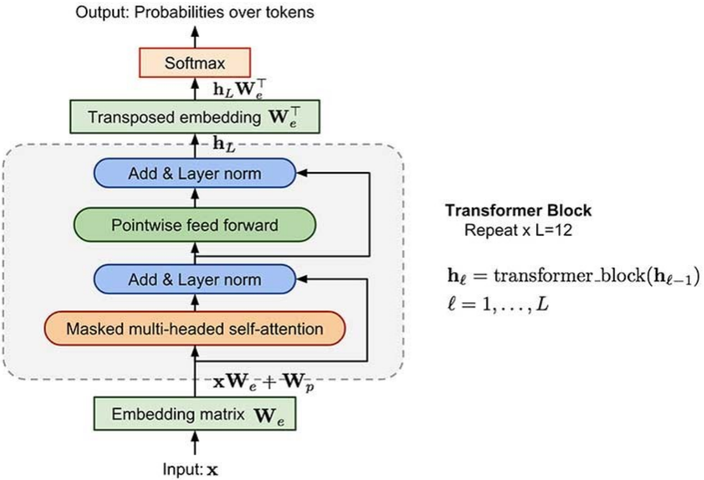
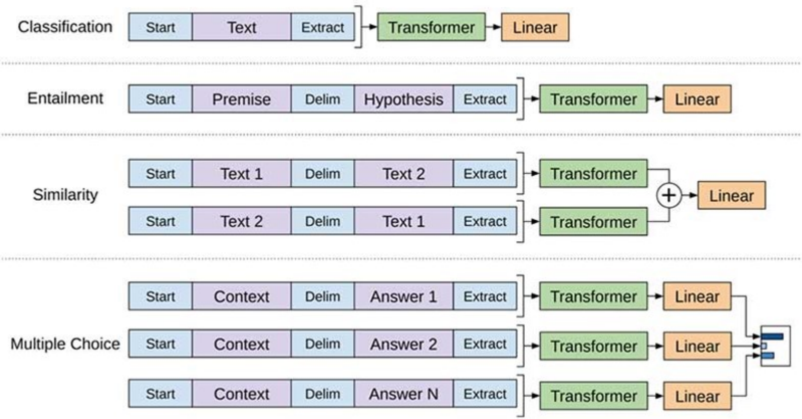
GPT: Transformer Decoder as Language Model


Output: Probabilities over tokens
Softmax
$\mathbf{h}_L \mathbf{W}_e^\top$
Transposed embedding $\mathbf{W}_e^\top$
$\mathbf{h}_L$
Add & Layer norm
Pointwise feed forward
Add & Layer norm
Masked multi-headed self-attention
$\mathbf{x}\mathbf{W}_e + \mathbf{W}_p$
Embedding matrix $\mathbf{W}_e$
Input: $\mathbf{x}$
Transformer Block
Repeat x L=12
$\mathbf{h}_\ell = \text{transformer\_block}(\mathbf{h}_{\ell-1})$
$\ell = 1, \ldots, L$

Liliang Wen, Generalized Language Models: Ulmfit & OpenAI GPT (blog)

GPT: supervised fine-tuning



Liliang Wen, Generalized Language Models: Ulmfit & OpenAI GPT (blog)

GPT-2: Introduction

- The OpenAI GPT-2 language model is a direct successor to GPT. GPT-2 has 1.5B parameters, 10x more than the original GPT, and it achieves SOTA results on 7 out of 8 tested language modeling datasets in a zero-shot transfer setting without any task-specific fine-tuning.
- The pre-training dataset contains 8 million Web pages collected by crawling qualified outbound links from Reddit. Large improvements by OpenAI GPT-2 are specially noticeable on small datasets and datasets used for measuring long-term dependency.

Liliang Wen, Generalized Language Models: Bert & OpenAI GPT-2 (blog)

GPT-2: zero-shot transfer

The pre-training task for GPT-2 is solely language modeling. All the downstream language tasks are framed as predicting conditional probabilities and there is no task-specific fine-tuning.

- **Text generation** is straightforward using LM.
- **Machine translation** task, for example, English to Chinese, is induced by conditioning LM on pairs of "English sentence = Chinese sentence" and the sentence to be translated "English sentence =" at the end.

  (to be continued...)

Liliang Wen, Generalized Language Models: Bert & OpenAI GPT-2 (blog)

GPT-2: zero-shot transfer

(...continued)
For example, the conditional probability to predict might look like:

$$p\left(?\,\middle|\,\begin{array}{r}\text{I like green apples.} = \text{我喜欢绿苹果。}\\ \text{A cat meows at him.} = \text{一只猫对他喵。}\\ \text{It is raining cats and dogs.} = \end{array}\right)$$

- **QA** task is formatted similar to translation with pairs of questions and answers in the context.
- **Summarization** task is induced by adding TL;DR: after the articles in the context.

Liliang Wen, Generalized Language Models: Bert & OpenAI GPT-2 (blog)

GPT-2: Model Modifications

Compared to GPT, other than having many more transformer layers and parameters, GPT-2 incorporates only a few architecture modifications:

- Layer normalization was moved to the input of each sub-block, similar to a residual unit of type “building block” (differently from the original type “bottleneck”, it has batch normalization applied before weight layers).
- An additional layer normalization was added after the final self-attention block.

Liliang Wen, Generalized Language Models: Bert & OpenAI GPT-2 (blog)

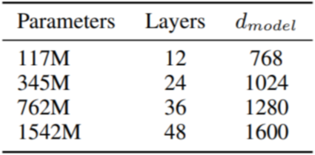
GPT-2: Model Modifications

Compared to GPT, other than having many more transformer layers and parameters, GPT-2 incorporates only a few architecture modifications:

- A modified initialization was constructed as a function of the model depth.
- The weights of residual layers were initially scaled by a factor of $\frac{1}{\sqrt{N}}$ where N is the number of residual layers.
- Use larger vocabulary size and context size.

Liliang Wen, Generalized Language Models: Bert & OpenAI GPT-2 (blog)

GPT-2: Models

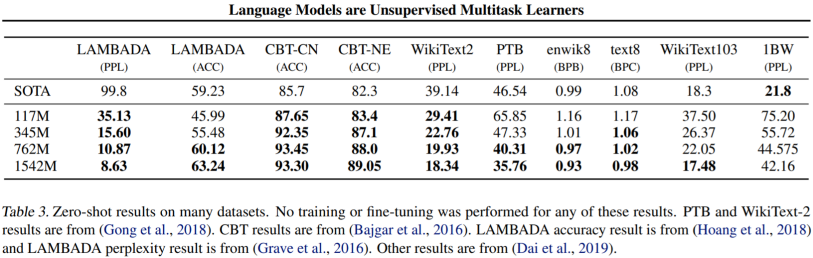


Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

Liliang Wen, Generalized Language Models: Bert & OpenAI GPT-2 (blog)

GPT-2: Results



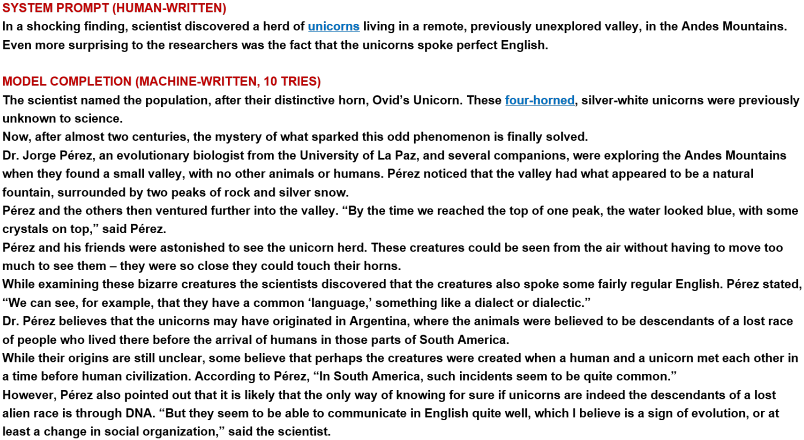
Language Models are Unsupervised Multitask Learners

	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).

Alec Radford, et al.,Language Models are Unsupervised Multitask Learners, OpenAI Blog 1.8 (2019): 9

GPT-2: an example of generated text



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

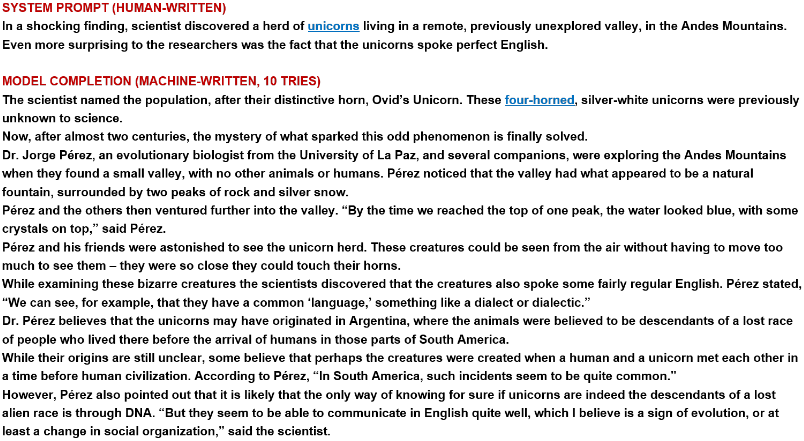
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

The generated text is suprisingly fluent and cohesive except for the common sense error (highlighted).

Alec Radford, et al.,Language Models are Unsupervised Multitask Learners, OpenAI Blog 1.8 (2019): 9

GPT-2: an example of generated text



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

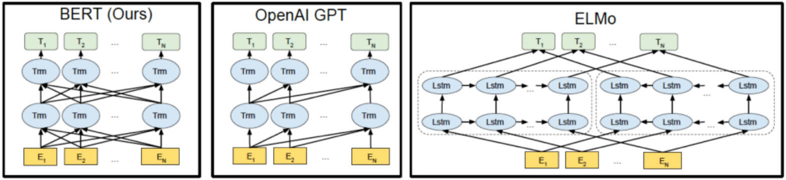
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

The generated text is suprisingly fluent and cohesive except for the common sense error (highlighted).

Alec Radford, et al.,Language Models are Unsupervised Multitask Learners, OpenAI Blog 1.8 (2019): 9

Differences: BERT, GPT & ELMo



Differences in pre-training model architectures: BERT, OpenAI GPT, and ELMo

Devlin et al., 2018, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Slides)

Content

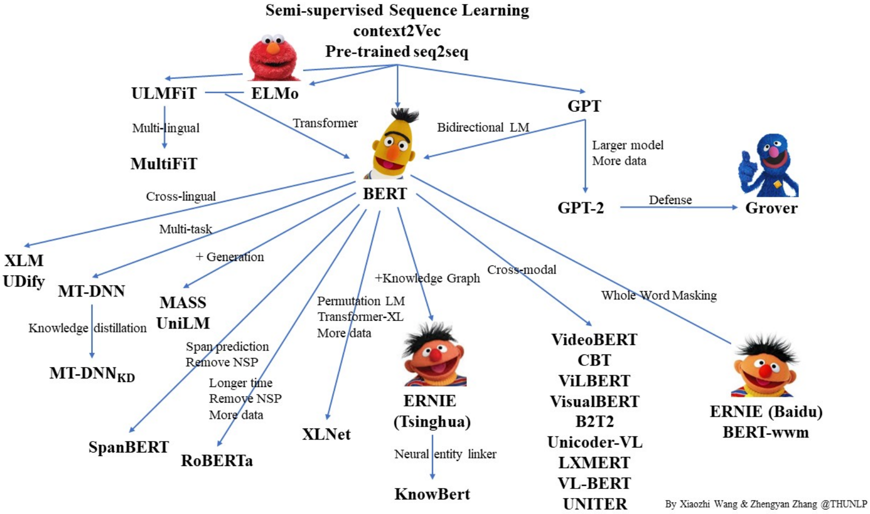
3 Pre-trained language models (PLMs)

- General introduction
- BERT
- GPT
- Recent progress and applications

The success of PLMs

- PLMs become another huge success in using deep learning in NLP after NMT:
  - PLMs have refreshed the state-of-the-art of most of the NLP tasks.
  - PLMs simplified the design of NN structures for downstream NLP tasks.
- Using PLMs has become a new paradigm for NLP research.

PLM recent progress: model family



https://github.com/thunlp/PLMpapers

Recent progress and applications

- Improved PLMs
- Larger PLMs
- Downsized PLMs
- Knowledge-aware PLMs
- Multilingual PLMs
- Multimodal PLMs
- Text Generation with PLMs
- Go beyond NLP

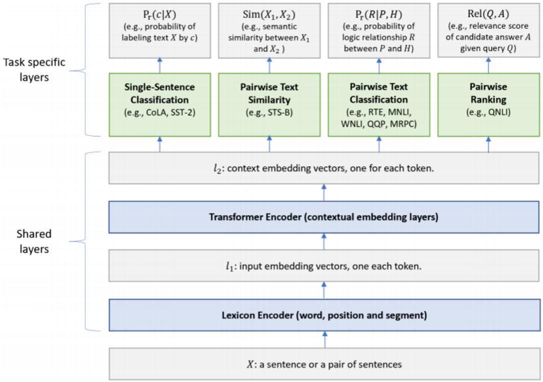
Multi-task Learning with BERT

- Pre-training
  - learning universal language representations by leveraging large amounts of unlabeled data
- Multi-task Learning
  - leveraging supervised data from many related tasks
  - regularization effect; universal representations across tasks
- Four types of tasks:
  - Single-Sentence Classification
  - Text Similarity
  - Pairwise Text Classification
  - Relevance Ranking

Liu et al., 2019, Multi-Task Deep Neural Networks for Natural Language Understanding

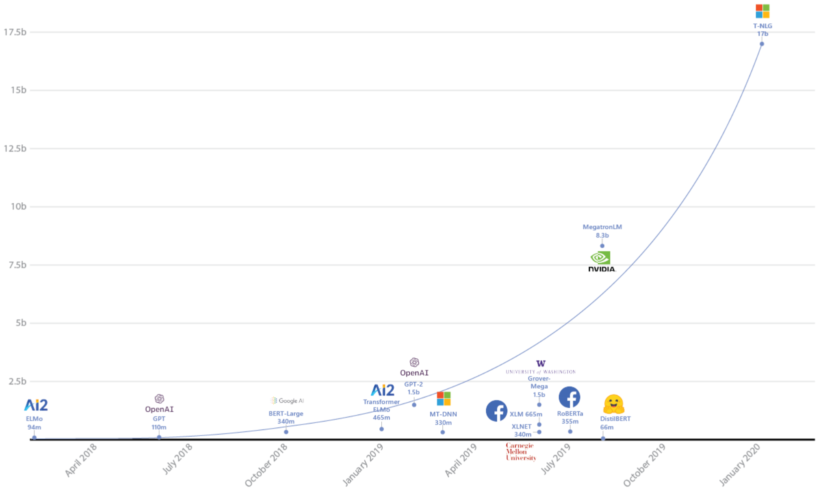
Multi-task Learning with BERT

- Multi-task learning happens in the fine-tuning phase.
- New SOTA on ten NLU tasks


Task specific layers
$P_r(c|X)$ (e.g., probability of labeling text X by c)
$\mathrm{Sim}(X_1, X_2)$ (e.g., semantic similarity between X_1 and X_2)
$P_r(R|P,H)$ (e.g., probability of logic relationship R between P and H)
$\mathrm{Rel}(Q,A)$ (e.g., relevance score of candidate answer A given query Q)
Single-Sentence Classification (e.g., CoLA, SST-2)
Pairwise Text Similarity (e.g., STS-B)
Pairwise Text Classification (e.g., RTE, MNLI, WNLI, QQP, MRPC)
Pairwise Ranking (e.g., QNLI)
Shared layers
l_2: context embedding vectors, one for each token.
Transformer Encoder (contextual embedding layers)
l_1: input embedding vectors, one each token.
Lexicon Encoder (word, position and segment)
X: a sentence or a pair of sentences

Liu et al., 2019, Multi-Task Deep Neural Networks for Natural Language Understanding

PLM recent progress: model size



Turing-NLG: A 17-billion-parameter language model by Microsoft, Microsoft Research Blog

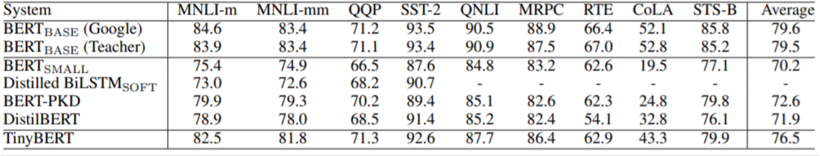
TinyBERT: Distilling BERT for NLU

- TinyBERT is proposed to execute on resource-restricted devices, for example, mobile phones.
- A novel transformer distillation method specially designed for knowledge distillation (KD) for transformer-based models.
- A new two-stage learning framework, which performs at both the pre-training and task-specific learning stages.
- A novel data augmentation technique is used in the second stage distillation.

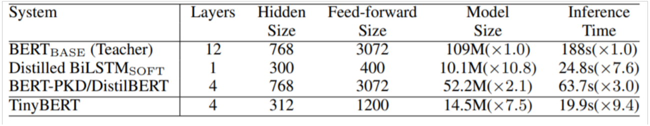
Jiao et al., Distilling BERT for Natural Language Understanding, https://arxiv.org/abs/1909.10351, 2019

TinyBERT: Distilling BERT for NLU

- TinyBERT is empirically effective and achieves more than 96% the performance of teacher BERT-base on GLUE benchmark while being 7.5x smaller and 9.4x faster on inference.



System	MNLI-m	MNLI-mm	QQP	SST-2	QNLI	MRPC	RTE	CoLA	STS-B	Average
BERT$_{\text{BASE}}$ (Google)	84.6	83.4	71.2	93.5	90.5	88.9	66.4	52.1	85.8	79.6
BERT$_{\text{BASE}}$ (Teacher)	83.9	83.4	71.1	93.4	90.9	87.5	67.0	52.8	85.2	79.5
BERT$_{\text{SMALL}}$	75.4	74.9	66.5	87.6	84.8	83.2	62.6	19.5	77.1	70.2
Distilled BiLSTM$_{\text{SOFT}}$	73.0	72.6	68.2	90.7	-	-	-	-	-	-
BERT-PKD	79.9	79.3	70.2	89.4	85.1	82.6	62.3	24.8	79.8	72.6
DistilBERT	78.9	78.0	68.5	91.4	85.2	82.4	54.1	32.8	76.1	71.9
TinyBERT	82.5	81.8	71.3	92.6	87.7	86.4	62.9	43.3	79.9	76.5

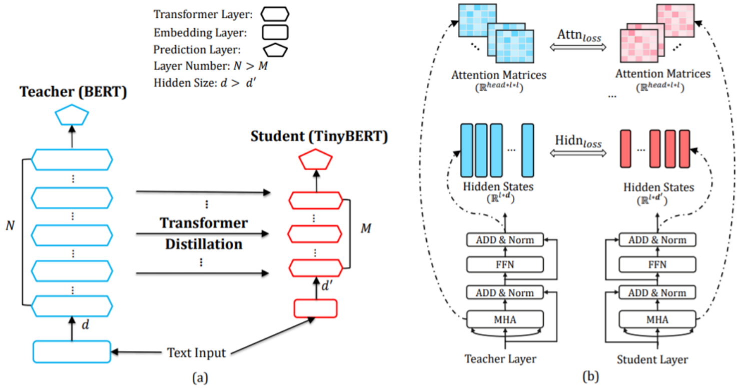


System	Layers	Hidden	Feed-forward	Model	Inference
		Size	Size	Size	Time
BERT$_{\text{BASE}}$ (Teacher)	12	768	3072	109M(×1.0)	188s(×1.0)
Distilled BiLSTM$_{\text{SOFT}}$	1	300	400	10.1M(×10.8)	24.8s(×7.6)
BERT-PKD/DistilBERT	4	768	3072	52.2M(×2.1)	63.7s(×3.0)
TinyBERT	4	312	1200	14.5M(×7.5)	19.9s(×9.4)

Jiao et al., Distilling BERT for Natural Language Understanding, https://arxiv.org/abs/1909.10351, 2019

TinyBERT: Distilling BERT for NLU

- Transformer distillation:



Jiao et al., Distilling BERT for Natural Language Understanding, https://arxiv.org/abs/1909.10351, 2019

TinyBERT: Distilling BERT for NLU

- Two-stage learning framework:

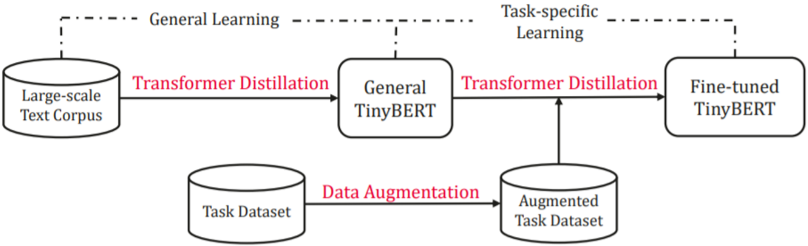
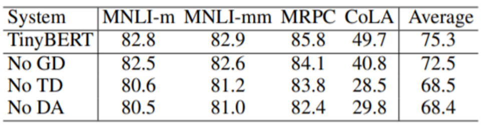


Figure 2: The illustration of TinyBERT learning

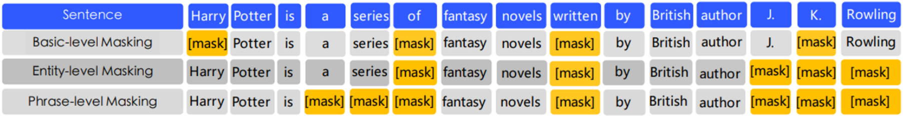


System	MNLI-m	MNLI-mm	MRPC	CoLA	Average
TinyBERT	82.8	82.9	85.8	49.7	75.3
No GD	82.5	82.6	84.1	40.8	72.5
No TD	80.6	81.2	83.8	28.5	68.5
No DA	80.5	81.0	82.4	29.8	68.4

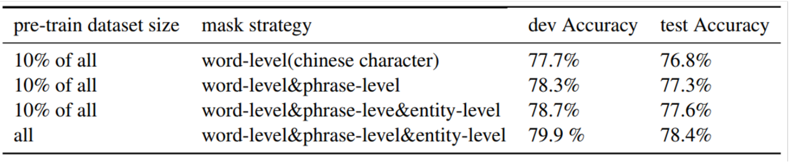
Jiao et al., Distilling BERT for Natural Language Understanding, https://arxiv.org/abs/1909.10351, 2019

ERNIE-Baidu

- ERNIE-Baidu
  - Entity-level masking
  - Phrase-level masking
- Outperform BERT on 5 Chinese language processing



Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

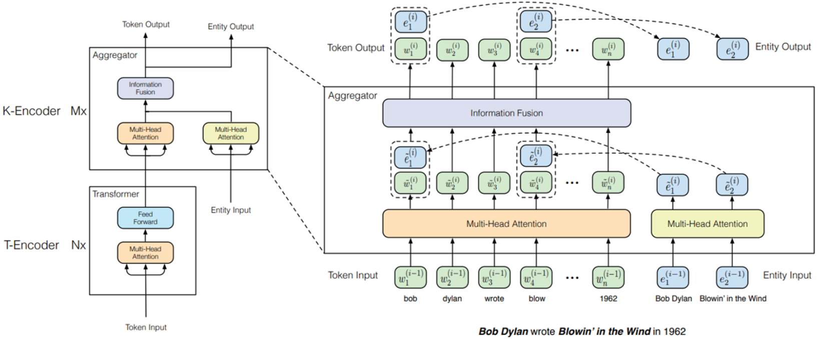


pre-train dataset size	mask strategy	dev Accuracy	test Accuracy
10% of all	word-level(chinese character)	77.7%	76.8%
10% of all	word-level&phrase-level	78.3%	77.3%
10% of all	word-level&phrase-leve&entity-level	78.7%	77.6%
all	word-level&phrase-level&entity-level	79.9 %	78.4%

Sun et al., 2019, ERNIE: Enhanced Representation through Knowledge Integration

ERNIE-Tsinghua

- ERNIE-Tsinghua
  - Entities from knowledge-graph
  - The information fusion adopt a two-layer feed-forward network


Token Output
Entity Output
Aggregator
Information Fusion
K-Encoder Mx
Multi-Head Attention
Multi-Head Attention
Transformer
Feed Forward
Entity Input
T-Encoder Nx
Multi-Head Attention
Token Input
$e_1^{(i)}$
$e_2^{(i)}$
Token Output
$w_1^{(i)}$ $w_2^{(i)}$ $w_3^{(i)}$ $w_4^{(i)}$ ... $w_n^{(i)}$
$e_1^{(i)}$ $e_2^{(i)}$ Entity Output
Aggregator
Information Fusion
$\tilde{e}_1^{(i)}$
$\tilde{e}_2^{(i)}$
$\tilde{w}_1^{(i)}$ $\tilde{w}_2^{(i)}$ $\tilde{w}_3^{(i)}$ $\tilde{w}_4^{(i)}$ ... $\tilde{w}_n^{(i)}$
$\tilde{e}_1^{(i)}$ $\tilde{e}_2^{(i)}$
Multi-Head Attention
Multi-Head Attention
Token Input
$w_1^{(i-1)}$ $w_2^{(i-1)}$ $w_3^{(i-1)}$ $w_4^{(i-1)}$ ... $w_n^{(i-1)}$
$e_1^{(i-1)}$ $e_2^{(i-1)}$ Entity Input
bob dylan wrote blow 1962
Bob Dylan Blowin' in the Wind
Bob Dylan wrote ***Blowin' in the Wind*** in 1962

(a) Model Achitecture

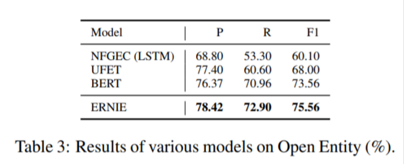
(b) Aggregator

Zhang et al., 2019, ERNIE: Enhanced Language Representation with Informative Entities

ERNIE-Tsinghua

- ERNIE-Tsinghua
  - Entity embeddings are pre-trained with TransE
  - Knowledge encoder + Entity prediction task
- Results
  - Comparable performance on normal NLP tasks
  - Better performance on knowledge-driven tasks

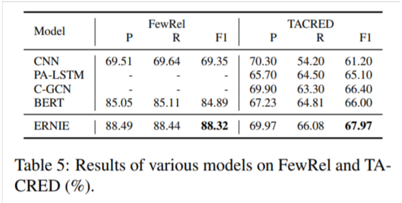
Entity Typing tasks:



Model	P	R	F1
NFGEC (LSTM)	68.80	53.30	60.10
UFET	77.40	60.60	68.00
BERT	76.37	70.96	73.56
ERNIE	**78.42**	**72.90**	**75.56**

Table 3: Results of various models on Open Entity (%).

Relation Classification tasks:



Model	FewRel			TACRED		
	P	R	F1	P	R	F1
CNN	69.51	69.64	69.35	70.30	54.20	61.20
PA-LSTM	-	-	-	65.70	64.50	65.10
C-GCN	-	-	-	69.90	63.30	66.40
BERT	85.05	85.11	84.89	67.23	64.81	66.00
ERNIE	88.49	88.44	**88.32**	69.97	66.08	**67.97**

Table 5: Results of various models on FewRel and TACRED (%).

Zhang et al., 2019, ERNIE: Enhanced Language Representation with Informative Entities

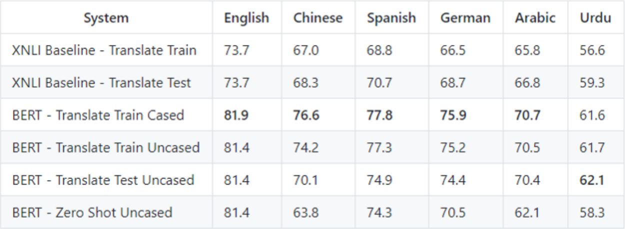
Multilingual BERT

- It is impractical to have individual models for each language
  - Learning a universal model for all languages
- Multilingual BERT
  - The languages chosen were the top 100 languages with the largest Wikipedias.
  - The entire Wikipedia dump for each language (excluding user and talk pages) was taken as the training data for each language
  - Balance the data: high-resource languages like English will be under-sampled, and low-resource languages like Icelandic will be over-sampled.

`https://github.com/google-research/bert/blob/master/multilingual.md`

Multilingual BERT

- Multilingual BERT (continued)
  - For tokenization, a 110k shared WordPiece vocabulary was used.
  - The word counts are weighted the same way as the data, so low-resource languages are upweighted by some factor.
  - We intentionally do not use any marker to denote the input language (so that zero-shot training can work).



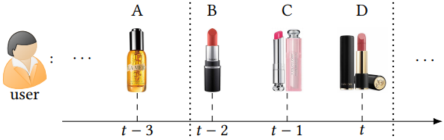
System	English	Chinese	Spanish	German	Arabic	Urdu
XNLI Baseline - Translate Train	73.7	67.0	68.8	66.5	65.8	56.6
XNLI Baseline - Translate Test	73.7	68.3	70.7	68.7	66.8	59.3
BERT - Translate Train Cased	**81.9**	**76.6**	**77.8**	**75.9**	**70.7**	61.6
BERT - Translate Train Uncased	81.4	74.2	77.3	75.2	70.5	61.7
BERT - Translate Test Uncased	81.4	70.1	74.9	74.4	70.4	**62.1**
BERT - Zero Shot Uncased	81.4	63.8	74.3	70.5	62.1	58.3

However, for high-resource languages, the multilingual model is somewhat worse than a single-language model.

`https://github.com/google-research/bert/blob/master/multilingual.md`

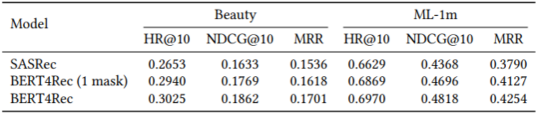
BERT for Recommendation: BERT4Rec

- Drawbacks of sequence modeling
  - Sequential dependencies over long time scales (e.g., from A to [B,C,D]) vs random actions in a short period (e.g., [B,C,D])



- BERT: jointly conditioning on both left and right context

Table 3: Analysis on bidirection and Cloze with $d = 256$



Model	Beauty			ML-1m		
	HR@10	NDCG@10	MRR	HR@10	NDCG@10	MRR
SASRec	0.2653	0.1633	0.1536	0.6629	0.4368	0.3790
BERT4Rec (1 mask)	0.2940	0.1769	0.1618	0.6869	0.4696	0.4127
BERT4Rec	0.3025	0.1862	0.1701	0.6970	0.4818	0.4254

Sun et al., 2019, BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer

Content

1. Subword level and character level NMT
2. Transformer-based NMT
3. Pre-trained language models (PLMs)