Improving Language Understanding by Generative Pre-Training (GPT-1)

Paper Review by Grace

1. Introduction

> *Most deep learning methods require substantial amounts of manually labled data, which restricts their applicability in many domains that suffer from a darth of annotated resources.*

- Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce.
- 모델을 훈련하기 위한 labeled 데이터가 부족하고 unlabeled 데이터는 상당히 많음. labeled 데이터의 양은 한정적이므로 모델의 적용 범위가 한정됨.

👩‍💻 So why is it hard to use unlabeld data?

- Unlabeled data has no labels or targets to predict, only features to represent them. Basically RAW data.

> *First, it is unclear what type of optimization objectives are most effective at learning text representations that are useful for transfer.*

> *Second, There is no consensus on the most effective way to transfer these learned representations to the target task.*

- Semisupervised language model, Labeled 데이터가 부족하여 새로운 기법을 제시한 것.

👩‍💻 What this paper presents

> *A semi-supervised approach using a combination of unsupervised pre-training and supervised fine-tuning.*

> *First, we use a language modeling objective on the unlabeled data to learn the initial parameters of a neural network model.*

> *Subsequently, we adapt these parameters to a target task using the corresponding supervised objective.*

- 모델 학습이 unlabeled 데이터를 이용한 unsupervised pre-training 단계 및 구체적인 task (목적함수)로 fine-tuning 을 한 supervised 훈련 단계로 나뉨. 즉, semi-supervised 이다!

> *Our goal is to learn a universal representation that transfers with little adaption to a wide range of tasks.*

- 양이 많은 unlabled 데이터로 훈련을 하였기 때문에 조금 더 보편적인 모델이 되면서, 구체적인 task 으로 fine-tuning 과정을 거쳤기 때문에 더 정확도가 증가한다고 볼 수 있다.

👩‍🏫 필요한 사전지식

Language model: Predicts the next token using given tokens.

Ex) The search recommendations in Google!

Generative VS Discriminative Model

https://youtu.be/HHNESCbZqUg

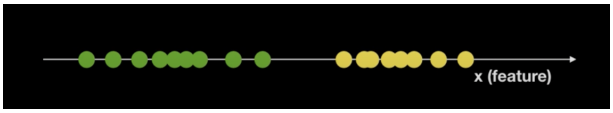
- Generative models model the distribution of individual classes. 각각 클래스의 분포도를 이용한다! 두 클래스가 어떻게 다른가에 초점을 맞추고 있음.
  - Generative models are interested in “How X(feature) and Y(label) occur at the same time”

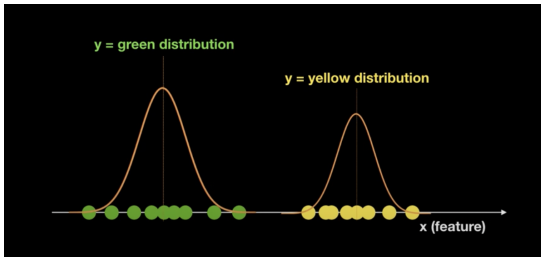
$$p(x, y) = p(y) * p(x|y)$$

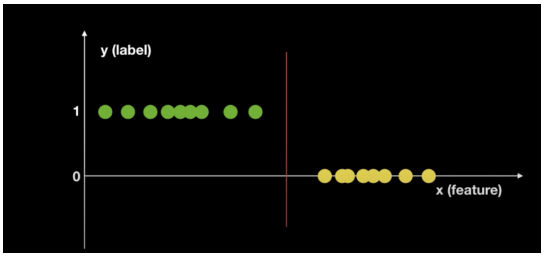
- Discriminative models learn the bounaries between classes.

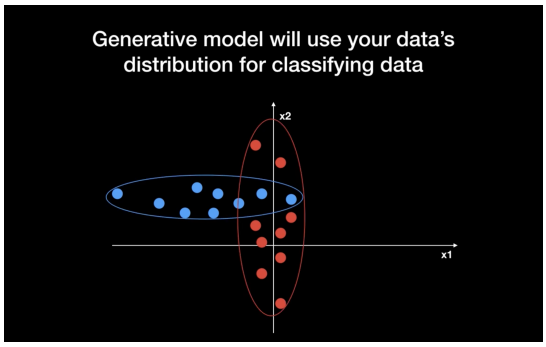
$$p(y|x)$$

👩‍💻 Example: y1=green, y2=yellow 라면 분류하고자 하는 데이터 (x) 와의 joint probability 가 높은 것으로 간주!





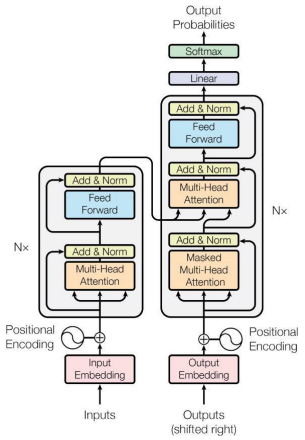




	Generative Model	Discriminative Model
개념	데이터를 생성하기 위한 모델이며 데이터가 생성될 확률인 P(x,y) 를 이용함.	데이터를 구별하기 위한 모델이며, 데이터가 주어졌을 때 특정 클래스에 속할 확률인 P(y|x) 이용함.
Labeling	unlabel 가능	needs label
Example	강아지, 고양이 이미지 각각의 특징을 학습한 후에 강아지와 고양이 이미지로 분류하는 과정	강아지, 고양이 이미지의 차이를 학습한 후에 강아지와 고양이로 분류하는 과정
In GPT?	Unlabled data 를 사용해 보편적인 표현 (universal representation) 을 학습함.	Labeled data 를 활용해 특정 task 에 fine-tuning 하는 과정에 적용

- Advantages of generative models: less chance of overfitting and misclassifying!
- Disadvantages of generative models: outliers can affect your model performance. Need enough data and computation.

Transformer



- The model design introduced in “Attention is All You Need”.
- Encoder, Decoder architecture.
- By using attention rather than sequences, transformers can take the input at once.
- Before transformers: Model takes tokens in one by one sequentially in the decoder, and the product enters the deocder RNN cell.
- With transformers: By using positional encoding, the model can understand where the word is without it being entered sequentially. Replaces the RNN layers, and has multi attention layers.

- Work before? Pre-training phase with LSTM —> restricts the prediction ability to a short range.

---

2. GPT-1 Framework

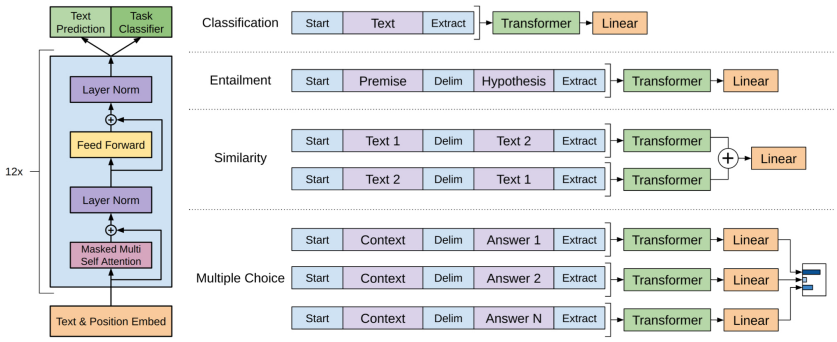
2.1 Unsupervised pre-training

- Transformer 의 Decoder 구조를 변형하여 사용함.
- Transformer base model 은 6개의 layer 를 가지고 있는 반면, GPT-1 은 12개 임.
- Standard language modeling objective of maximizing the likelihood.

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

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

2.2 Supervised fine-tuning



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

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

But for better generalization and faster convergence, we have an auxiliary learning objective as below.

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

- Pre-trained model 에 Linear Output Layer 을 추가함.
- Pre-training 에 사용된 linear model 의 최적화 함수를 함께 사용하여 더 빨리 수렴하고 일반화 능력도 향상됨.

> *Certain tasks have structured inputs such as ordered sentence pairs, or triplets of document, question, and answers. We convert structured inputs into an ordered sequence that our pre-trained model can process. These input transformations allow us to avoid making extensive changes to the architecture across tasks.*

- Pre-trained 된 모델은 unlabled 로 학습하였기 때문에, 입력을 연속적인 단어의 형태로 받아야함. 하지만, labled 데이터는 그러한 형태가 아니기 때문에, 연속적인 단어의 형태가 될 수 있도록 입력의 형태를 바꾸어주어 fine-tuning 을 진행함.

<Specific Task 의 종류>

a. **Natural Language Inference**
   - 문장 간 관계의 종류: entailment, contradiction, netural
   - 문장의 참, 거짓으로 구분 지음.
   - Ex) A man is sleeping. (TRUE) && A man is eating dinner. (FALSE) —> contradiction 관계이다!

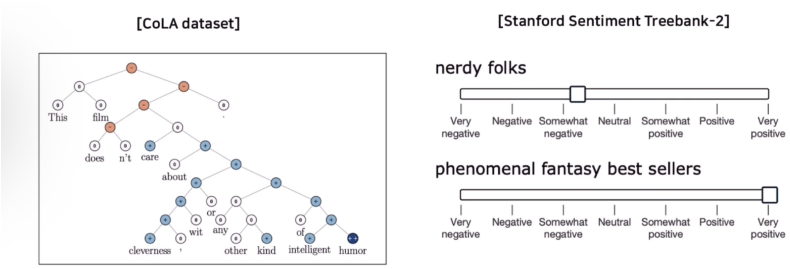
b. **Semantic Similarity**

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.

- 주어진 두 문장 간 유사한 정도를 계산하여 점수로 나타냄.
- Predicticting whether the two sentences are semantically equivalent or not.

c. Question Answering & Commonsense reasoning

d. Classfication



- Grammatically correct or not?
- Binary Classficiation Task

---

3. Experiments (훈련)

Training Sets

- Unsupervised pre-training : BooksCorpus, 1B Word Benchmark.
- Supervised fine-tuning
- Supervised vs unsupervised learning

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]

Model Specifications & Fine-tuning Details

- 12-layer decoder-only decoder transformer with masked self-attention heads

$$L3(C) = L2(C) + 0.5L1(C)$$

Results

- Natural Language Inference Tasks

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

- Question answering and commmonsense 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**

- Semantic similarity and classficiation results

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**

- *New state-of-the art results in 9 out of the 12 datasets!*

Conclusion

- GPT shows that we can transfer unlabeled data to solving specific tasks!
- Focus more on unsupervised learning.

++ GPT-2, GPT-3, GPT-4??

The Journey of Open AI GPT models

Generative Pre-trained Transformer (GPT) models by OpenAI have taken natural language processing (NLP) community by storm by introducing very powerful language models. These models can perform various NLP tasks like question answering, textual entailment, text summarisation etc. without any

https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2

- GPT-2 (Task Conditioning & Zero-shot learning)

$$P(output|input, task)$$

같은 input 에 대해서도 task 이 다르면 다른 output 이 나오도록

Task 가 무엇인지 정확히 명시하지 않아도 알아서 수행하도록 하는 zero-shot learning (ex: 한국어 + 영어 문장 —> 번역이다!)

- GPT-3

> https://www.youtube.com/watch?v=jz78fSnBG0s

++ 구현해보기

https://paul-hyun.github.io/gpt-01/