Generative Pre-Training in NLP & Its Generalization

Speaker：Cong MA

Date : 14th, April, 2019

中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

Outline

1. Introduction
2. The birth of GPT
3. Advanced version: GPT-2
4. Using pre-training to enhance NMT
5. Take-home-message

Outline

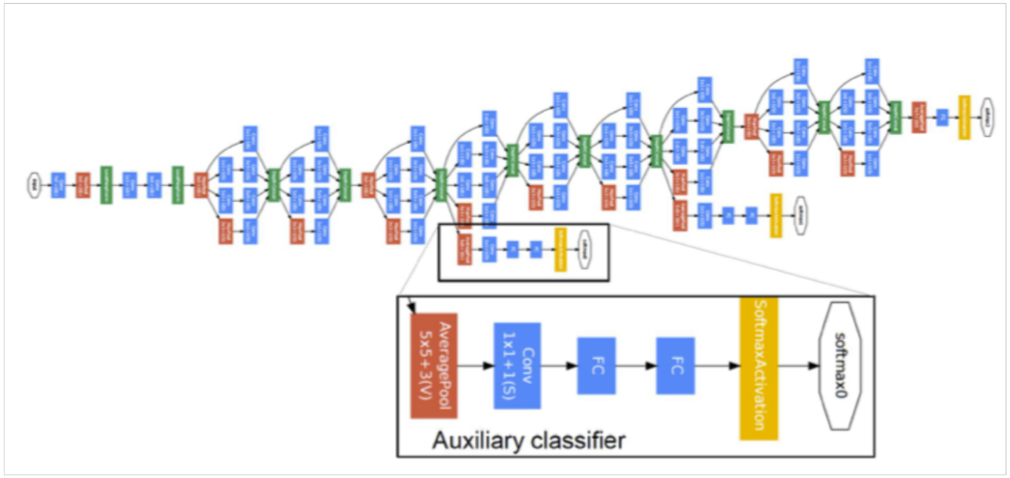
1. **Introduction**
2. The birth of GPT
3. Advanced version: GPT-2
4. Using pre-training to enhance NMT
5. Take-home-message

1.1 Pre-Training

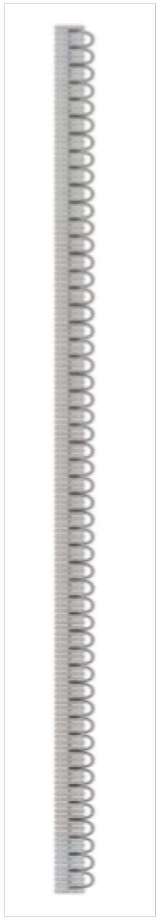
Pre-Training in Computer Vision Area

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

VGG



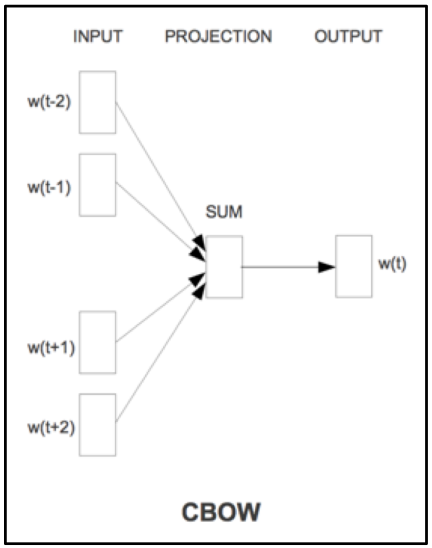
GoogLe Net

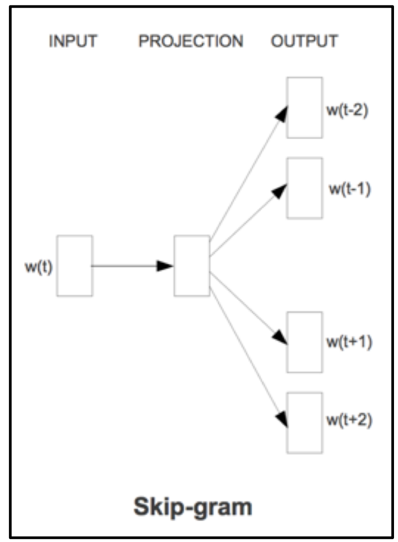


ResNet

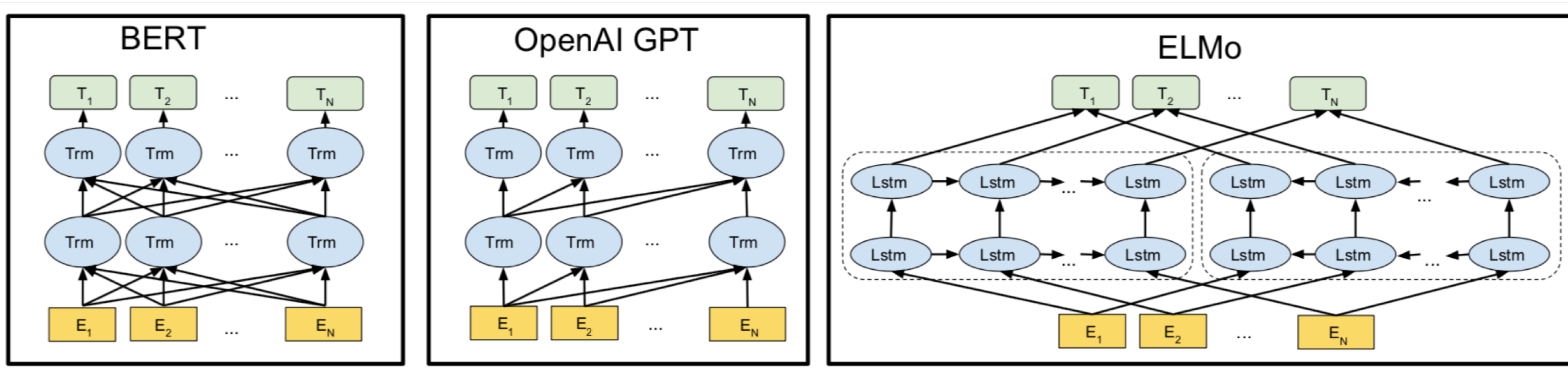
1.1 Pre-Training

Pre-Training in Natural Language Processing





GloVe
Tencent wrod2vec
sent2vec
doc2vec
…

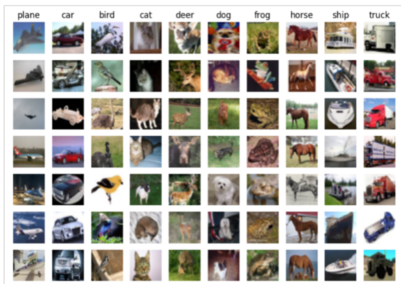


1.1 Pre-Training

Why do we need pre-training?

1.1 Pre-Training

Why do we need pre-training?



Images

Hello World.

Natural Language

Data Transfer

Feature Representation

Model

Classification

Regression

Generation

…

Raw Data: Source Space

Feature: Vector Space

Output Data: Target Space

1.2 Language Model

Language Model Task in Natural Language Processing

Given a sequence of N tokens, $(t_1, t_2, ..., t_N)$

Uni-directional Language Model:

Forward Language Model: $$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1}).$$

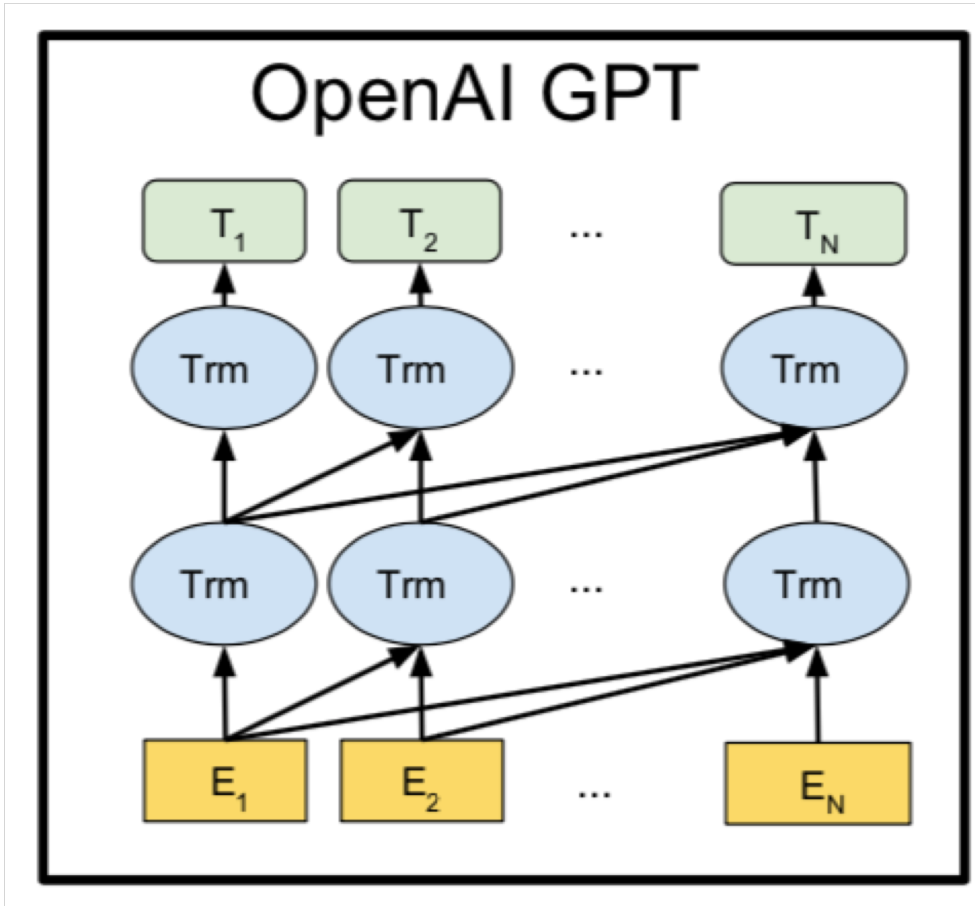
Backward Language Model: $$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N).$$

Bi-directional Language Model: $$\sum_{k=1}^{N} \left(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \right).$$

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer: Deep Contextualized Word Representations. NAACL-HLT 2018: 2227-2237

1.2 Language Model

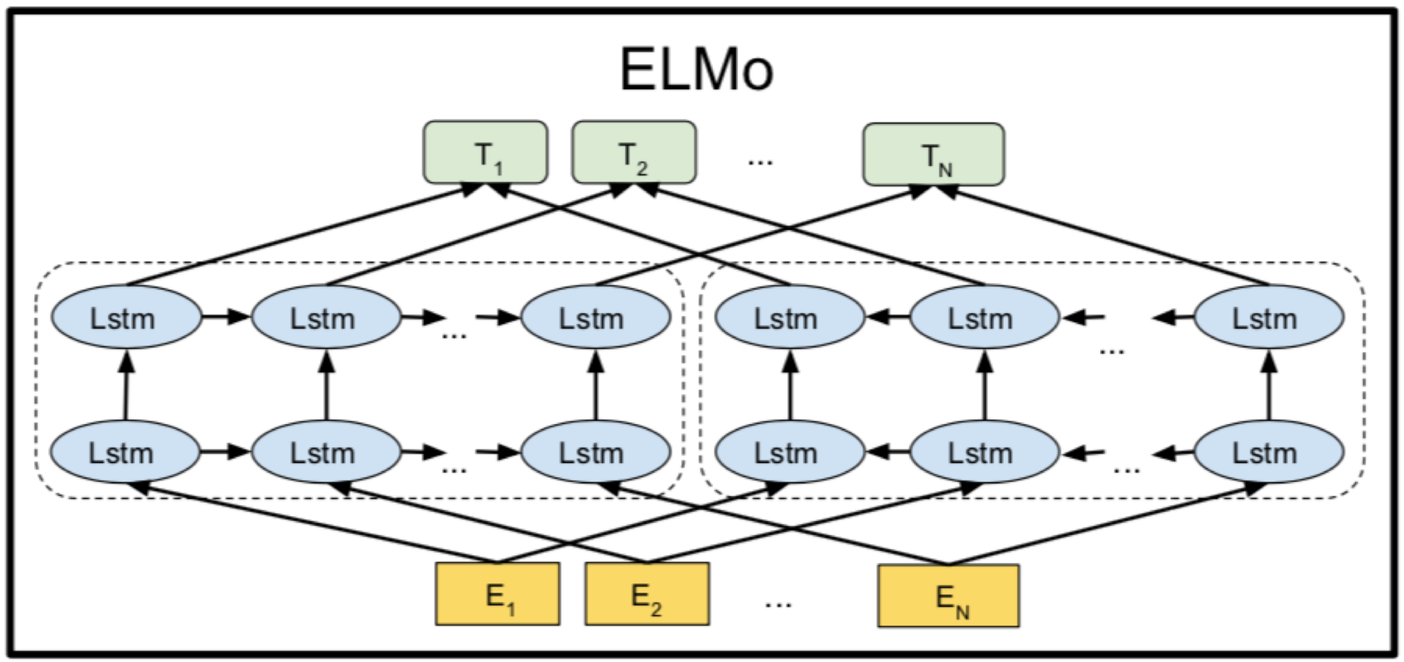
Pre-training Uni-directional Language Model in Natural Language Processing



Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

1.2 Language Model

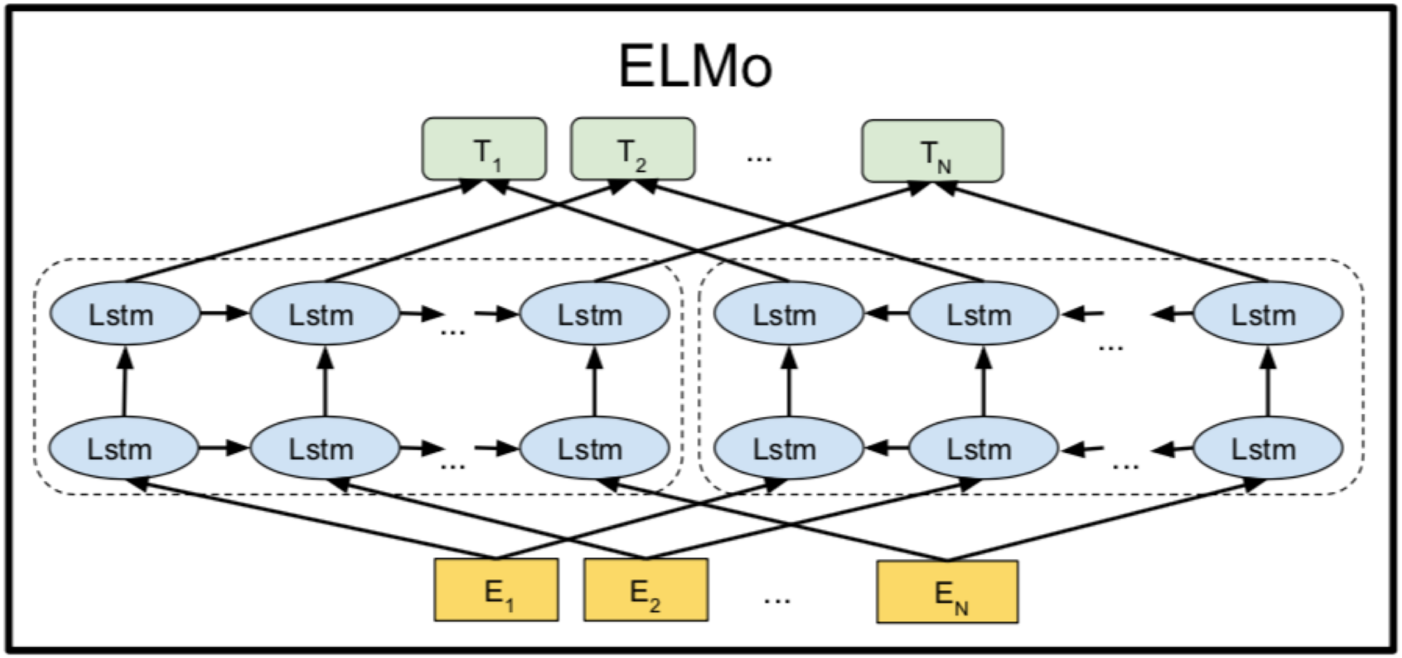
Pre-training Bi-directional Language Model in Natural Language Processing



Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer: Deep Contextualized Word Representations. NAACL-HLT 2018: 2227-2237

1.2 Language Model

Pre-training Bi-directional Language Model in Natural Language Processing



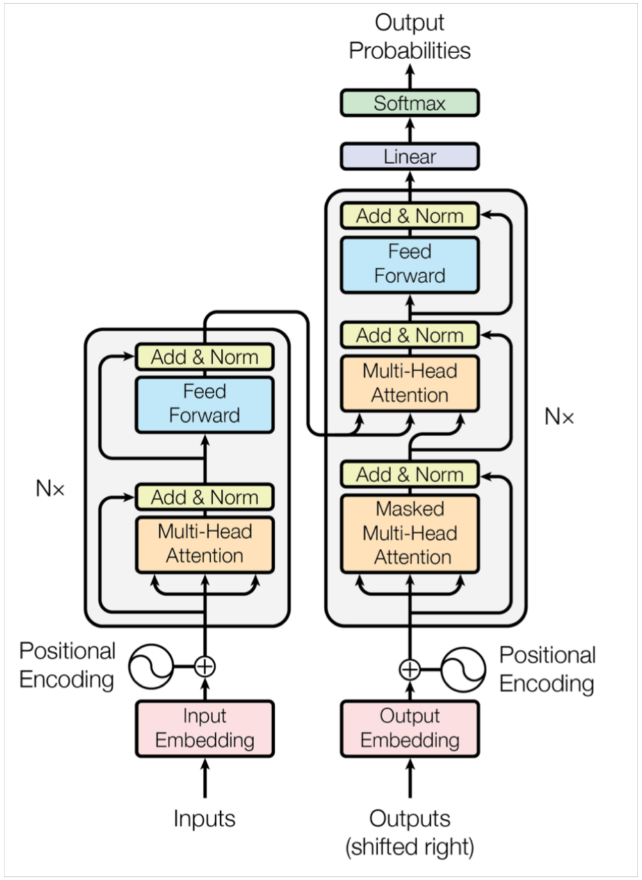
$$\begin{aligned} R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\ &= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\}, \end{aligned}$$

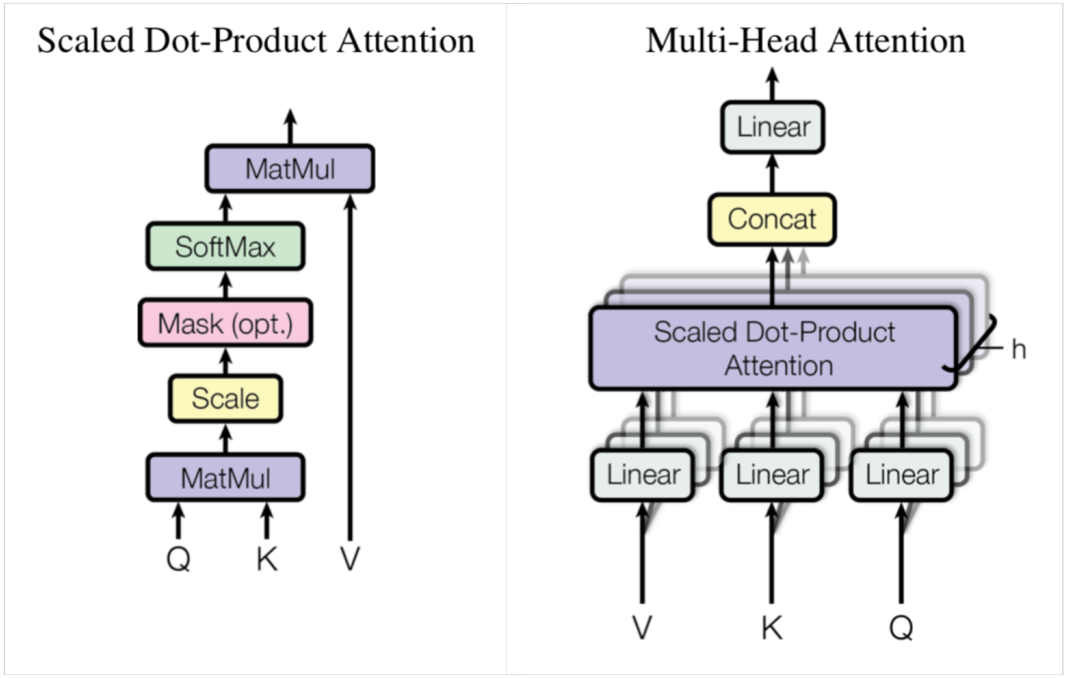
$$\mathrm{h}_{k,j}^{LM} = [\overrightarrow{\mathrm{h}}_{k,j}^{LM}; \overleftarrow{\mathrm{h}}_{k,j}^{LM}]$$

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer: Deep Contextualized Word Representations. NAACL-HLT 2018: 2227-2237

1.2 Language Model

Language Model + Transformer: BERT, GPT, GPT-2 ...





Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention is All you Need. NIPS 2017: 6000-6010

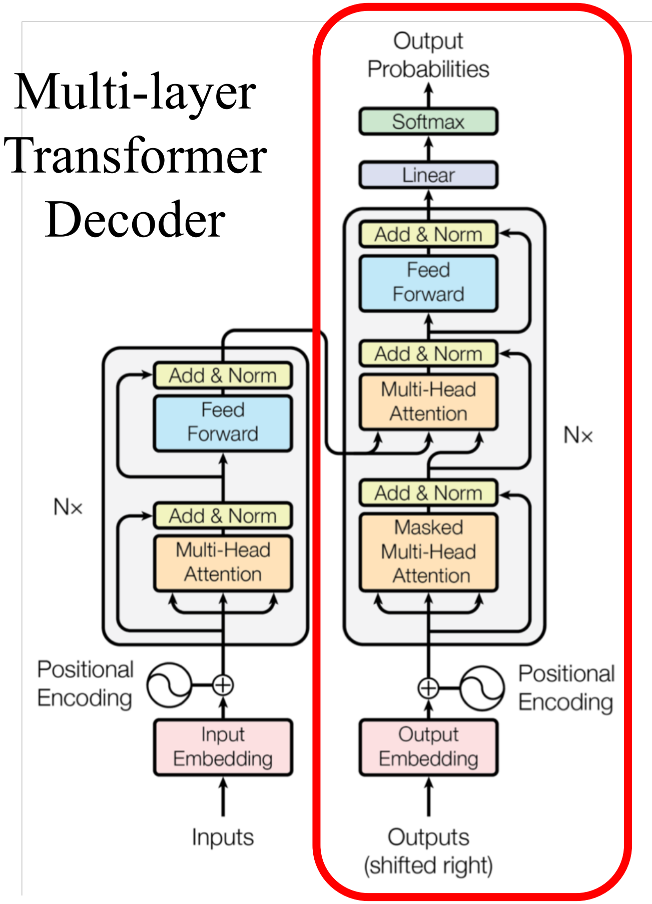
Outline

1. Introduction
2. **The birth of GPT**
3. Advanced version: GPT-2
4. Using pre-training to enhance NMT
5. Take-home-message

2. The Birth of GPT

Unsupervised pre-training.

Multi-layer Transformer Decoder



$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1})$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

$U = (u_{-k}, \ldots, u_{-1})$ is the context vector of tokens,
n is the number of layers,
W_e is the token embedding matrix,
W_p is the position embedding matrix.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Supervised fine-tuning.

Predict the y of the specific task:

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

Unsupervised Language Model Objective:

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

Task-specific Objective:

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

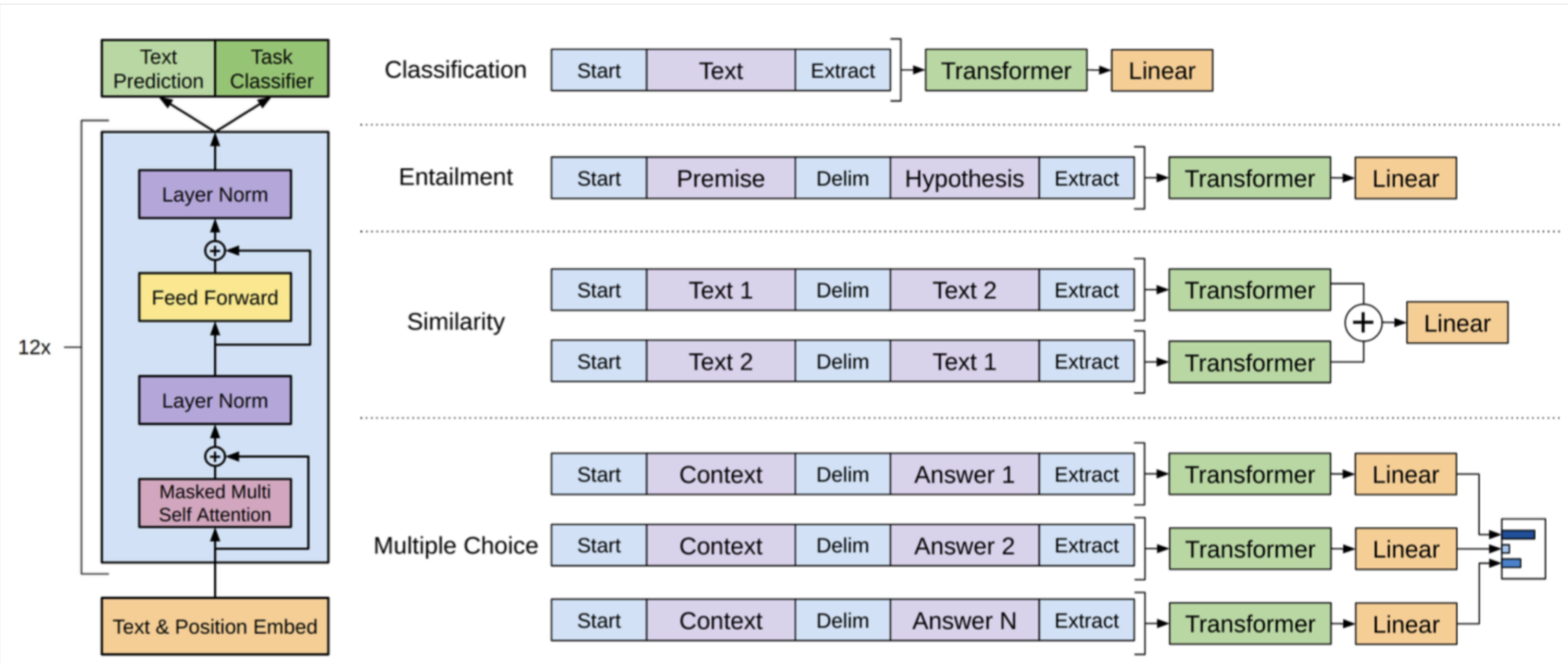
Auxiliary Objective:

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Model Architecture.



Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Examples of fine-tuning tasks in GPT Blog.

DATASET	EXAMPLE	LABEL
SNLI	1. A black race car starts up in front of a crowd of people. 2. A man is driving down a lonely road.	Contra.
MNLI	1. At the other end of Pennsylvania Avenue, people began to line up for a White House tour. 2. People formed a line at the end of Pennsylvania Avenue.	Entails
SciTail	1. Because type 1 diabetes is a relatively rare disease, you may wish to focus on prevention only if you know your child is at special risk for the disease. 2. Diabetes is unpreventable in the type one form but may be prevented by diet if it is of the second type.	Neutral
QNLI	Context: In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. Statement: What causes precipitation to fall?	Entails
RTE	1. Passions surrounding Germany's final match turned violent when a woman stabbed her partner because she didn't want to watch the game. 2. A woman passionately wanted to watch the game.	Contra.
STS-B	1. They flew out of the nest in groups. 2. They flew into the nest together.	Similarity 2/5
QQP	1. What are natural numbers 2. What is the least natural number	Not same
MRPC	1. If people took the pill daily, they would lower their risk of heart attack by 88 percent and of stroke by 80 percent, the scientists claim. 2. Taking the pill would lower the risk of heart attack by 88 percent and of stroke by 80 percent, the scientists said.	Same

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Examples of fine-tuning tasks in GPT Blog.

DATASET	EXAMPLE	LABEL
RACE	In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to. "Here's a letter for Miss Alice Brown," said the mailman. " I'm Alice Brown," a girl of about 18 said in a low voice. Alice looked at the envelope for a minute, and then handed it back to the mailman. "I'm sorry I can't take it, I don't have enough money to pay it", she said. A gentleman standing around were very sorry for her. Then he came up and paid the postage for her. When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it." "Really? How do you know that?" the gentleman said in surprise. "He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news." The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter. "The postage to be paid by the receiver has to be changed," he said to himself and had a good plan. "The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it. The girl handed the letter back to the mailman because: 1. she didn't know whose letter it was 2. she had no money to pay the postage 3. she received the letter but she didn't want to open it 4. she had already known what was written in the letter	4
ROCStories	Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating. 1. Karen became good friends with her roommate. 2. Karen hated her roommate.	1

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Examples of fine-tuning tasks in GPT Blog.

DATASET	EXAMPLE	LABEL
ROCStories	Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating. 1. Karen became good friends with her roommate. 2. Karen hated her roommate.	1
COPA	The man broke his toe. What was the CAUSE of this? 1. He got a hole in his sock. 2. He dropped a hammer on his foot.	2
SST-2	Just the labor involved in creating the layered richness of the imagery in this chiaroscuro of madness and light is astonishing.	Positive
CoLA	As you eat the most, you want the least.	Not acceptable

Analysis on CoLA:
https://nyu-mll.github.io/CoLA/

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Experimental Results.

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]

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

Accuracy

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Experimental Results.

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]

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**

Accuracy

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Experimental Results.

mc – Mathews Correlation
acc – Accuracy
pc – Pearson correlation

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]

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**

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

2. The Birth of GPT

Ablation Experimental Results.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

Outline

1. Introduction
2. The birth of GPT
3. **Advanced version: GPT-2**
4. Using pre-training to enhance NMT
5. Take-home-message

3. Advanced Version: GPT-2

Title: Language Models are Unsupervised Multitask Learners.

Improvements:

- Higher quality and larger quantity of data: WebText, around 8 million documents for a total of 40GB of text.
- More parameters in GPT-2: around 1542M parameters.
- Changes in Model Architecture:
  - Layer normalization was moved to the input of each sub-block and an additional layer normalization was added after the final self-attention block
  - A modified initialization on the residual part.
- Larger vocabulary: expand to 50257
- Larger context size and larger batch size.

GPT-2 zero-shots to state-of-the-art performance on 7 out of 8 tested language modeling dataset.

	Parameters	Layers	d_{model}
GPT-1	117M	12	768
BERT_Large	345M	24	1024
	762M	36	1280
GPT-2	1542M	48	1600

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

3. Advanced Version: GPT-2

Experimental Results on Language Modeling Task.

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

Table 6. Percentage of test set 8 grams overlapping with training sets.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019.
Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

3. Advanced Version: GPT-2

Experimental Results on multi NLP tasks with unsupervised method.

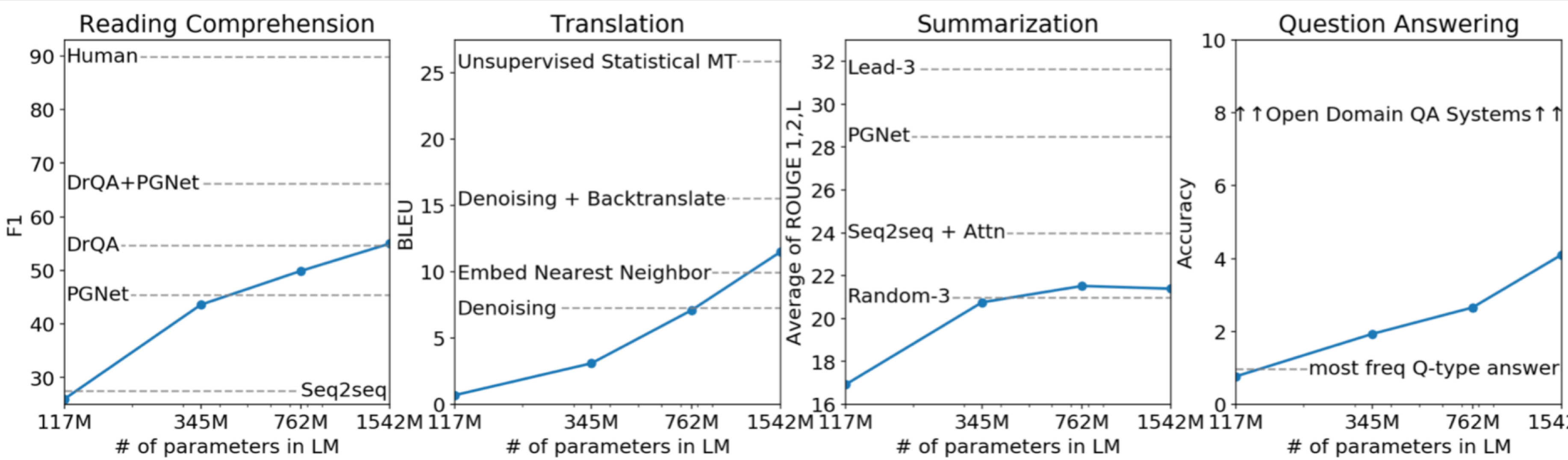


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

For Machine Translation: using [english sentence = french sentence]

For Summarization: using [TL;DR:](reprent Too Long; Don’t Read:)

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

3. Advanced Version: GPT-2

Experimental Results on unsupervised Summarization.

For Summarization: using [TL;DR:](reprent Too Long; Don't Read:)

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019.
Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

3. Advanced Version: GPT-2

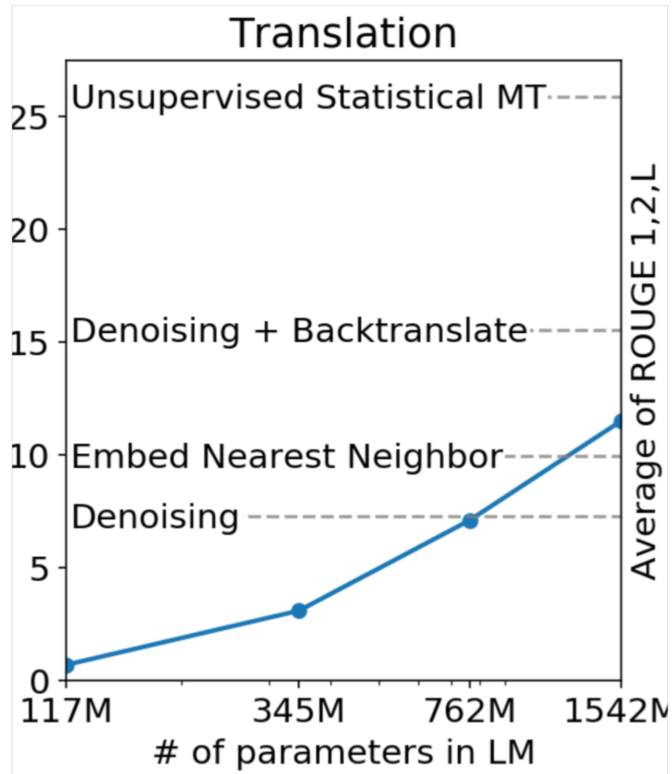
Experimental Results on multi NLP tasks with unsupervised method.

For Machine Translation: using [english sentence = french sentence]

For Machine Translation:
using [english sentence = french sentence]

A byte-level language detector is run on WebText, and just detected only 10MB (around 0.025% of WebText) of data in the French.

However, GPT-2 gets 5 BLEU on WMT-15 EnFr test set, which is slightly worse than a word-by-word substitution with a bilingual lexicon.



Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019.
Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

3. Advanced Version: GPT-2

Experimental Results on multi NLP tasks with unsupervised method.

Reading papers for more analysis on unsupervised test with GPT-2.

GPT-2 also show advantages of generating texts with system prompt[Human-Written], which is showed both on paper and blog.
https://www.openai.com/blog/better-language-models/#sample1

It is unclear whether the additional training data and capacity of GPT-2 is sufficient to overcome the inefficiencies of uni-directional representations demonstrated by Bert.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019.
Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Outline

1. Introduction
2. The birth of GPT
3. Advanced version: GPT-2
4. **Using pre-training to enhance NMT**
5. Take-home-message

4. Using pre-training to enhance NMT

Facebook AI Research[22th Mar.]: Replace learned input word embeddings in the encoder network with the output of the language model(SRC-FT). Specifically, they use the language model representation of the layer before the softmax and feed it to the encoder.

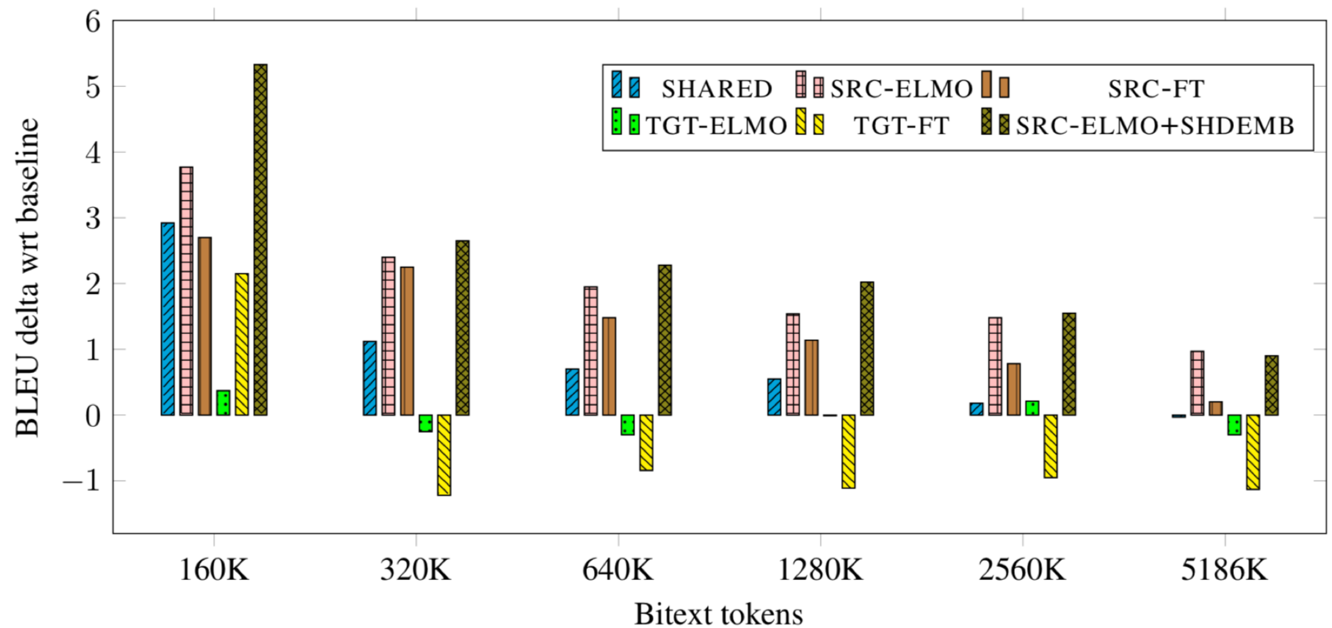


Figure 1: BLEU difference to a bitext-only baseline when adding pre-trained language model representations to a neural machine translation model in different simulated bitext settings. Results are based on averaging newstest2012-2017 of WMT English-German translation.

Sergey Edunov, Alexei Baevski, Michael Auli: Pre-trained Language Model Representations for Language Generation. CoRR abs/1903.09722 (2019).

4. Using pre-training to enhance NMT

	160K	640K	5186K
baseline	21.4	33.1	40.1
SRC-ELMO	26.6	35.6	41.8
SRC-FT	24.3	34.9	40.8
TGT-ELMO	21.3	31.9	40.5
TGT-FT	24.2	31.4	38.8
SRC-ELMO+SHDEMB	29.0	36.2	41.8

Table 1: BLEU on newstest2018 of WMT English-German in three simulated bitext size scenarios.

	news2017	news2018
baseline	9.8	9.5
SRC-ELMO	12.0	11.3
SRC-ELMO+SHDEMB	12.9	11.8

Table 2: WMT English-Turkish translation results in terms of BLEU on newstest2017 (valid) and newstest2018 (test) with ELMo inputs to the encoder.

	ROUGE 1	ROUGE 2	ROUGE L
Lead-3	40.34	17.70	36.57
See et al. (2017)	39.53	17.28	36.38
Gehrmann et al. (2018)	41.22	18.68	38.34
baseline	40.07	17.61	36.78
SRC-ELMO+SHDEMB	41.56	18.94	38.47

Table 3: Abstractive summarization results on CNN-DailyMail. ELMo inputs achieve a new state of the art.

Sergey Edunov, Alexei Baevski, Michael Auli: Pre-trained Language Model Representations for Language Generation. CoRR abs/1903.09722 (2019).

Outline

1. Introduction
2. The birth of GPT
3. Advanced version: GPT-2
4. Using pre-training to enhance NMT
5. **<u>Take-home-message</u>**

5. Take-Home-Message

1. Using Pre-trained representations to improve task-specific performance
2. GPT-2: Using Language Model as an Unsupervised Multitask Learner
3. Improve low-resource task performance with the help of Language Model
4. More data, More complex architecture, Better Performance

Thanks