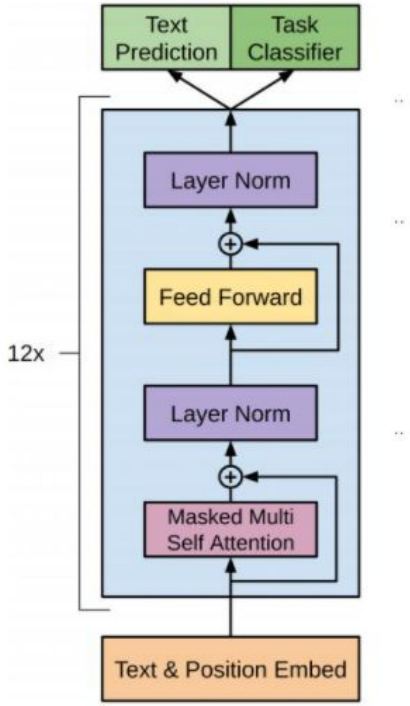
Email : iamtrueline@gmail.com

GitHub : https://github.com/iamtrueline

Abstract

- 일반적으로 Labeled Data보단 Unlabeled Data의 절대량이 많은 현실.
- Unlabeled Data라고 해서 사용할 수 없는 것은 아니나, 두 가지 문제점이 존재.
- 무엇을 학습해야 할지? (Pre-train objective)
- 어떻게 일반화해야 할지? (전이)
- 본 논문에서 이를 Semi-Supervised Approach로 정리.

Introduction



- Semi-Supervised Approach = Unsupervised pre-training + Supervised fine-tuning
- Transfomer Decoder 사용. cf) Attention is all you need
- 실험 결과 최소한의 수정만으로(최소 파라미터 추가) 효과적인 결과 도출.
- 12개 테스크 중 9개 SOTA 달성.

Framework

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

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

- Unsupervised pre-training

Framework

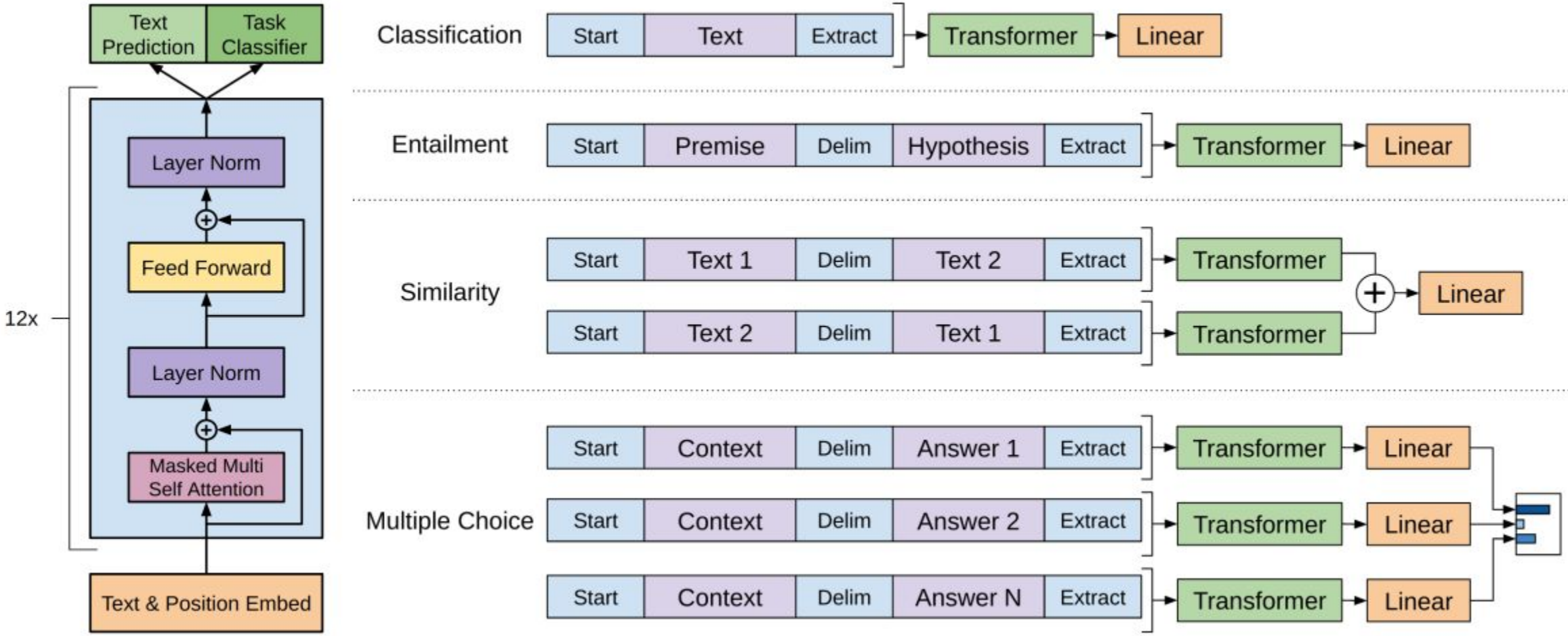
$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y).$$

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

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

- Supervised fine-tuning

Framework



- Task에 따라 구조 변경.

Experiments

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

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**

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**

Further + Conclusion

Table 5: Analysis of various model ablations on different tasks. Avg. score is a unweighted average of all the results. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

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

- 레이어 개수 비교 : 레이어 추가시 성능 상승. 12에서 수렴.
- Pre-trained 유무 비교 : Pre-trained 단이 있을 경우가 성능 15%상승. 단, Dataset의 크기가 작을 경우 fine-tunning만 하는 것이 더 좋은 성능을 보임.
- LSTM 비교 : 모든 Task에서 LSTM보다 좋은 성능.