Improving Language Understanding by Generative Pre-Training

GPT - 1

Alec Radford ~ Karthik Narasimhan ~ Tim Salimans ~ Ilya Sutskever

OpenAI

Overview

Introduction

Introduction

- GPT-1 was launched in 2018
- by OpenAI

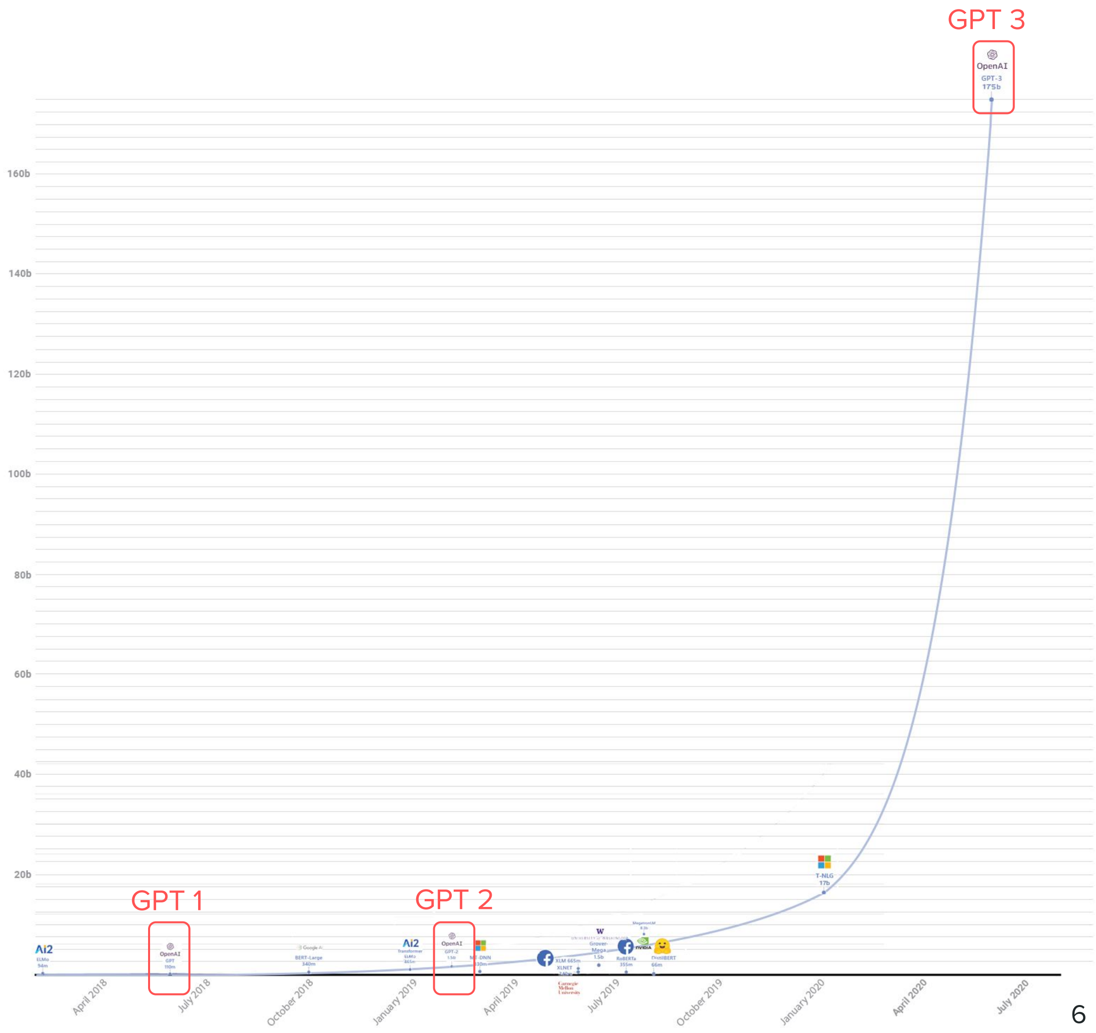
Introduction: Why ?

- NLP models were heavily trained on large amounts of annotated data for a particular task

- The NLP models were limited to what they have been trained for and failed to perform out-of-the-box tasks

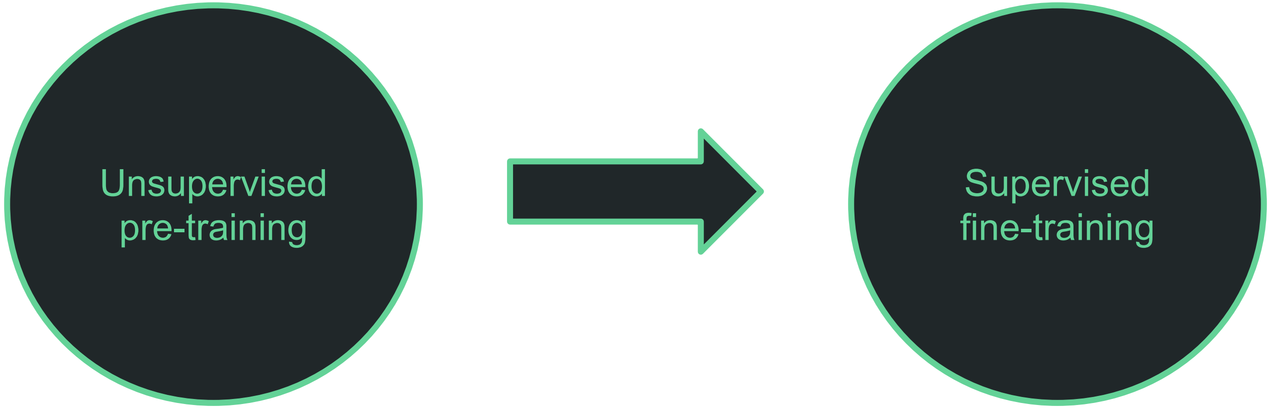
Introduction: Path of GPT

	GPT-1	GPT-2	GPT-3
Parameters	117 Million	1.5 Billion	175 Billion
Decoder Layers	12	48	96
Context Token Size	512	1024	2048
Hidden Layer	768	1600	12288
Batch Size	64	512	3.2M



Methodology

Methodology: Framework



Methodology: Unsupervised pre-training

1. For a unsupervised corpus of tokens **U = {u1, . . . , un}**
   a. Standard language modeling objective to maximize the this likelihood

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

   k is the size of the context window

   b. Conditional probability is **P** is modeled using a neural network with parameters **Θ**
   c. Parameters are trained using stochastic gradient descent [2]

2. Used multi-layer transformer decoder [3]
   a. Applies multi-headed self attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens

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

   U = (u−k, . . . , u−1) is the context vector of tokens,
   n is the number of layers, We is the token embedding matrix
   Wp is the position embedding matrix

[2] H. Robbins and S. Monro. A stochastic approximation method. The annals of mathematical statistics, pages 400–407, 1951
[3] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences. ICLR, 2018.

Methodology: Supervised fine-tuning

1. Assumes a labeled dataset **C** where each instance consists of input tokens, **x1, . . . , xm,** along with a label **y**
2. The inputs are passed through the pre-trained model to obtain the final transformer block's activation h_l^m , which is then fed into an added linear output layer with parameters **Wy** to predict **y**

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

3. Above gives the objective to maximize

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

4. Included language modeling as an auxiliary objective to fine-tuning helped learning by
    a. Improving generalization of the supervised model
    b. Accelerating convergence

    Optimized the following objective (with weight λ)

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

Methodology: Task specific input transformation

1. Some (Text classification) task can be directly fine tuned
2. But some are not. like,
    a. Question answering
    b. Textual entailment
    c. Have structured inputs such as ordered sentence pairs, Triplets of document, Question and answers

=> need modifications

Modifications

Use a traversal-style approach [4]

- 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.
- All transformations include adding randomly initialized start and end tokens (⟨s⟩, ⟨e⟩).

=> start and end tokens were added to input sequence
=> delimiter token was added between different parts of example so that input could be sent as ordered sequence.

E.g. a training example comprised of sequences for context, question and answer for question answering task.

[4] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013

Methodology: Input transformation

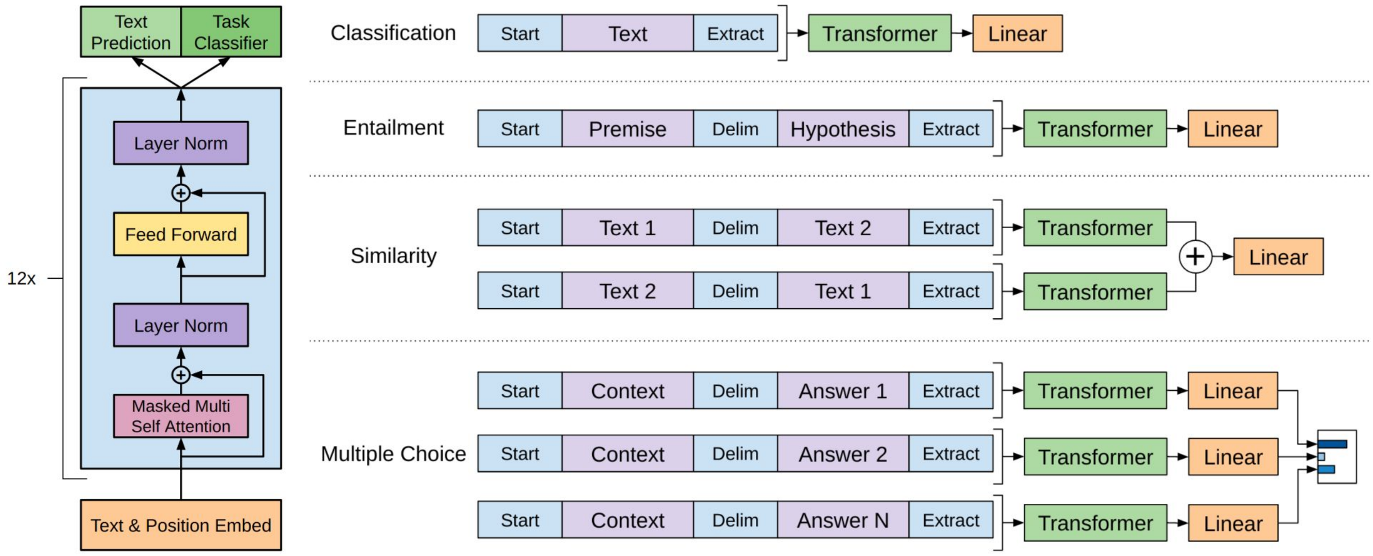


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Methodology: Dataset

1. Used BookCorpus dataset for training the language model.

   (Over 7000 unpublished books)

[5] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27, 2015.

Methodology: Model Specification > Unsupervised Training

1. Model largely follows the original transformer work [6]
2. Trained 12 layer decoder-only transformer with masked self-attention heads (768 dimensional states and 12 attention heads)
3. For the position-wise feed-forward networks, used 3072 dimensional inner states
4. Used Adam optimization scheme [7] with max learning rate of 2.5e-4
5. The learning rate was increased linearly from 0 over the first 2000 updates and annealed to 0 using a cosine schedule
6. Train for 100 epochs on minibatches of 64 randomly sampled, contiguous sequences of 512 tokens
7. Since layernorm [8] is used extensively throughout the model, a simple weight initialization of N (0, 0.02) was sufficient
8. Used a bytepair encoding (BPE) vocabulary with 40,000 merges and residual, embedding, and attention dropouts with a rate of 0.1 for regularization
9. Employed a modified version of L2 regularization proposed in [9], with w = 0.01 on all non bias or gain weights
10. Gaussian Error Linear Unit (GELU) [10] was used as the activation function
11. Used learned position embeddings instead of the sinusoidal version proposed in the original work
12. Used the ftfy library2 to clean the raw text in BooksCorpus, standardize some punctuation and whitespace
13. Used the spaCy tokenizer

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin.Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010, 2017.
[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[8] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.
[9] I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101,2017.
[10] D. Hendrycks and K. Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. arXiv preprint arXiv:1606.08415, 2016.

Methodology: Model Specification > Supervised Fine-tuning

1. 3 epochs for most of the downstream tasks

2. Most of the hyper parameters from unsupervised pre-training were used for fine-tuning.

Experiments

Experiments: Natural Language Inference Tasks

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

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

Experiments: Question Answering and Commonsense Reasoning Tasks

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

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**

Experiments: Semantic Similarity and Classification

Table 4: Semantic similarity and classification results, comparing our model with current state-of-the-art methods. All task evaluations in this table were done using the GLUE benchmark. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

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**

Analysis

Analysis

- GPT-1 proved that language model served as an effective pre-training objective which could help model generalize well.

- The architecture facilitated transfer learning and could perform various NLP tasks with very little fine-tuning.

- This model showed the power of generative pre-training and opened up avenues for other models which could unleash this potential better with larger datasets and more parameters.

References

References

[1] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).
[2] H. Robbins and S. Monro. A stochastic approximation method. The annals of mathematical statistics, pages 400–407, 1951
[3] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences. ICLR, 2018.
[4] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013
[5] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27, 2015.
[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin.Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010, 2017.
[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[8] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.
[9] I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101,2017.
[10] D. Hendrycks and K. Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. arXiv preprint arXiv:1606.08415, 2016.

Thank You