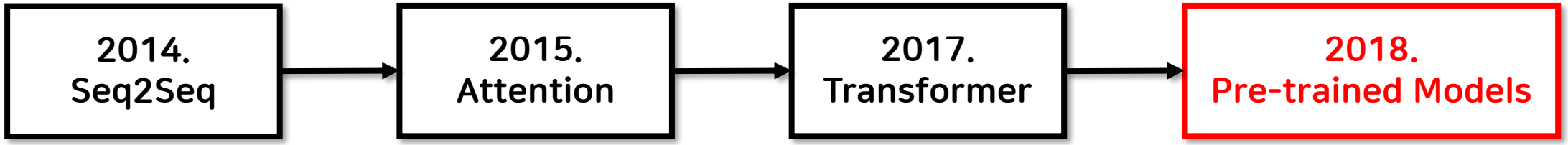
Improving Language Understanding by Generative Pre-Training

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever

단국대학교 모바일시스템공학과 양윤성

Contents

1-(1). Background



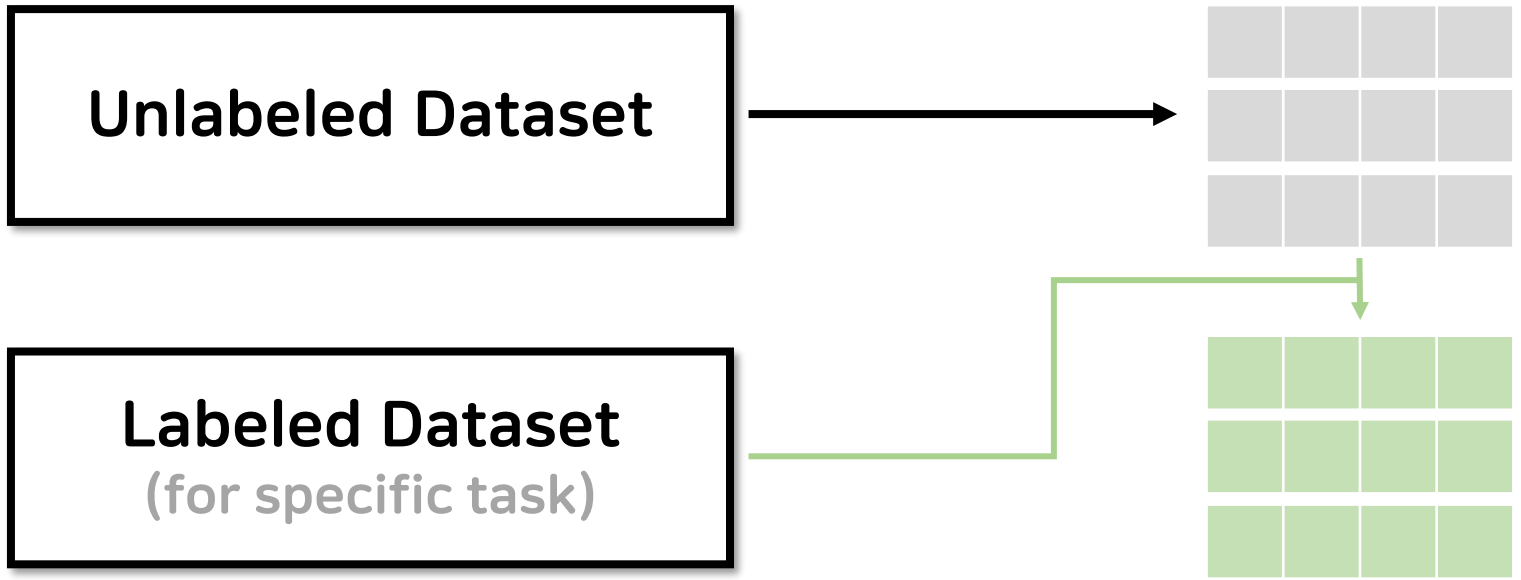
- Word representation learning search (ex. word embedding)
- How we apply transformer to downstream tasks?

➢ Advent of pre-trained language model

1-(2). Main Idea

- Previously, labeled dataset was used for learning
- However, unlabeled dataset > labeled dataset
- Idea : Use unlabeled dataset to increase model performance
- Make task-agnostic model that can be transferred to many tasks

1-(2). Main Idea



- Find embedding vector from unlabeled dataset to use GPT model
- Then discriminative fine-tuning on labeled task → much better results

Contents

2-(1). Unsupervised pre-training

- Learning with lots of unlabeled data
- Pre-training language features (grammar, vocabulary, context…)
- Given an unsupervised corpus of tokens, use L1 function

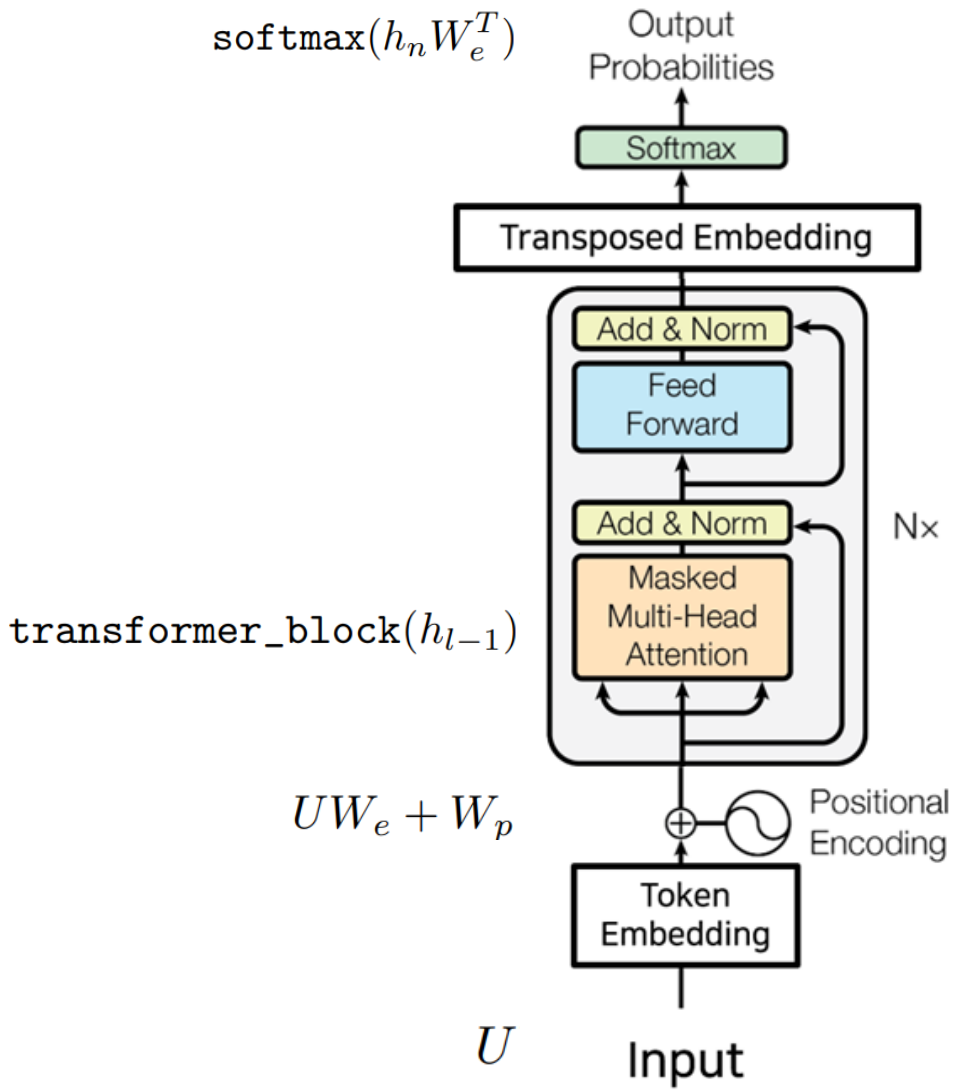
$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

- L1 : Objective functions for learning model → Maximize likelihood
- k : size of context window (number of token)
- P : conditional probability

2-(1). Unsupervised pre-training

- Multi-layer transformer's decoder
- Except encoder-decoder self attention

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



2-(2). Supervised Fine-Tuning

- Adjust to be more suitable for actual task (fine-tuning)
  - Label y : Input tokens {x1, ⋯, xm}
  - h_l^m : Final transformer block's activation (output)

➢ L2 : Objective functions for supervised learning

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

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

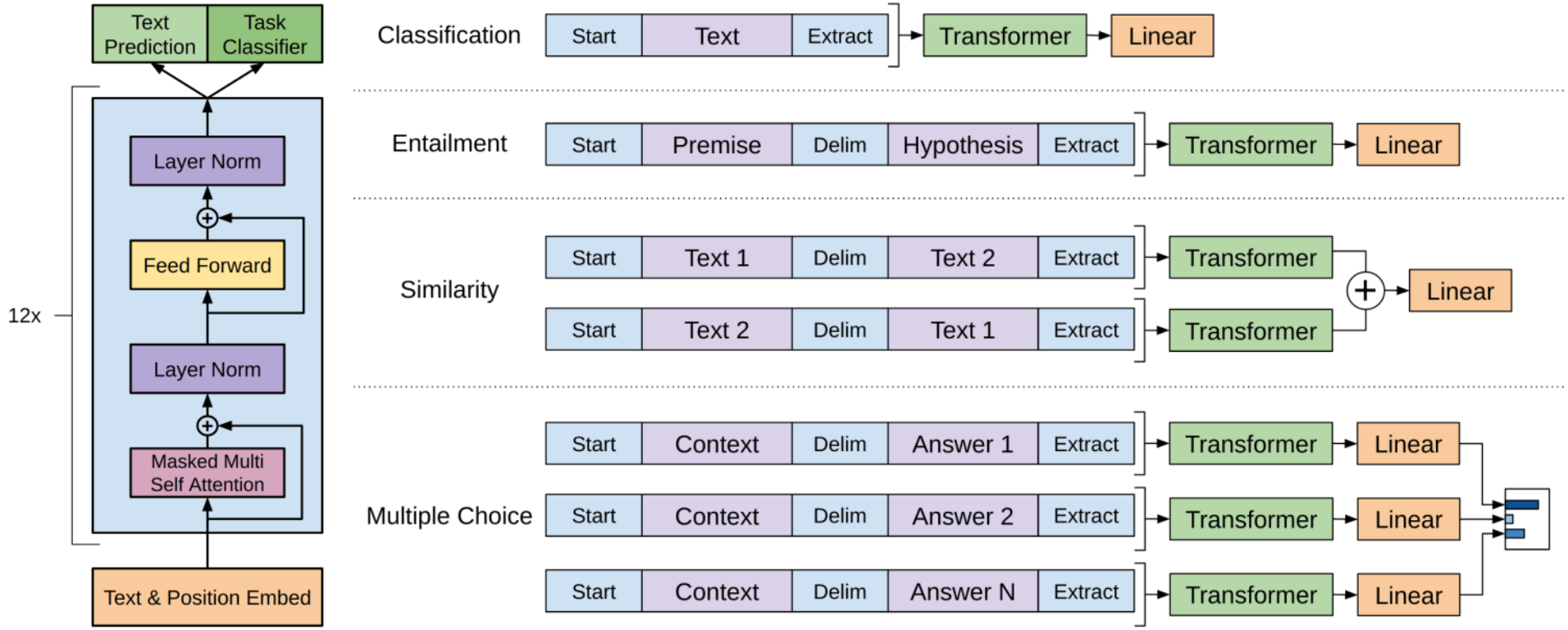
2-(2). Supervised Fine-Tuning

- Method of maximize learning efficiency
  - Re-update L1 with supervised corpus(C) and multiply weight
  - Combine with L2(C)
- Improving generalization of the supervised model
- Accelerating convergence

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

2-(3). Task-Specific Input Transformations

- Fine-tuning to different tasks
- Each tasks have different structure of token sequence(input)



2-(4). Evaluation Tasks

1. Natural Language Inference

- Find relationship between text and hypothesis
- Express relationship as “judgements” (Contradiction/Neutral/Entailment)

Text	Judgement	Hypothesis
A black race car starts up in front of a crowd of people	Contradiction	A man is driving a lonely road
A smiling costumed woman is holding an umbrella	Neutral	A happy woman in a fairy costume holds an umbrella
A soccer game with multiple males playing	Entailment	Some men are playing a sport

2-(4). Evaluation Tasks

2. Semantic Similarity

- Calculate similarity between two sentences as score

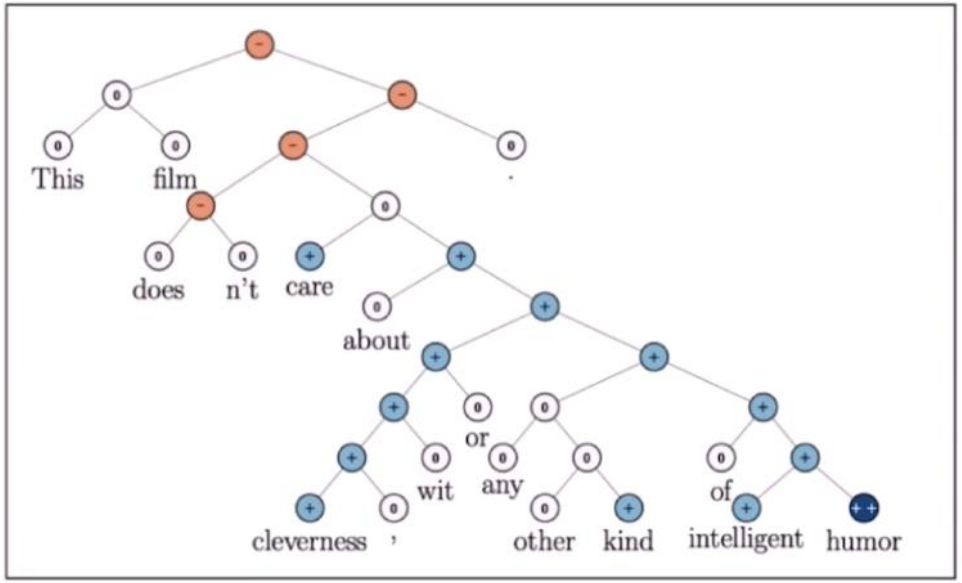
Score	English	Spanish
5/4	*The two sentences are completely equivalent, as they mean the same thing.*	
	The bird is bathing in the sink. Birdie is washing itself in the water basin.	El pájaro se esta bañando en el lavabo. El pájaro se está lavando en el aguamanil.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*	
	In May 2010, the troops attempted to invade Kabul. The US army invaded Kabul on May 7th last year, 2010.	
3	*The two sentences are roughly equivalent, but some important information differs/missing.*	
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.	John dijo que él es considerado como testigo, y no como sospechoso. "Él ya no es un sospechoso," John dijo.
2	*The two sentences are not equivalent, but share some details.*	
	They flew out of the nest in groups. They flew into the nest together.	Ellos volaron del nido en grupos. Volaron hacia el nido juntos.
1	*The two sentences are not equivalent, but are on the same topic.*	
	The woman is playing the violin. The young lady enjoys listening to the guitar.	La mujer está tocando el violín. La joven disfruta escuchar la guitarra.
0	*The two sentences are completely dissimilar.*	
	John went horse back riding at dawn with a whole group of friends. Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.	Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. La salida del sol al amanecer es una magnífica vista que puede presenciar si usted se despierta lo suficientemente temprano para verla.

2-(4). Evaluation Tasks

3. Classification

- CoLA dataset : Task of classifying whether grammar is correct or incorrect
- SST-2 dataset : Task of classifying sentiments in a sentence

Label	Sentence
*	The more books I ask to whom he will give, the more he reads.
✓	I said that my father, he was tight as a hoot-owl.
✓	The jeweller inscribed the ring with the name.
*	many evidence was provided.
✓	They can sing.
✓	The men would have been all working.
*	Who do you think that will question Seamus first?
*	Usually, any lion is majestic.
✓	The gardener planted roses in the garden.
✓	I wrote Blair a letter, but I tore it up before I sent it.



2-(4). Evaluation Tasks

4. Question Answering

- Find answer to question about fingerprint

지문	We report a case of a 72-year-old Caucasian woman with pl-7 positive antisynthetase syndrome. Clinical presentation included interstitial lung disease, myositis, mechanic's hands and dysphagia. As lung injury was the main concern, treatment consisted of prednisolone and cyclophosphamide. Complete remission with reversal of pulmonary damage was achieved, as reported by CT scan, pulmonary function tests and functional status. [...]
질문	Therefore, in severe cases an aggressive treatment, combining ____ and glucocorticoids as used in systemic vasculitis, is suggested.
정답	cyclophoshamide

Contents

3-(1). Compare Scores of Tasks

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]

- **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

3-(1). Compare Scores of Tasks

- **Question Answering**

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**

- **Classification and Semantic Similarity**

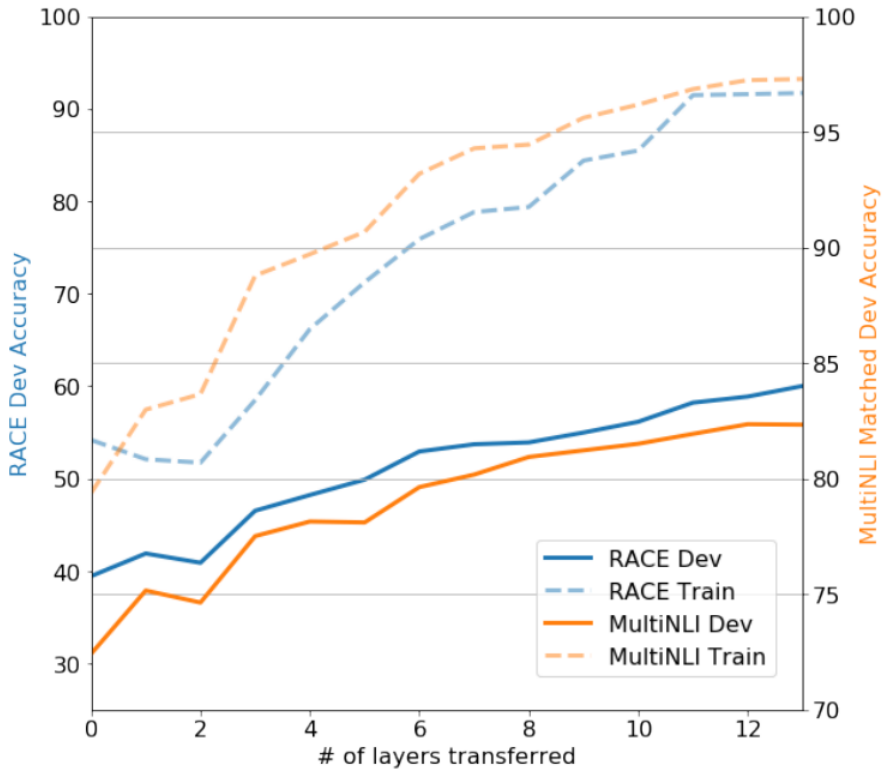
Method	**Classification**		**Semantic Similarity**			**GLUE**
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task 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**

Better performance than traditional models

(except RTE, SST-2, MRPC)

3-(2). Analysis Graph

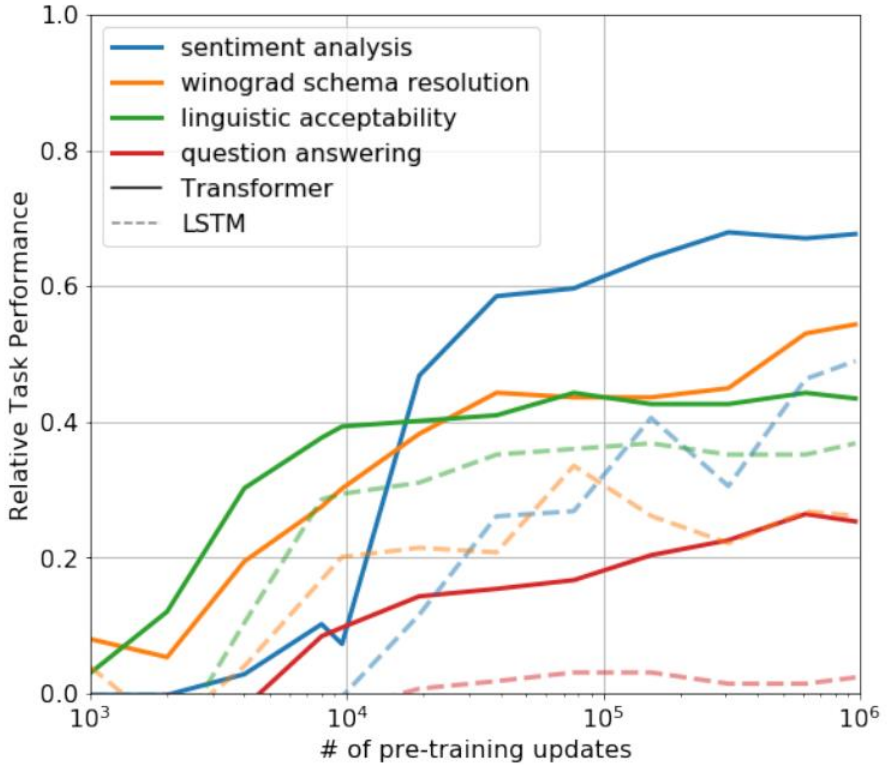
- **Impact of number of layered transferred**



- Test how many transformer decoder blocks should be stacked for efficiency
- Accuracy increases as it accumulates, but converges from 12

3-(2). Analysis Graph

- **Zero-shot behaviors**



- Comparison of unsupervised learning first and only supervised learning
- Better performance than LSTM across all parts

3-(3). Model ablations on different tasks

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

- Performance drops by 15% without pre-training
- Auxiliary objective function : Only benefit from large datasets

Contents

4. Conclusion

- Active use of unsupervised learning
  - Use unlabeled data → More data available
  - Pre-Training and Fine-Tuning → Increased learning accuracy
- Successful application of Transformer architecture

Thank You!

Improving Language Understanding by Generative Pre-Training