GPT 1 to 3

리뷰

GPT1

Abstract

GPT1은 비지도 사전학습, task별 적합한 fine-tuning

Introduction

- unlabeled text로 부터 word-level 정보를 넘어선 활용에는 2가지 어려움
  - text를 잘 표현하기 위한 학습에 어떤 loss가 사용되는 것이 최적인지 불분명
  - task별로 가장 효과적인 사전학습 representation이 다름.
- 본 논문에서는 비지도 사전학습과 지도 파인튜닝의 조합으로 다양한 범주의 task에 범용하고자 함.
- 2 step 학습과정을 거침
  - 1st, NN에 초기 파라미터를 설정하기 위해 unlabeled data에 LM Objective를 사용함.
  - 이 파라미터를 target task에 해당 supervised objective를 사용함.
- Transformer구조를 사용하였고, transfer를 할 때, task에 적합한 input으로 하나의 연속토큰시퀀스를 사용함.

Framework

- Unsupervised pre-training

  표준 LM objective를 사용함. muti-layer Transformer decoder를 사용함.

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

LM을 학습하는 LOSS FUNCTION인데 product rule이 사용되었고, 언어의 문맥을 이해하는데 사용되는 것으로 추론됨.

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

- Supervised fine-tuning

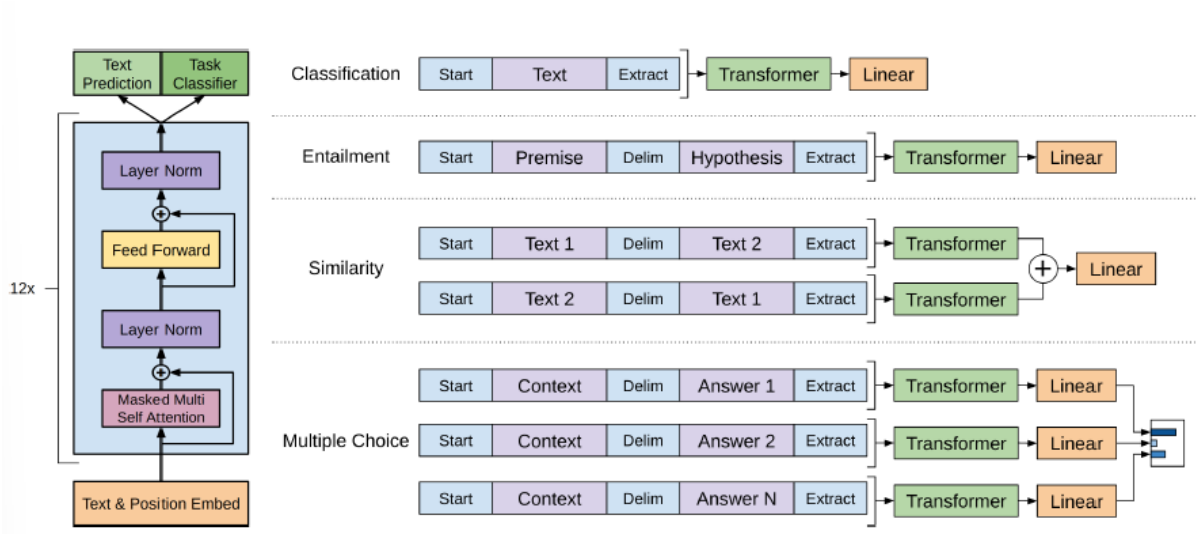
$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y).$$

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

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

사전학습된 모델에서 input이 h_l을 출력하고 새로운 layer를 추가하여 W_y를 학습한다. 그 때의 출력이 2번째 objective function이고 기존의 L1과 L2를 복합적으로 반영하여 L3를 구성하고 이를 학습시켜 W_y 가중치를 학습한다.

- 이렇게 학습할 경우 사전학습된 모델의 input과 유사한 형태의 task만 타파가능한 문제가 발생. 이를 해결하고자 fine tuning을 위한 task에 맞게 input을 커스텀하는 방법을 제안했다.



- Classification의 경우는 그냥 그대로 사용하면 됨.

- Textual entailment의 경우는 a delimiter token ($) in between premise and hypothesis를 해서 이진 분류 task로 변환하면 됨.
- Similarity의 경우는 2문장사이의 고유한 순서가 없기 때문에 2가지 경우에 대해 모두 objective를 추출하여 concat하고 finetuning하는 방법을 적용함.
- QA, Commonsense Reasoning의 경우는 문서 z, 질문 q, 잠재 정답 ak가 있을 때, [z; q; $; ak] 다음과 같이 concat을 하고 모든 잠재 정답에 대해서 가각의 output을 도출하여 가장 확률이 높은 잠재 정답을 정답으로 출력하는 방식을 적용함.

Conclusion

the state of the art on 9 of the 12 datasets

We hope that this will help enable
new research into unsupervised learning, for both natural language understanding and other domains,
further improving our understanding of how and when unsupervised learning works.

GPT2(Language Models are Unsupervised Multitask Learners)

Abstract

GPT-2는 대량의 웹 페이지 데이터셋(WebText)을 통해 훈련되며, 이를 통해 기계 번역, 질문 응답, 독해, 요약 등 다양한 자연어 처리 작업을 명시적인 감독 없이도 수행할 수 있게 되었다. zero-shot task transfer를 가능케했고, 모델 용량을 늘릴수록 다양한 작업에서 로그-선형적으로 성능이 개선됨을 보였다.

Introduction

transfer learning에 대해 architecture나 parameter 수정 없이 zero-shot learning을 가능하게하고자함.

Approach

multitask라는 것은 동일한 input에 대해 다양한 task를 수행하기 때문에 입력값에는 input과 task가 됨.

하지만 GPT2는 비지도로도 task의 구분이 가능해짐.

LM은 기본적으로 문장을 생성하거나 이해하는 작업을 수행한다.

Approach-Dataset

데이터셋은 약 4500만개의 link에서 추출한 text로 구성되어있고, 총 800만개가 넘는 문서, 40GB의 text로 구성되어있다.

Approach-Input Representation

유니코드 문자열을 UTF-8 바이트의 시퀀스로 처리

Byte Pair Encoding (BPE)는 글자와 단어 수준 언어 모델 사이의 실용적인 중간 방식으로, 빈번한 기호 열에 대해 단어 수준 입력을, 드물게 등장하는 기호 열에 대해 글자 수준 입력을 효과적으로 보완

→한글자단위로 token을 생성하되 자주 붙는 글자들은 하나로 합쳐서 token을 생성함.

Approach-Model

GPT1의 모델을 따르되 약간의 수정을 거쳤다.

- 수정 사항
  - Layer Norm이 각각의 sub-block의 input에 위치함.
  - 최종 self-attention layer이후에 LN이 추가됨.
  - 단어수는 50,257. the context size가 512에서 1024 tokens이 되었고, batchsize는 512로 설정함.

Conclusion

 충분히 크고 다양한 데이터셋에서 LLM을 훈련시키면 여러 domain과 dataset에서 잘 수행할 수 있음.

지도 학습 없이 제로샷에서 성능이 나오기 시작한 모델.

GPT3

Abstract

모든 taks에 대해서 GPT3는 gradient updates나 fine-tuning 없이 많은 NLP task에서 우수한 성능을 보였다고 한다.

Meta-Learning은 학습 과정에서 다양한 스킬과 패턴인식 능력을 동시에 키워, Inference 단계에서 원하는 task에 빠르게 적응할 수 있도록 모델을 학습시키는 방법이다.

GPT-2에서 모델의 크기, 데이터셋의 크기, 다양성, 학습 횟수를 전반적으로 늘린 모델에 불과하다.

Introduction

NLP task에서 많은 발전을 이뤘지만, task-specific dataset과 fine-tuning이 필요했다.

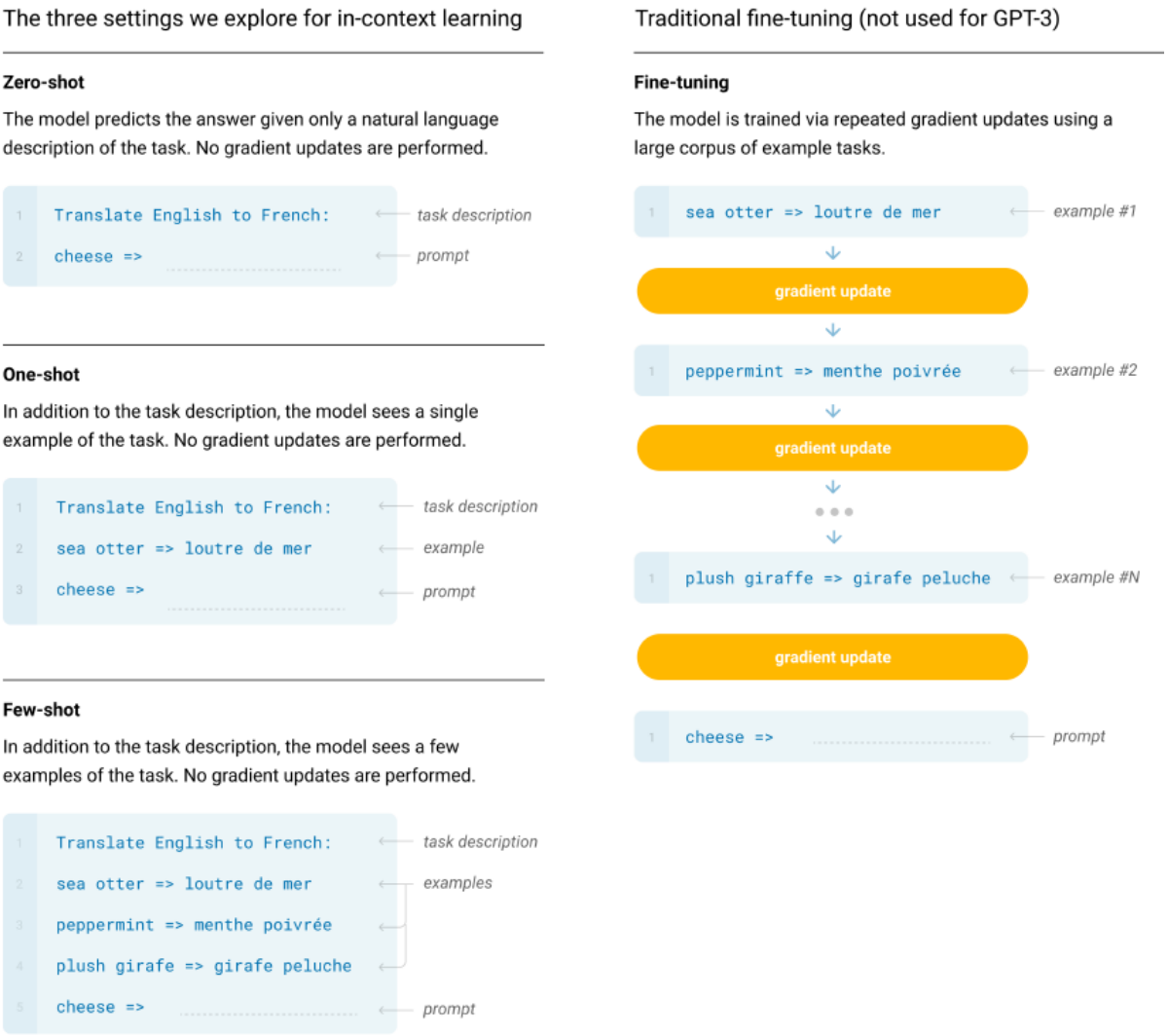
인간→ language task를 처리하기 위해 large supervised dataset이 필요하지 않다.

이러한 점이 현재 NLP기술의 conceoptual limitation을 시사한다. →이러한 문제를 해결하기 위해 meta-learning 도입.(fine-tuning보단 별로)→ 그러나 transformer를 사용하고 상당한 parameter를 도입함에 따라 컨텍스트 내 학습 능력이 규모에 따라 크게 향상될 가능성을 보였다.

본 논문에서는 1750억개 파라미터를 가진 AR language model로 in-context learning을 측정하였다. GPT3는 few-shot learning에 대해 SOTA를 달성하거나 능가한 task도 존재하고, one-shot이나 zero-shot에 대해 높은 성능을 보였다.

GPT3는 다양한 능력을 보여주었지만, bias, fairness, and broader societal impacts도 중요한 논의 내용이다.

Approach



FS(Few-Shot)은 weight update없이 task에 대한 설명과 examples들을 부여받고 수행.

1S(One-Shot)은 few-shot가 모두 동일하되, one-example만 부여받고 수행.

0S(Zero-Shot)은 어떤 task인지에 대한 설명만 있음.

- Model and Architectures
  - GPT2와 유사한 구조
  - Transformer의 각 층마다 dense attention과 locally banded sparse attention을 번갈아 사용했음.
  - 다양한 크기의 training model을 사용했음.

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

- Training Dataset
  - 거의 1조단어에 달하는 Common Crawl dataset은 품질이 낮은 경향이 있기에 평균 품질을 향상시키기 위해 1)고품질 참조 corpus와 유사성을 기반으로 다운로드하고 필터링을 진행하였다. 2) 중복을 방지하고 유효성 검증 세트의 무결성을 유지했다.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

오염을 줄이기 위해 개발 및 테스트 세트와의 중복을 찾아 제거하려고 했음.

Training Process

더 큰 batch size 사용, 더 작은 학습률 필요, V100 GPU을 사용.

Evaluation

few-shot learning : 각 task의 trainset에서 K개의 예시를 무작위로 추출하여 testset의 각 예시를 평가함. LAMBADA와 Storycloze와 같이 supervised trainset이 없는 경우 development set에서 예시를 추출하고 testset에서 평가함.

Result

그냥 검증 과정들 포함. (생략)

https://supkoon.tistory.com/27

Limitations

GPT-3는 Auto-regressive 언어 모델입니다.

또한 GPT-3의 구조는 Transformer의 Decoder block으로 구성 되기 때문에, Masked self-Attention에 기반하고 있음.

Masked-self Attention은 엄밀히 따지면 bi-directional한 구조가 아닙니다. 또한 본 모델은 de-noising을 고려하는 목적함수를 설정하지 않았습니다.

이러한 구조, 알고리즘적인 문제가 일부 분야에서의 성능적 한계를 불러왔을 가능성이 존재
→만약 GPT-3와 비슷한 크기의 양방향(bidirectional) 모델이 있다면, GPT-3보다 뛰어난 fine-tuning 성능을 보였을 것으로 예상

Pre-training 과정에서 기존의 목적함수는 본질적인 한계가 존재

GPT-3는 Pre-training 동안 인간이 한평생 보는 것 보다 많은 양의 텍스트를 학습합니다. 테스트 과정에서는 Few-shot으로 샘플링 효율성이 인간에 가깝기는 하지만, 추가적인 정보나 알고리즘의 개선을 통해 사전훈련 과정의 샘플링 효율성을 개선할 필요가 있다.

Conclusion

GPT-3는 대규모의 데이터와 모델을 바탕으로 한 Auto-regressive Pre-trained language model.

GPT-3의 가장 큰 Contribution은 기존의 언어모델들과 다르게 Fine-tuning을 사용하지 않고도 Meta learning의 한가지 방법인 in-context learning을 통해 높은 Few-shot 성능을 보였다는 것.

심지어 일부의 Task에서는 Zero-shot, One-shot 환경에서도 기존의 SOTA 모델을 능가함.

구현

https://www.youtube.com/watch?v=kCc8FmEb1nY&t=22s