Boulder

Natural Language Processing; Language Modeling

YouTube Playlist

Maziar Raissi

Assistant Professor

Department of Applied Mathematics

University of Colorado Boulder

maziar.raissi@colorado.edu

Deep contextualized word representations

YouTube Video

ELMo: Embeddings from Language Models

$(t_1, t_2, \ldots, t_N) \rightarrow$ sequence of N tokens

Forward Language Model

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

$x_k^{LM} \rightarrow$ context-independent token representation

$x_k^{LM} \triangleright$ L layers of forward LSTMs

$\overrightarrow{h}_{k,j}^{LM}, j = 1, \ldots, L$

$\overrightarrow{h}_{k,L}^{LM} \triangleright$ softmax $\rightarrow$ predict next token t_{k+1}

Backward LM

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

$\overleftarrow{h}_{k,j}^{LM}, j = 1, \ldots, L$

Bidirectional LM

token softmax

$$\sum_{k=1}^{N} (\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s))$$

ELMo

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

$$h_{k,0}^{Lm} = x_k^{LM}$$

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

$$\text{ELMo}_k = E(R_k; \Theta_e)$$

$$E(R_k) = h_{k,L}^{LM}$$

$$\text{ELMo}_k^{\text{task}} = E(R_k; \Theta^{\text{task}}) = \gamma^{\text{task}} \sum_{j=0}^{L} s_j^{\text{task}} h_{k,j}^{LM}$$

$s_j^{\text{task}} \rightarrow$ softmax-normalized weights

$[x_k; \text{ELMo}_k^{\text{task}}] \rightarrow$ input to task RNN

$[h_k; \text{ELMo}_k^{\text{task}}] \rightarrow$ output of task RNN

Downstream Tasks

1. Question Answering
2. Textual Entailment
   Determine whether a "hypothesis" is true given a "premise"
3. Semantic Role Labeling
   who, did what, to whom, at where
4. Coreference Resolution
   clustering mentions in the text that refer to the same underlying real world entities
5. Named Entity Extraction
   (e.g., entity types PER, LOC, ORG, MISC
6. Sentiment Analysis

Task	Previous SOTA			Our baseline	ELMo + baseline	Increase (absolute/ relative)
Stanford Question Answering Dataset ← SQuAD	Liu et al. (2017)	F_1	84.4	81.1	85.8	4.7 / 24.9%
Stanford Natural Language Inference ← SNLI	Chen et al. (2017)	accuracy	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
Semantic Role Labeling ← SRL	He et al. (2017)	F_1	81.7	81.4	84.6	3.2 / 17.2%
Coreference Resolution ← Coref	Lee et al. (2017)	average F_1	67.2	67.2	70.4	3.2 / 9.8%
Named Entity Recognition ← NER	Peters et al. (2017)	F_1	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
Stanford Sentiment Treebank ← SST-5	McCann et al. (2017)	accuracy	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {…}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {…}	{…} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Peters, Matthew E., et al. "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365* (2018).

Improving Language Understanding by Generative Pre-Training

YouTube Video

Unsupervised pre-training

$\mathcal{U} = \{u_1, u_2, \ldots, u_n\} \rightarrow$ unsupervised corpus of tokens

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

$k \rightarrow$ size of the context window

Transformer decoder

$U = (u_{-k}, \ldots, u_{-1}) \rightarrow$ context vector of tokens

$$h_0 = UW_e + W_p$$

$W_p \rightarrow$ position embedding matrix

$W_e \rightarrow$ token embedding matrix

$$h_l = \texttt{transformer\_block}(h_{l-1}), l = 1, \ldots, L$$

$L \rightarrow$ number of layers

$$P(u) = \texttt{softmax}(h_L W_e^T)$$

Supervised fine-tuning

$\mathcal{C} \rightarrow$ labeled dataset

$x^1, \ldots, x^m \rightarrow$ input tokens

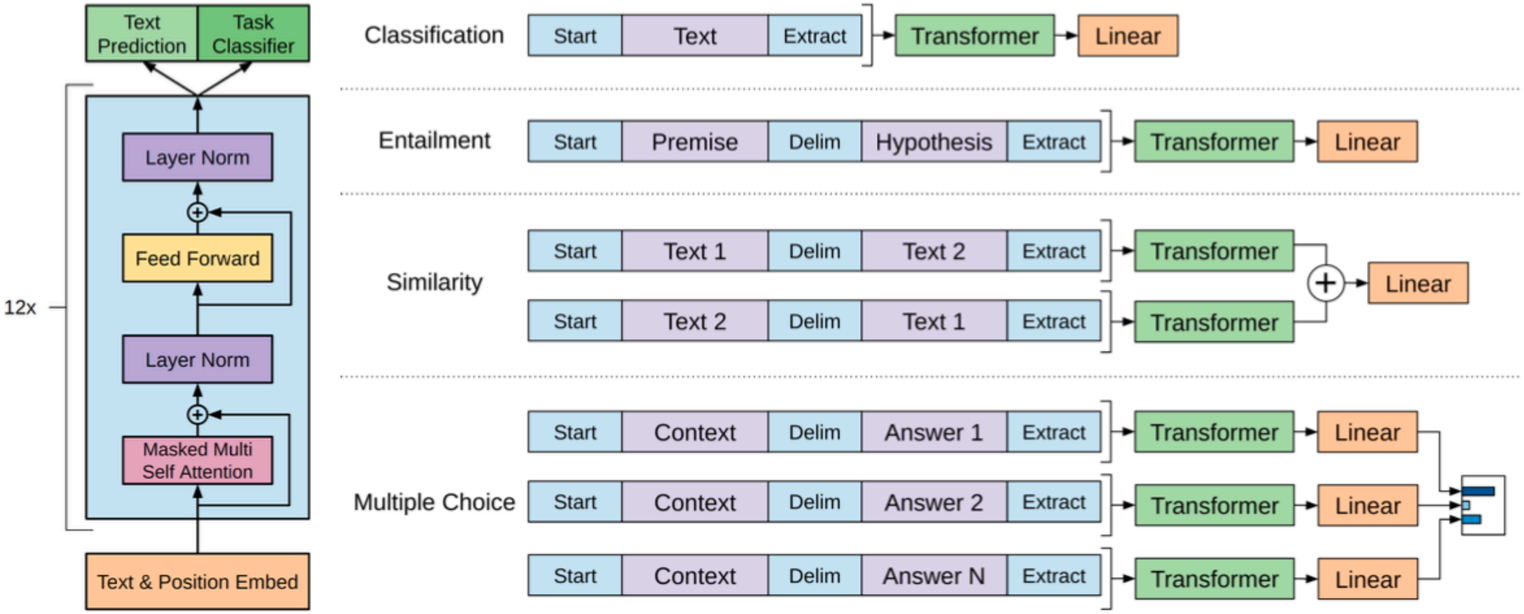
$y \rightarrow$ label

$h_L^m \rightarrow$ final transformer block's activation

$$p(y | x^1, \ldots, x^m) = \text{softmax}(h_L^m W_y)$$

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

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



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]

Natural Language Inference `entailment`, `contradiction` or `neutral`

image captions (SNLI)

transcribed speech, popular fiction, and government reports (MNLI)

Wikipedia articles (QNLI)

science exams (SciTail)

news articles (RTE)

Corpus of Linguistic Acceptability (CoLA)

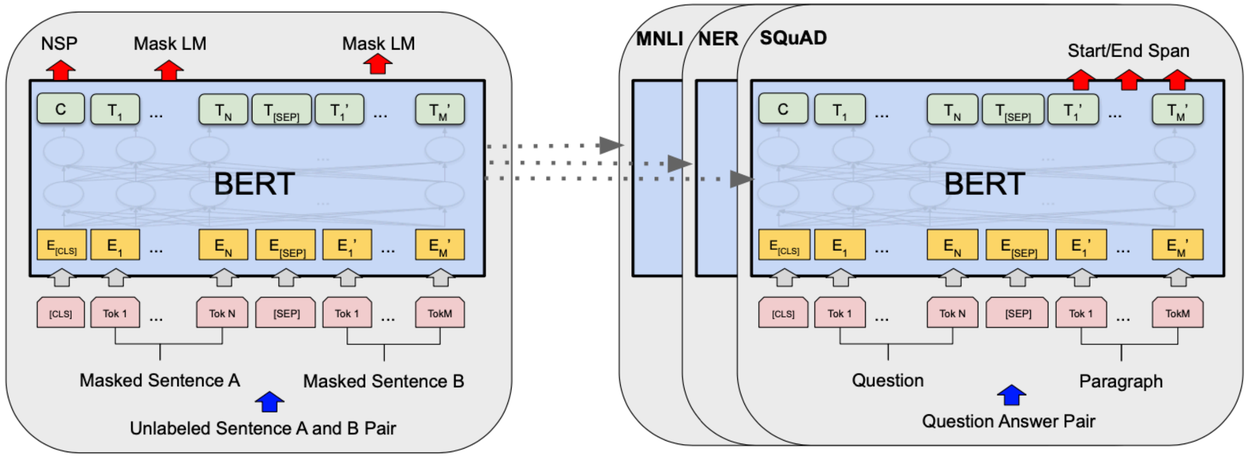
Semantic Textual Similarity (STS-B)

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018): 12.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

YouTube Playlist

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

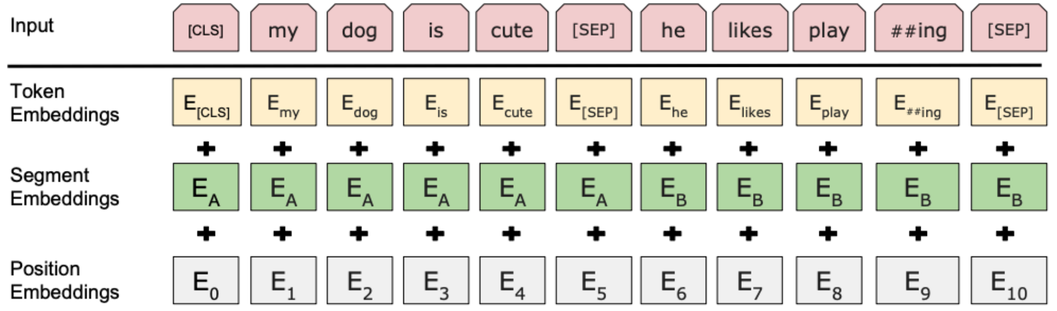


Task 1: Masked LM

mask 15% of all WordPiece tokens in each sequence at random

Task 2: Next Sentence Prediction (NSP)

IsNext vs NotNext



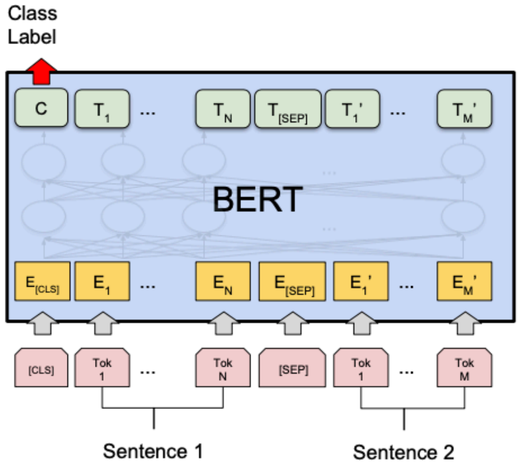
$S \in \mathbb{R}^H \rightarrow$ start

$E \in \mathbb{R}^H \rightarrow$ end

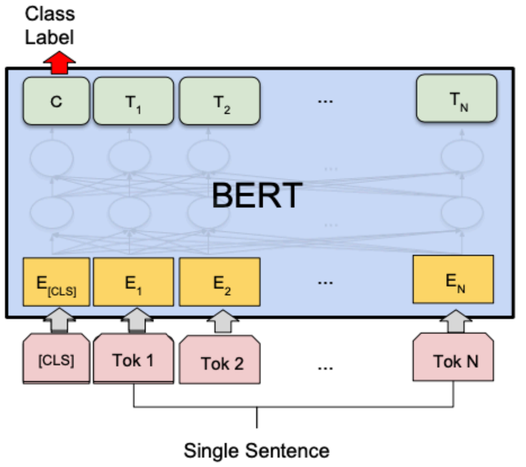
$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$$

$S \cdot T_i + E \cdot T_j$

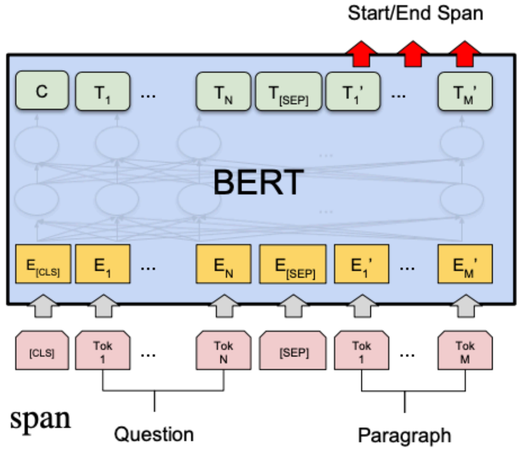
maximum scoring span



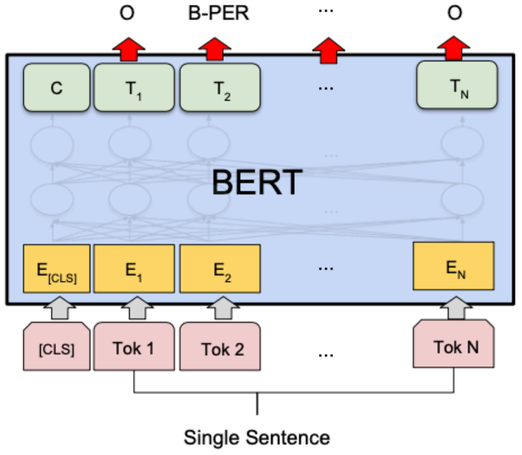
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Language Models are Unsupervised Multitask Learners

YouTube Video

– Question Answering
– Machine Translation
– Reading Comprehension
– Summarization

GPT-2 → a 1.5B parameter Transformer

$(x_1, x_2, \ldots, x_n)$ → set of examples

$x = (s_1, s_2, \ldots, s_n)$ → variable-length sequence of symbols

$$p(x) = \prod_{i=1}^{n} p(s_n|s_1, \ldots, s_{n-1})$$

$$p(s_{n-k}, \ldots, s_n|s_1, \ldots, s_{n-k-1})$$

$$p(output|input)$$

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

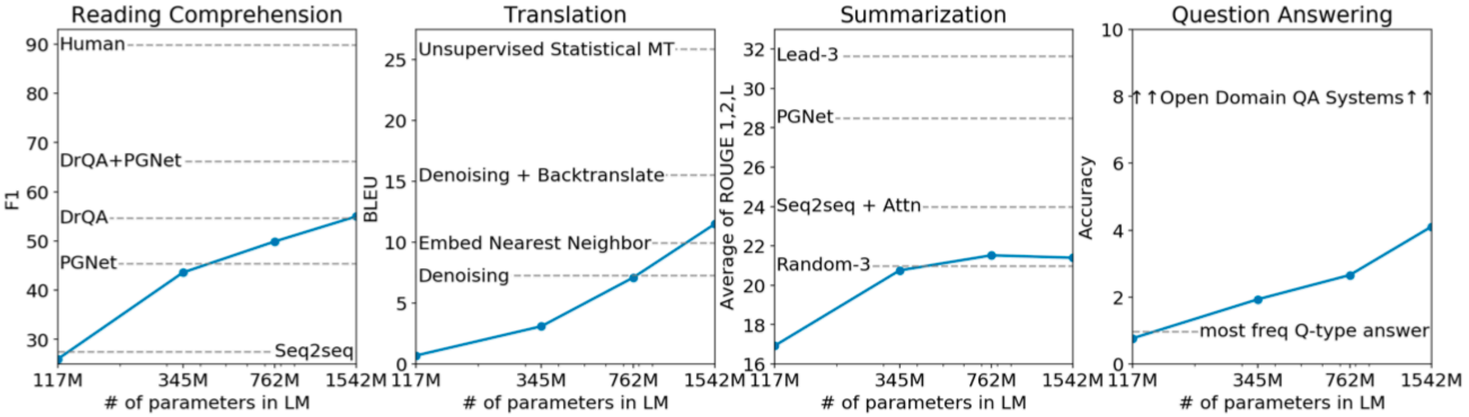
Task conditioning: 1. architectural level
2. algorithmic level
(e.g., Model-Agnostic Meta-Learning)

translation training example
`(translate to french, english text, french text)`

reading comprehension training example
`(answer the question, document, question, answer)`

pre-processing steps: lower-casing,
tokenization, and out-of-vocabulary tokens

Byte Pair Encoding (BPE): a practical middle ground
between character and word level language modeling



"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I'm not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose**," which translates as, "**Lie lie and something will always remain.**"

"I hate the word '**perfume**,"' Burr says. 'It's somewhat better in French: '**parfum**.'

If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

"**Brevet Sans Garantie Du Gouvernement**", translated to English: "**Patented without government warranty**".

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?
A: 54

Q: where does she live?
A:

Model answer: Stockholm
Turker answers: Sweden, Sweden, in Sweden, Sweden

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI Blog* 1.8 (2019): 9.

ALBERT: A Lite BERT for Self-supervised Learning of Language Representations

Boulder

YouTube Video

$V \rightarrow$ vocabulary size ($\approx$ 30,000)
$E \rightarrow$ vocabulary embedding size
$L \rightarrow$ number of encoder layers
$H \rightarrow$ hidden size
$4H \rightarrow$ feedforward/filter size
$H/64 \rightarrow$ number of attention heads

Factorized embedding parameterization

Reduce the embedding parameters from $\mathcal{O}(V \times H)$ to $\mathcal{O}(V \times E + E \times H)$

This parameter reduction is significant when $H \gg E$.

Cross-layer parameter sharing

Inter-sentence coherence loss

In addition to the masked language modeling (MLM) loss, BERT uses an additional loss called next-sentence prediction (NSP).

Positive examples:
consecutive segments from the corpus

Negative examples:
pairing segments from different documents

sentence-order prediction (SOP) loss

Positive examples:
consecutive segments from the corpus

Negative examples:
swapped consecutive segments

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

The effect of vocabulary embedding size on the performance of ALBERT-base.

Model	E	Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base not-shared	64	87M	89.9/82.9	80.1/77.8	82.9	91.5	66.7	81.3
	128	89M	89.9/82.8	80.3/77.3	83.7	91.5	67.9	81.7
	256	93M	90.2/83.2	80.3/77.4	84.1	91.9	67.3	81.8
	768	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base all-shared	64	10M	88.7/81.4	77.5/74.8	80.8	89.4	63.5	79.0
	128	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1
	256	16M	88.8/81.5	79.1/76.3	81.5	90.3	63.4	79.6
	768	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8

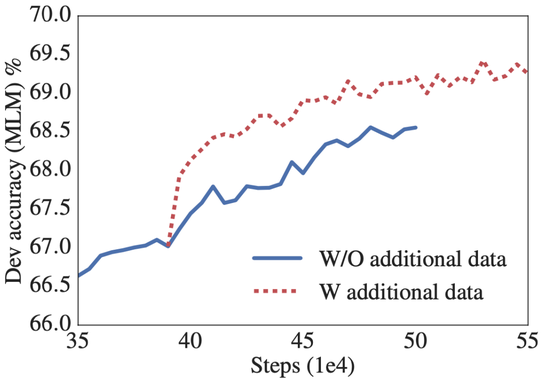
The effect of cross-layer parameter-sharing strategies, ALBERT-base configuration.

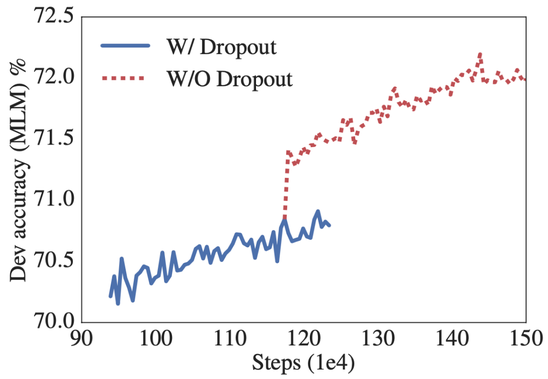
Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
ALBERT base E=768	all-shared	31M	88.6/81.5	79.2/76.6	82.0	90.6	63.3	79.8
	shared-attention	83M	89.9/82.7	80.0/77.2	84.0	91.4	67.7	81.6
	shared-FFN	57M	89.2/82.1	78.2/75.4	81.5	90.8	62.6	79.5
	not-shared	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3
ALBERT base E=128	all-shared	12M	89.3/82.3	80.0/77.1	82.0	90.3	64.0	80.1
	shared-attention	64M	89.9/82.8	80.7/77.9	83.4	91.9	67.6	81.7
	shared-FFN	38M	88.9/81.6	78.6/75.6	82.3	91.7	64.4	80.2
	not-shared	89M	89.9/82.8	80.3/77.3	83.2	91.5	67.9	81.6

NSP vs. SOP

	Intrinsic Tasks		
SP tasks	MLM	NSP	SOP
None	54.9	52.4	53.3
NSP	54.5	90.5	52.0
SOP	54.0	78.9	86.5

Downstream Tasks					
SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
88.6/81.5	78.1/75.3	81.5	89.9	61.7	79.0
88.4/81.5	77.2/74.6	81.6	**91.1**	62.3	79.2
89.3/82.3	**80.0/77.1**	**82.0**	90.3	**64.0**	**80.1**





Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." *arXiv preprint arXiv:1909.11942* (2019).

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Boulder

Robustly **o**ptimized **BERT** **a**pproach
replication study of BERT pre-training
BERT
$x_1, x_2, \ldots, x_N \rightarrow$ segment 1 (sequence of tokens)
$y_1, y_2, \ldots, y_M \rightarrow$ segment 2 (sequence of tokens)
$[CLS], x_1, x_2, \ldots, x_N, [SEP], y_1, y_2, \ldots, y_M, [EOS] \rightarrow$ single input sequence
$M + N < T \rightarrow$ maximum sequence length during training
$L \rightarrow$ number of layers
$A \rightarrow$ number of self-attention heads
$H \rightarrow$ hidden dimension
– masked language modeling (MLM)
15% of input tokens are selected for possible replacement. Of the selected tokens, 80% are replaced with [MASK], 10% are left unchanged, and 10% are replaced by a randomly selected vocabulary token.
– next sentence prediction (NSP)
positive examples (consecutive sentences from the text corpus)
negative examples (pairing segments from different documents)
The NSP objective was designed to improve performance on downstream tasks, such as Natural Language Inference
$T = 512 \rightarrow$ maximum length
$S = 1,000,000 \rightarrow$ updates
$B = 256 \rightarrow$ minibatch
16GB of uncompressed text (BOOKCORPUS + English WIKIPEDIA)
RoBERTa's modifications are simple, they include: (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; (4) dynamically changing the masking pattern applied to the training data; and (5) collecting a large new dataset (CC-NEWS) of comparable size to other privately used datasets, to better control for training set size effects.

160GB of uncompressed text:
– BOOKCORPUS + English WIKIPEDIA (16GB)
– CC-NEWS (76GB)
– OPENWEBTEXT (38GB)
– STORIES (31GB)

masking once during data preprocessing ←

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Our reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

Model	SQuAD 1.1 EM	SQuAD 1.1 F1	SQuAD 2.0 EM	SQuAD 2.0 F1
Single models on dev, w/o data augmentation				
$\text{BERT}_{\text{LARGE}}$	84.1	90.9	79.0	81.8
$\text{XLNet}_{\text{LARGE}}$	**89.0**	94.5	86.1	88.8
RoBERTa	88.9	**94.6**	**86.5**	**89.4**
Single models on test (as of July 25, 2019)				
$\text{XLNet}_{\text{LARGE}}$			86.3†	89.1†
RoBERTa			86.8	89.8
XLNet + SG-Net Verifier			**87.0**†	**89.9**†

batch size ←

bsz	steps	lr	ppl	MNLI-m	SST-2
256	1M	1e-4	3.99	84.7	92.7
2K	125K	7e-4	**3.68**	**85.2**	**92.9**
8K	31K	1e-3	3.77	84.6	92.8

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

Results on the RACE test set.

Model	Accuracy	Middle	High
Single models on test (as of July 25, 2019)			
$\text{BERT}_{\text{LARGE}}$	72.0	76.6	70.1
$\text{XLNet}_{\text{LARGE}}$	81.7	85.4	80.2
RoBERTa	**83.2**	**86.5**	**81.3**

Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context

YouTube Playlist

$x = (x_1, x_2, \ldots, x_T) \rightarrow$ corpus of tokens

$p(x) = \prod_t p(x_t|x_{<t}) \rightarrow$ Language Modeling

encode the context $x_{<t}$
into a fixed size hidden state

$$\begin{cases} s_\tau = [x_{\tau,1}, x_{\tau,2}, \ldots, x_{\tau,L}] \\ s_{\tau+1} = [x_{\tau+1,1}, x_{\tau+1,2}, \ldots, x_{\tau+1,L}] \end{cases}$$

consecutive segments of length L

$h_\tau^n \in \mathbb{R}^{L \times d}$

n-th layer hidden state sequence produced for the τ-th segment s_τ

$\tilde{h}_{\tau+1}^{n-1} = [\text{SG}(h_\tau^{n-1}), h_{\tau+1}^{n-1}] \in \mathbb{R}^{2L \times d}$

stop gradient

$q_{\tau+1}^n = h_{\tau+1}^{n-1} W_q^T$

$k_{\tau+1}^n = \tilde{h}_{\tau+1}^{n-1} W_k^T$

$v_{\tau+1}^n = \tilde{h}_{\tau+1}^{n-1} W_v^T$

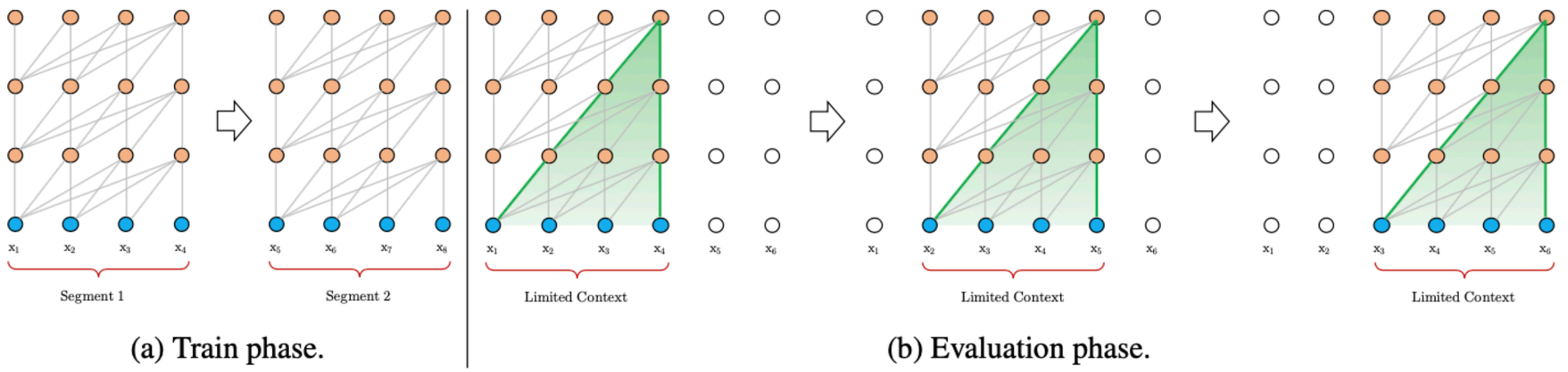
$h_{\tau+1}^n = \text{transformer-layer}(q_{\tau+1}^n, k_{\tau+1}^n, v_{\tau+1}^n)$

$m_\tau^n \in \mathbb{R}^{M \times d}$ (e.g., $m_\tau^n = h_\tau^n$) $\rightarrow$ memory

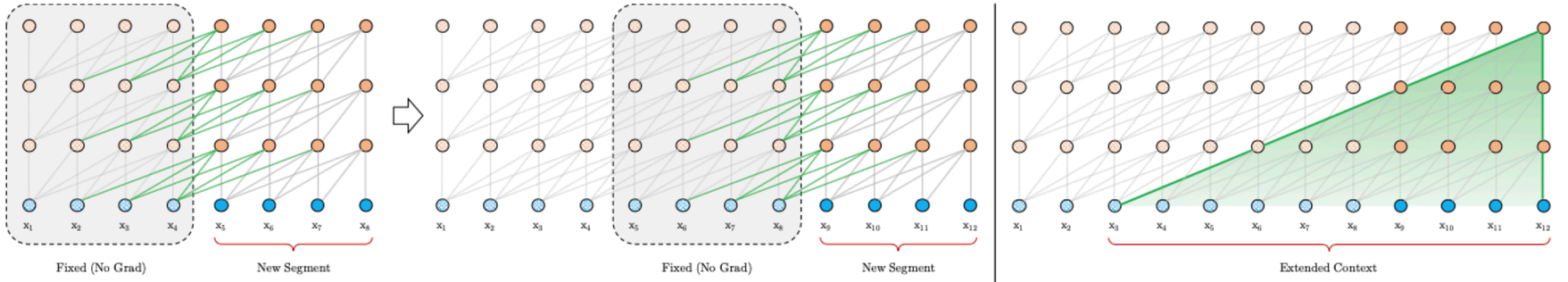
Relative Positional Encodings

$U \in \mathbb{R}^{L_{\max} \times d}$

"absolute" positional encodings



(a) Train phase. (b) Evaluation phase.



(a) Training phase. (b) Evaluation phase.

The model has no information to distinguish the positional difference between $x_{\tau,j}$ and $x_{\tau+1,j}$ for any $j = 1, \ldots, L$

$R \in \mathbb{R}^{L_{\max} \times d}$ i-th row: relative distance of i between two positions

"relative" positional encoding

$$\mathbf{A}_{i,j}^{\text{abs}} = \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{U}_j}_{(b)} + \underbrace{\mathbf{U}_i^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{E}_{x_j}}_{(c)} + \underbrace{\mathbf{U}_i^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{U}_j}_{(d)}.$$

attention score between query q_i and key vector k_j

$$\mathbf{A}_{i,j}^{\text{rel}} = \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,E} \mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,R} \mathbf{R}_{i-j}}_{(b)} + \underbrace{u^\top \mathbf{W}_{k,E} \mathbf{E}_{x_j}}_{(c)} + \underbrace{v^\top \mathbf{W}_{k,R} \mathbf{R}_{i-j}}_{(d)}.$$

Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." *arXiv preprint arXiv:1901.02860* (2019).

XLNet: Generalized Autoregressive Pretraining for Language Understanding

Boulder

YouTube Video

Pre-training objectives:
- Autoencoding (AE) Language Modeling
- Autoregressive (AR) Language Modeling

$x = (x_1, x_2, \ldots, x_T)$

$$p(x) = \prod_{t=1}^{T} p(x_t|x_{<t}) \rightarrow \text{forward product}$$

$$p(x) = \prod_{t=T}^{1} p(x_t|x_{>t}) \rightarrow \text{backward product} \Big\} \text{AR}$$

[MASK] → AE (pretrain-finetune discrepency)

all possible permutations of the factorization order

$$\max_\theta \log p_\theta(\mathbf{x}) = \sum_{t=1}^{T} \log p_\theta(x_t \mid \mathbf{x}_{<t})$$

$$= \sum_{t=1}^{T} \log \frac{\exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x_t)\right)}{\sum_{x'} \exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x')\right)}$$

$x \rightarrow \hat{x} \rightarrow \overline{x}$ masked tokens

corrupted ([MASK] 15%)

$$\max_\theta \log p_\theta(\bar{\mathbf{x}} \mid \hat{\mathbf{x}}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t \mid \hat{\mathbf{x}})$$

$m_t = 1$ indicates x_t is masked.

$$= \sum_{t=1}^{T} m_t \log \frac{\exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x_t)\right)}{\sum_{x'} \exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x')\right)}$$

$\mathcal{Z}_T \rightarrow$ set of all possible permutations of $\{1, 2, \ldots, T\}$

$z_t \rightarrow t$-th element of a permutation $z \in \mathcal{Z}_T$

$z_{<t} \rightarrow$ the first $t-1$ elements of a permutation $z \in \mathcal{Z}_T$

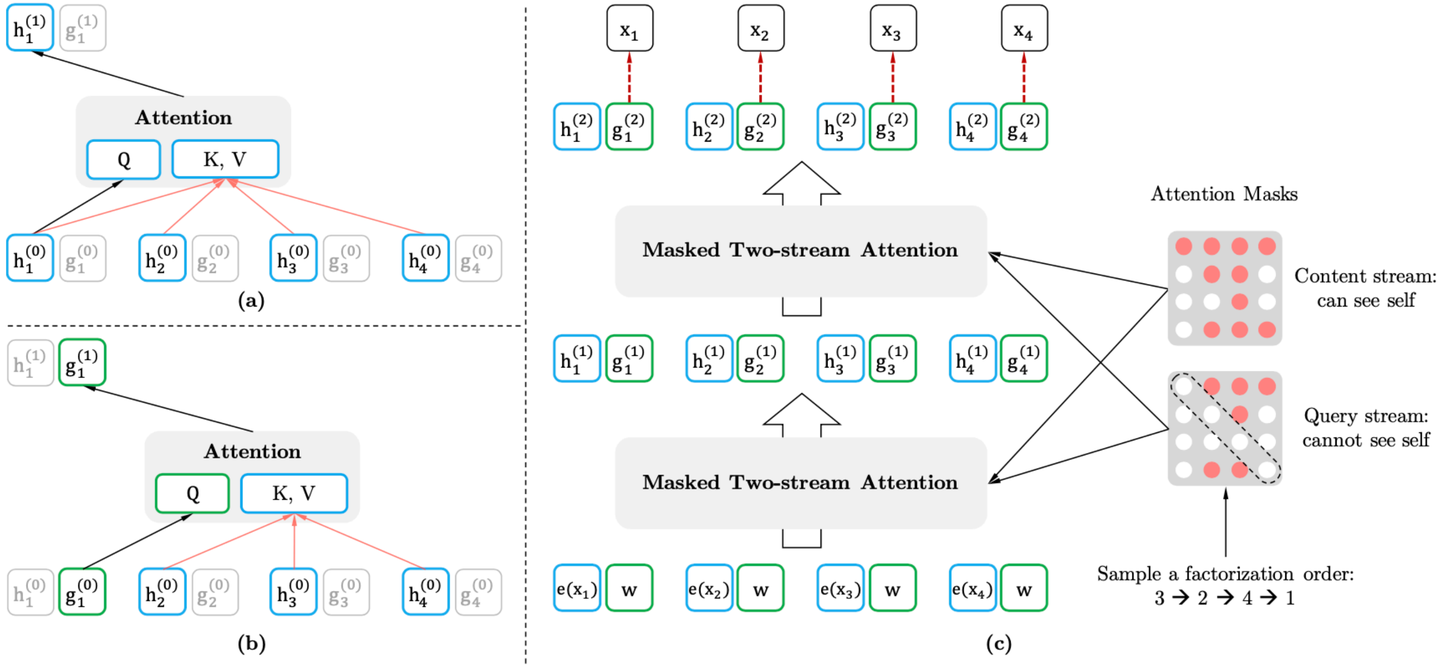
$$\mathcal{J}_{\text{BERT}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{is a city}),$$

$$\mathcal{J}_{\text{XLNet}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{New}, \text{is a city}).$$

Partial Prediction

$$\max_\theta \quad \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_T}\left[\sum_{t=1}^{T} \log p_\theta(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{<t}})\right] \qquad \max_\theta \quad \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_T}\left[\log p_\theta(\mathbf{x}_{\mathbf{z}_{>c}} \mid \mathbf{x}_{\mathbf{z}_{\le c}})\right] = \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_T}\left[\sum_{t=c+1}^{|\mathbf{z}|} \log p_\theta(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{<t}})\right]$$

Make the objective function depend on the position (i.e., the value of z_t) it will predict.



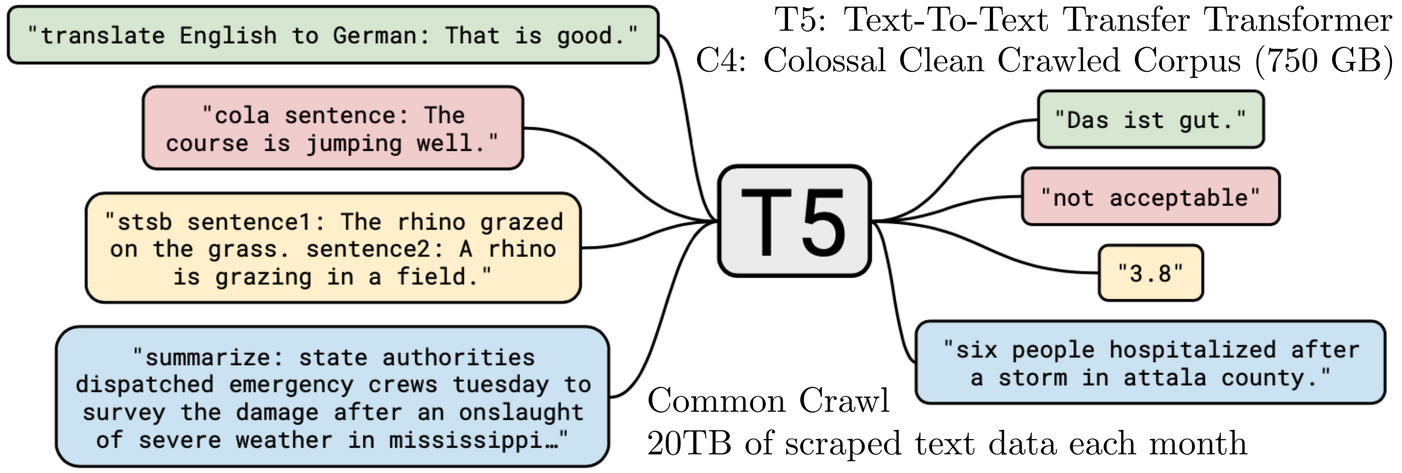
$$g_{z_t}^{(m)} \leftarrow \text{Attention}(\text{Q} = g_{z_t}^{(m-1)}, \text{KV} = \mathbf{h}_{\mathbf{z}_{<t}}^{(m-1)}; \theta), \quad \text{(query stream: use } z_t \text{ but cannot see } x_{z_t})$$

$$h_{z_t}^{(m)} \leftarrow \text{Attention}(\text{Q} = h_{z_t}^{(m-1)}, \text{KV} = \mathbf{h}_{\mathbf{z}_{\le t}}^{(m-1)}; \theta), \quad \text{(content stream: use both } z_t \text{ and } x_{z_t}).$$

Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems*. 2019.

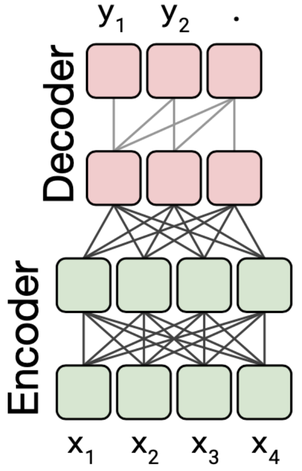
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

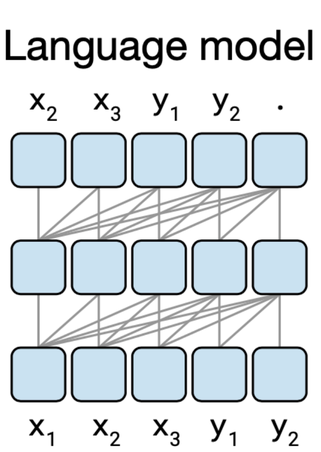
Boulder

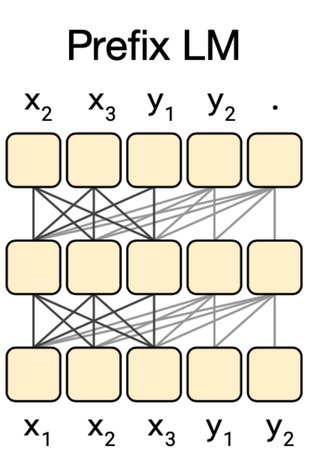


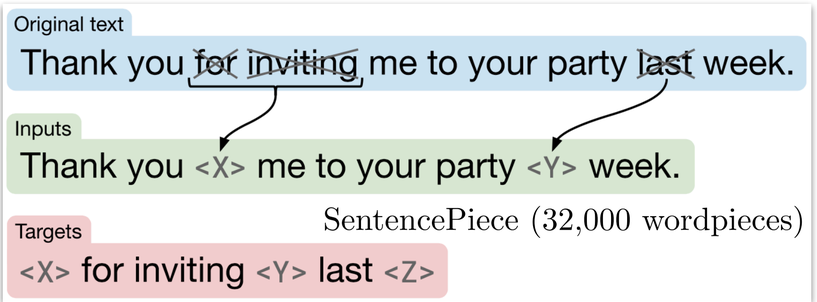
T5: Text-To-Text Transfer Transformer
C4: Colossal Clean Crawled Corpus (750 GB)

Common Crawl
20TB of scraped text data each month

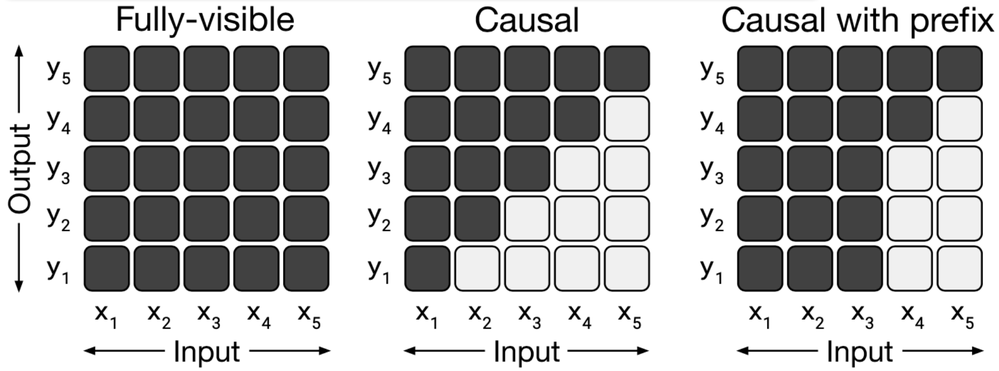








Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
MASS-style Song et al. (2019)	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>



$1/\sqrt{\max(n, 10^4)}$ → "inverse square root" learning rate schedule

AdaFactor Optimiser

Data set	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019).

Cross-lingual Language Model Pretraining

YouTube Video

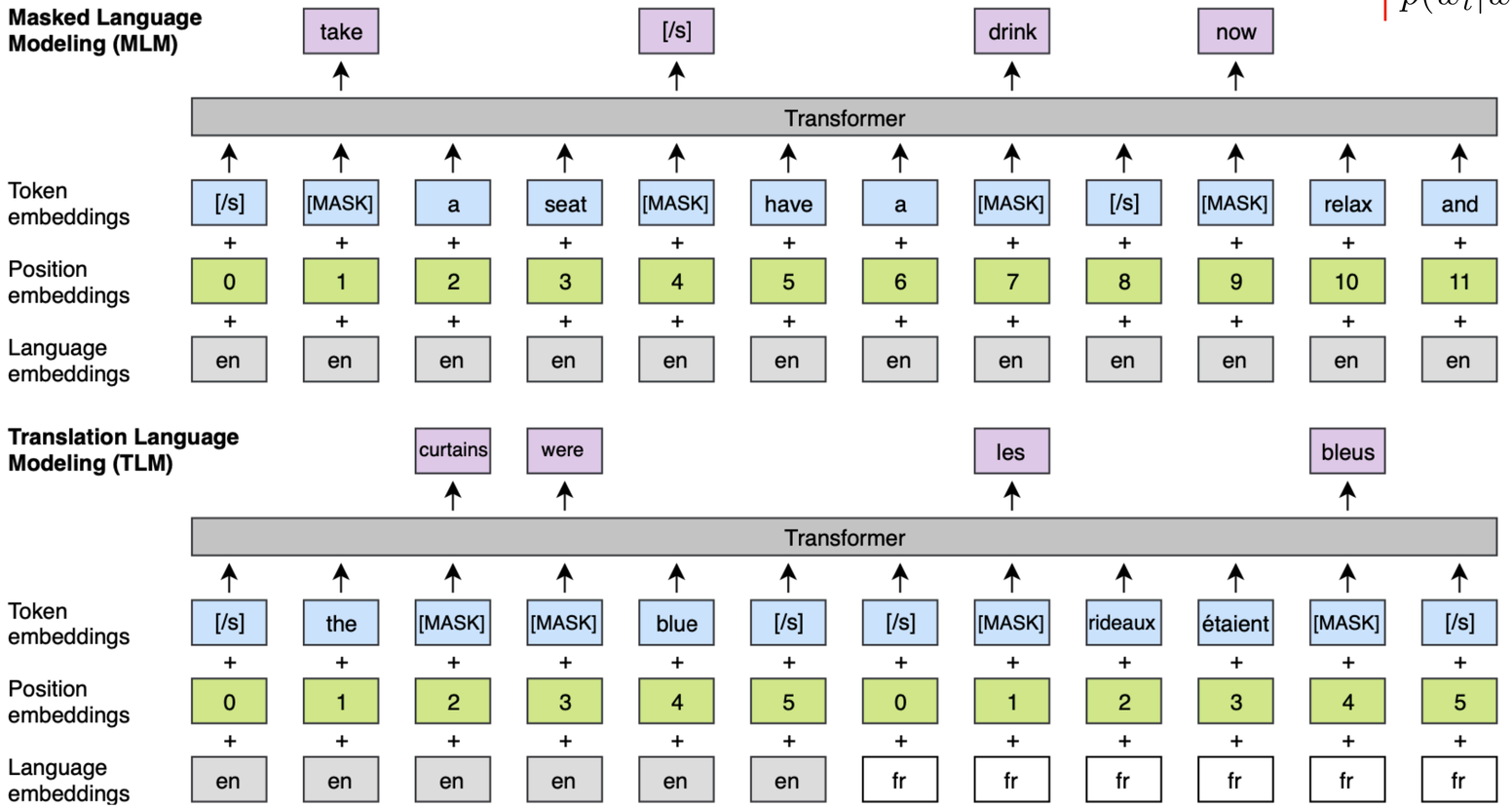
– multiple languages
– cross-lingual pretraining
cross-lingual language models (XLMs)
– monolingual data (unsupervised)
– parallel data (supervised)
$\{C_i\}_{i=1}^N \rightarrow N$ monolingual corpora
$n_i \rightarrow$ number of sentences in C_i

Shared sub-word vocabulary
Shared vocabulary created using Byte Pair Encoding (BPE)
Learn the BPE splits on the concatenation of sentences sampled randomly from the monolingual corpora

Sentences are sampled according to a multinomial distribution with probabilities $\{q_i\}_{i=1}^N$

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^N p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^N n_k}, \alpha = 0.5$$

Alleviates the bias towards high-resource languages!
Causal Language Modeling (CLM)
$p(w_t|w_1, \ldots, w_{t-1}; \theta) \rightarrow$ a Transformer language model



Results on cross-lingual classification accuracy. Test accuracy on the 15 XNLI languages.

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Results on unsupervised MT. BLEU

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

Results on supervised MT. BLEU scores

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro → en	28.4	31.5	35.3
ro ↔ en	28.5	31.5	35.6
ro ↔ en + BT	34.4	37.0	**38.5**

Low-resource language model

Training languages	Nepali perplexity
Nepali	157.2
Nepali + English	140.1
Nepali + Hindi	115.6
Nepali + English + Hindi	**109.3**

Lample, Guillaume, and Alexis Conneau. "Cross-lingual language model pretraining." *arXiv preprint arXiv:1901.07291* (2019).

Language Models are Few-Shot Learners

YouTube Video

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer        <- example #1
```
↓
gradient update
↓
```
peppermint => menthe poivrée      <- example #2
```
↓
gradient update
↓
...
↓
```
plush giraffe => girafe peluche   <- example #N
```
gradient update
```
cheese =>                         <- prompt
```

In-context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:      <- task description
cheese =>                         <- prompt
```

One-shot

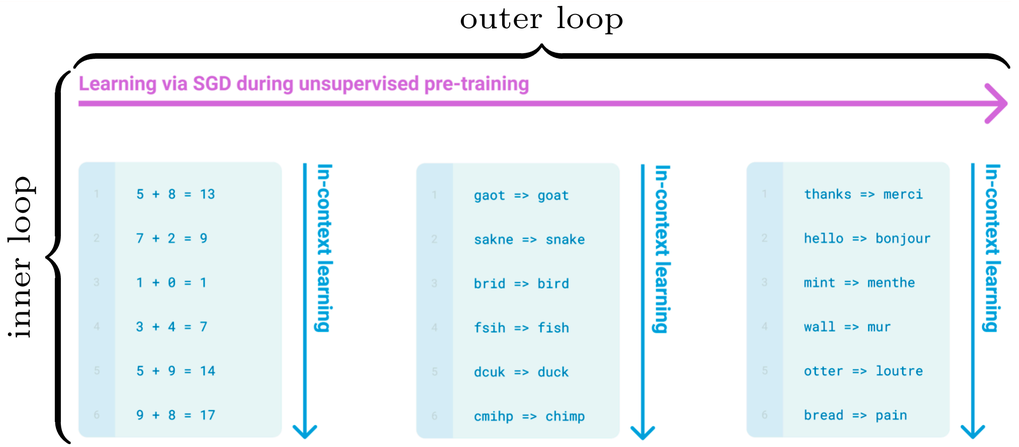
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:      <- task description
sea otter => loutre de mer        <- example
cheese =>                         <- prompt
```

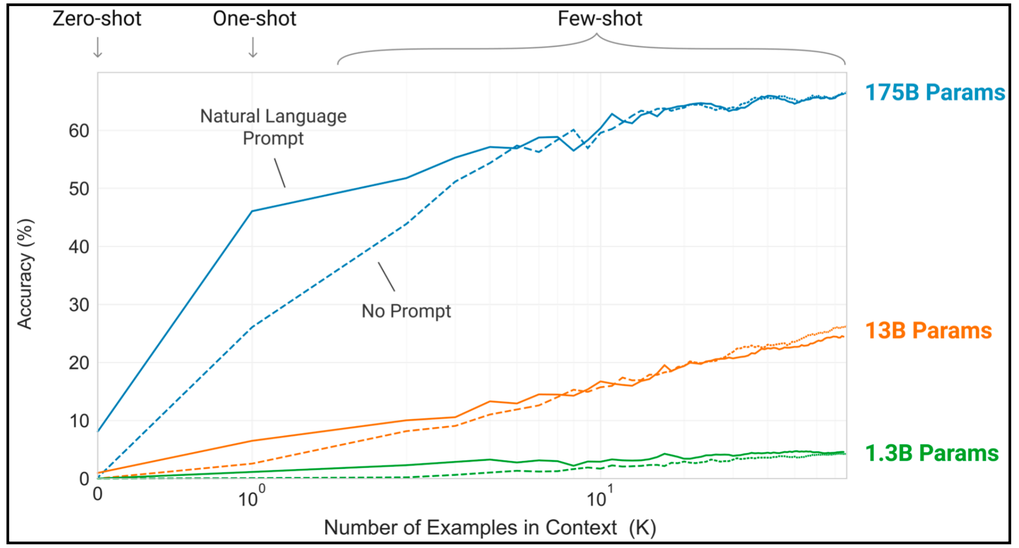
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:      <- task description
sea otter => loutre de mer        <- examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                         <- prompt
```



Larger models make increasingly efficient use of in-context information



– Common Crawl dataset (filtered) – Webtext dataset
– Two internet-based books corpora – English language Wikipedia

Brown, Tom B., et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

Boulder

Questions?

YouTube Playlist