Deep learning in NLP

Part 5: Bert Variants

Zeinab Rahimi

RoBERTa

- RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al, University of Washington and Facebook, 2019)
  - BERT را با تعداد epoch بیشتر و/یا داده‌های بیشتر آموزش داده است.
    - استفاده از ۱۶۰ گیگابایت متن به جای مجموعه داده ۱۶ گیگابایتی که در ابتدا برای آموزش BERT استفاده می شد
    - افزایش تعداد تکرارها را از ۱۰۰ هزار به ۳۰۰ هزار و سپس به ۵۰۰ هزار
  - تغییر و پویاسازی الگوی ماسک گذاری اعمال شده بر روی داده های آموزشی (در هر epoch کلمات mask متفاوت می شوند)
  - حذف هدف NSP از رویه آموزشی
  - نشان داد که دوره های بیشتر به تنهایی کمک می‌کند، حتی بر روی داده های مشابه
  - داده های بیشتر نیز کمک می کند.

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-

XLNet

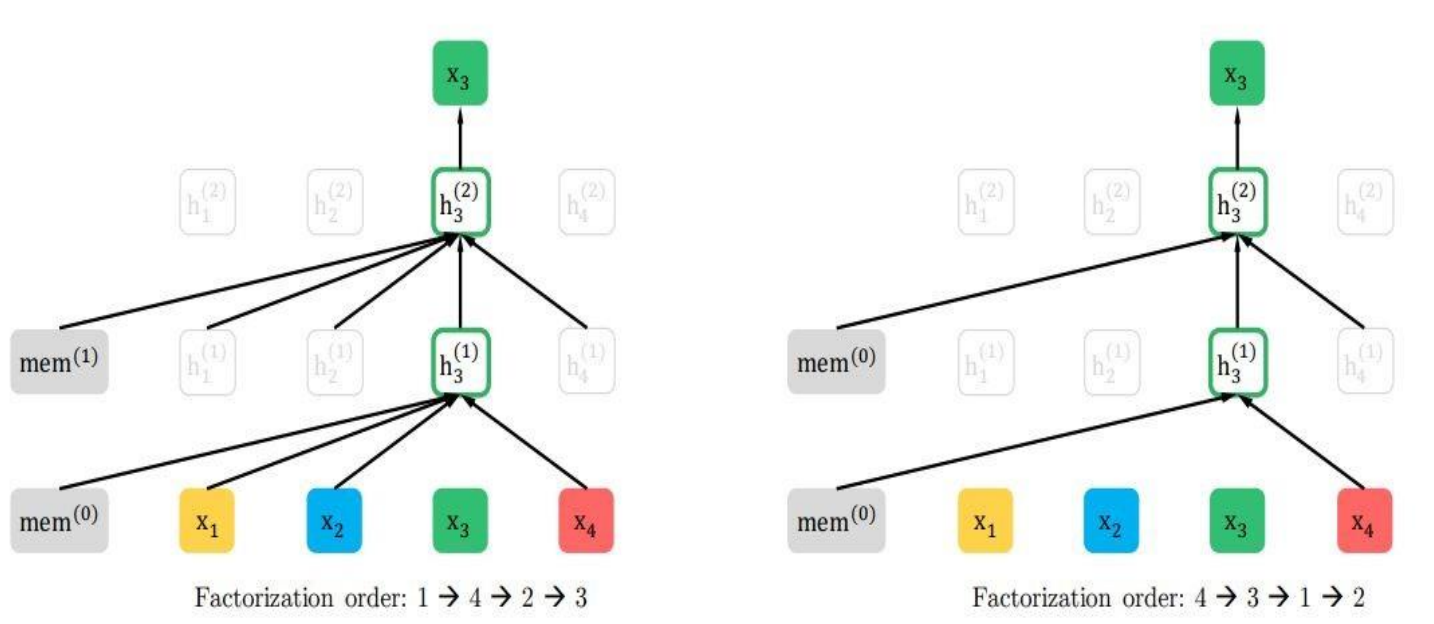
- XLNet: Generalized Autoregressive Pretraining for Language Understanding (Yang et al, CMU and Google, 2019)

- بر خلاف BERT استفاده از Auto regression (عدم وابستگی به ماسک گذاری)
- مثل BERT مبتنی بر شبکه دوجهته و نگاه به کلمات قبل و بعدی

- نوآوری شماره ۱: تعبیه موقعیت نسبی
  - مکانیسم بازگشت: فراتر از توالی فعلی برای شناسایی وابستگی های طولانی مدت.
  - استفاده از موقعیت نسبی برای کارکرد مکانیسم بازگشت.
  - با توجه به فواصل زوجی هر دو کلمه (بدون وجود محدودیت طول در دنباله ورودی)، اطلاعات موقعیت نسبی به عنوان یک جزء اضافی برای کلیدها به مدل ارائه می شود و در ماتریس های k و v در فرآیند self-attention یاد گرفته می شوند.

XLNet

- نوآوری شماره ۲: مدل زبانی جایگشتی
  - در یک مدل زبانی از چپ به راست، هر کلمه بر اساس تمام کلمات سمت چپ آن پیش‌بینی می‌شود.
  - در عوض: توالی هر جمله آموزشی را به‌طور تصادفی تغییر می دهد، معادل ماسک کردن، اما پیش‌بینی‌های بسیار بیشتری در هر جمله را می‌توان به طور مؤثر با ترنسفورمر انجام داد. (تقویت مدل با در نظرگرفتن احتمال همه جایگشت های ممکن)



XLNet

- همچنین از داده های بیشتر و مدل های بزرگتر استفاده کرد.
- نشان داد که اعمال نوآوری ها در BERT حتی با داده ها و اندازه مدل مشابه بهبود ایجاد کرده است.
- XLNet در ۲۰ تسک، BERT را با اختلاف بسیار زیاد شکست می دهد.
- نتایج :XLNet

Model	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B
Single-task single models on dev								
BERT [2]	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0
RoBERTa [21]	90.2/90.2	94.7	92.2	**86.6**	96.4	**90.9**	68.0	92.4
XLNet	**90.8/90.8**	**94.9**	**92.3**	85.9	**97.0**	90.8	**69.0**	**92.5**

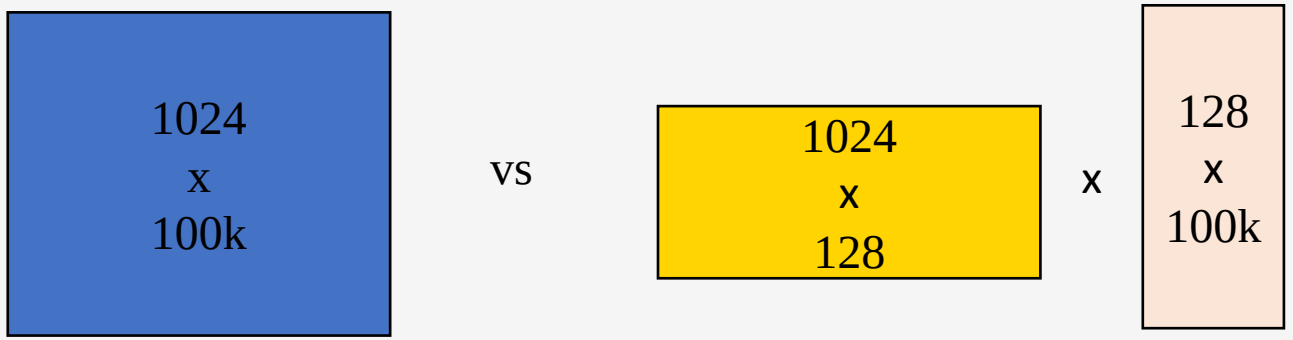
ALBERT

- ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (Lan et al, Google and TTI Chicago, 2019)

  - BERT مشکل بزرگی اندازه مدل‌های زبانی از پیش آموزش‌دیده را دارد که محدودیت‌های حافظه، زمان آماده‌سازی طولانی‌تر دارد.

  - نوآوری شماره ۱: پارامترگذاری جاسازی با فاکتورگیری

    - در BERT، بعد جاسازی به اندازه لایه پنهان گره خورده است. افزایش اندازه لایه پنهان دشوار است زیرا اندازه جاسازی و در نتیجه پارامترها را افزایش می دهد.
    - Albert از اندازه جاسازی کوچک (مثلاً ۱۲۸ بعد) استفاده می کند و سپس آن را با استفاده از ماتریس پارامتر به اندازه هیدن ترنسفورمر (مثلاً ۱۰۲۴) تبدیل می کند.



ALBERT

- نوآوری شماره ۲: به اشتراک گذاری پارامترهای متقابل
  - همه پارامترها را بین لایه های ترانسفورمر به اشتراک می گذارد. (افزایش کارایی)
- نتایج:

Models	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS
Single-task single models on dev								
BERT-large	86.6	92.3	91.3	70.4	93.2	88.0	60.6	90.0
XLNet-large	89.8	93.9	91.8	83.8	95.6	89.2	63.6	91.8
RoBERTa-large	90.2	94.7	**92.2**	86.6	96.4	**90.9**	68.0	92.4
ALBERT (1M)	90.4	95.2	92.0	88.1	96.8	90.2	68.7	92.7
ALBERT (1.5M)	**90.8**	**95.3**	**92.2**	**89.2**	**96.9**	**90.9**	**71.4**	**93.0**

- ALBERT از لحاظ تعداد پارامترها سبک تر است نه لزوما سرعت

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

T5

- Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (Raffel et al, Google, 2019)
  - تغییر و ارزیابی بسیاری از جنبه های pre-training:
  - سایز مدل
  - مقدار داده های آموزشی
  - دامنه/تمیز بودن داده های آموزشی
  - جزئیات تابع هزینه pre-training (به عنوان مثال، طول متن پوشانده شده)
  - دستور العمل تنظیم دقیق (به عنوان مثال، فقط به لایه های خاصی اجازه تنظیم دقیق می دهد)
  - آموزش مالتی تسک

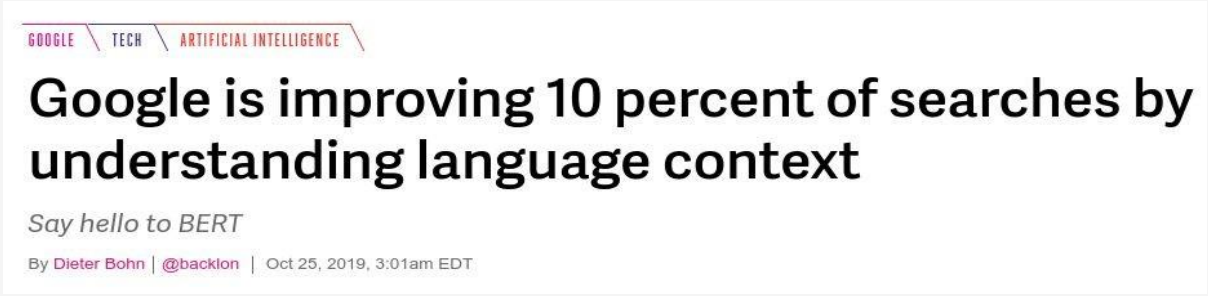
T5

- نتیجه گیری:
- افزایش اندازه مدل و مقدار داده های آموزشی کمک زیادی می کند.
- بهترین مدل پارامترهای ۱۱میلیون تاست (BERT-Large 330M است)، آموزش داده شده بر روی ۱۲۰میلیون کلمه متن خزش شده تمیز مشترک
- استراتژی دقیق masking چندان مهم نیست.
- نتایج عمدتا منفی برای تنظیم بهتر و استراتژی های چند وظیفه ای
- نتایج T5:

Rank	Name	Model	Score
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines	89.8
2	T5 Team - Google	T5	89.3
3	Zhuiyi Technology	RoBERTa-mtl-adv	85.7
4	Facebook AI	RoBERTa	84.6
5	IBM Research AI	BERT-mtl	73.5
6	SuperGLUE Baselines	BERT++	71.5
		BERT	69.0

Applying Models to Production Services

- BERT و سایر مدل های زبان از پیش آموزش دیده بسیار بزرگ و گران هستند.
- چگونه شرکت ها آنها را برای خدمات تولید با تاخیر کم اعمال می کنند؟


GOOGLE \ TECH \ ARTIFICIAL INTELLIGENCE

Google is improving 10 percent of searches by understanding language context

Say hello to BERT

By Dieter Bohn | @backlon | Oct 25, 2019, 3:01am EDT


Bing says it has been applying BERT since April

The natural language processing capabilities are now applied to all Bing queries globally.

George Nguyen on November 19, 2019 at 1:38 pm

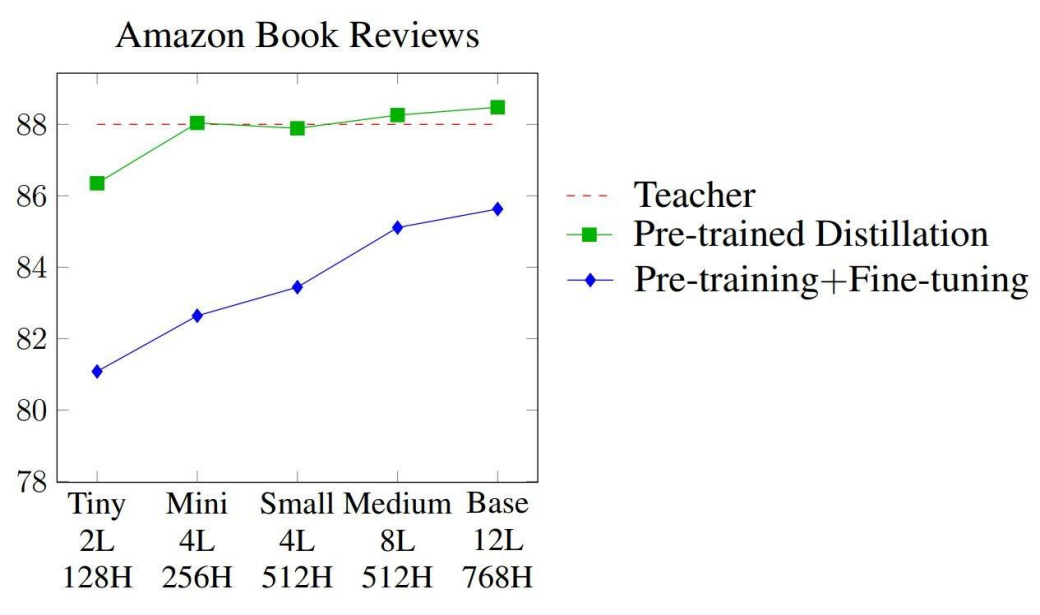
Distillation

- پاسخ: Distillation (مثلاً فشرده سازی مدل)
- ایده از قبل وجود داشته است:
  - فشرده سازی مدل (بوسیلا و همکاران، ۲۰۰۶)
  - فشرده سازی دانش در شبکه عصبی (هینتون و همکاران، ۲۰۱۵)

- یادگیری نسخه تقریبی BERT
  - عملکرد ۹۷٪ با فقط از نیمی از تعداد پارامترها و نیمی از لایه ها (۶ لایه انکودر)
  - شبکه عصبی بزرگ را با یک شبکه کوچکتر تقریب می کند.
  - تکنیک ساده:
    - آموزش "مدل معلم": برای آموزش مدل با حداکثر دقت، از مدل pre-train و تکنیک تنظیم دقیق استفاده می کند. (استفاده از وزن های برت به جای رندم)
    - مدل بسیار کوچکتر "دانش آموز" را (مثلاً ۵۰ برابر کوچکتر) برای تقلید از خروجی معلم آموزش می دهد.

Distillation

- Example distillation results
  - 50k labeled examples, 8M unlabeled examples



Well-Read Students Learn Better: On the Importance of Pre-training Compact Models (Turc et al, 2020)

- مدل های Distillation بهتر از pre-training + fine-tuning با مدل کوچکتر عمل می کنند.

نتیجه

- مدل های زبانی دو جهته از قبل آموزش دیده به طرز باورنکردنی خوب کار می کنند.
- با این حال، مدل ها بسیار سنگین و هزینه بر هستند.
- به نظر می‌رسد که بهبودها (متاسفانه) بیشتر از مدل‌های گران‌تر و داده‌های بیشتر ناشی می‌شوند.
- مشکل استنتاج بیشتر از طریق distillation حل می شود.

Name	Novel ideas	Size	Data	Training	Hardware	Speed	Performance
BERT	The first building block of all Pretrained LMs	Base: 110M Large: 340M	16GB	- Masked Language Modeling (MLM) - Next Sentence Prediction	Base: 16 TPU chips Large: 64 TPU chips	Both versions: 4 days	SOTA on GLUE and SquAD
Transformer-XL	- Introduce recurrence in attention-based models - Relave Positional Encoding						SOTA on WikiText-102, enwiki8, One Billion Word, Penn Treebank
XLNet	Combine Autoregressive and Bi-directional styles.	Comparable to BERT	158GB	Permutation Language Modeling	Large version: 512 TPU v3 chips	Large version: 5.5 days	Outperform BERT, SOTA on 20 tasks
RoBERTa	Better hyper-parameter tuning for BERT	The same as BERT	160GB	MLM	1024 32GB V100 GPUs	1 day	Outperform BERT, comparable to XLNet
DistilBERT	Distill from BERT	66M	Same as BERT	MLM with Distillation	8 16GB V100 GPUs	90 hours	97% of BERT BASE
ALBERT	- Factorized embedding parameterization - Cross-layer parameter sharing - Sentence Order Prediction	Base: 12M Large: 18M XLarge: 60M XXLarge: 235M	Union of data used for XLNet and RoBERTa	MLM and Sentence Order Prediction	64 to 512 TPU V3		Outperform BERT, RoBERTa, XLNet
BART	- Use the whole Transformer architecture - Reconstruct corrupted texts		Same as RoBERTa	Reconstruct corrupted texts			Comparable to RoBERTa, SOTA on some NLG tasks
MobileBERT	- Inverted-Bottleneck BERT - Careful optimizations for distillation	25.1M	Same as BERT	MLM and NSP with distillation	256 TPU v3 chips		Comparable to BERT
ELECTRA	Replaced Token Detection		Same as XLNet	Replaced Token Detection (RTD)	V100 GPUs	Match RoBERTa and XLNet performance with ¼ time	Outperform, RoBERTa, XLNet, ALBERT
ConvBERT	- Mixed attention and convolution - Span-based dynamic convolution - Grouped linear operator	Small: 14M Base: 106M	32GB	Replaced Token Detection		Outperform ELECTRA with ¼ time	Better performance and speed than ELECTRA
DeBERTa	- Disentangled attention - Enhanced mask decoder	Base: 134M	78GB	MLM (optionally RTD)	Base: 64 V100 Large: 96 V100	Base: 10 days Large: 20 days	Outperform ELECTRA, ALBERT
BigBird	Sparse attention		123GB	MLM	8 x 8 TPU v3		SOTA on long-text datasets

GPT-1

- اکثر مدل های پیشرفته NLP به طور خاص در مورد یک تسک خاص مانند طبقه بندی احساسات، رتبه بندی اسناد و غیره با استفاده از یادگیری نظارت شده آموزش دیده اند.

- ویژگی های مدل :GPT-1
  - یادگیری نیمه نظارتی (پیش آموزش بدون نظارت و به دنبال آن تنظیم دقیق نظارت شده) برای تسک های NLP
  - استفاده از مجموعه داده BooksCorpus برای تهیه مدل زبانی
    - BooksCorpus حدود ۷۰۰۰ کتاب منتشر نشده داشت که به تهیه مدل زبانی بر روی اطلاعات نادیده کمک کرد.
  - استفاده از دیکودر ۱۲ لایه با ماسک گذاری برای آموزش
  - نشان داد که مدل زبانی می تواند به مدل در جمع بندی خوب کمک کند.

GPT-2

- OpenAI یک مدل بسیار بزرگ ترانسفورماتور با ۱.۵ بیلیون پارامتر، بر روی مجموعه داده های عظیم و متفاوتی که حاوی متن خزش شده از ۴۵ میلیون صفحه وب سایت
- ایجاد گذرهای منطقی از متن و نتایج امیدوارکننده

- ویژگی های مدل GPT-2:
- آماده سازی مدل زبانی روی مجموعه داده عظیم و متنوع
  - انتخاب صفحات وب سایتی که توسط افراد غربال شده اند.
  - استفاده از مجموعه داده با بیش از ۸ میلیون گزارش و مجموع ۴۰ گیگابایت متن
  - استفاده از سطح رمزگذاری جفت بایت BPE برای ورودی
  - ساخت یک مدل مبتنی بر ترنسفورمر: بزرگترین مدل دارای ۱۵۴۲میلیون پارامتر و ۴۸ لایه
  - مدل اساساً از مدل OpenAI GPT پیروی می کند و تنظیمات چندانی ندارد.

GPT-3

- OpenAI نسخه دوم GPT را توسعه داد که قادر بود متونی طولانی و منسجم تولید کند که تمایز آن با نوشته‌ی انسان‌ها سخت بود.

- ویژگی های GPT3:
  - اندازه شبکه و اطلاعاتی که با آن آموزش داده می‌شود بسیار بزرگ‌تر از نسخه‌ها قبلی
  - GPT-3 در مقایسه با GPT-2 که ۱,۵ میلیارد پارامتر داشته، ۱۷۵ میلیارد پارامتر دارد.
  - برخلاف GPT-2 که بر روی ۴۰ میلیارد گیگابایت متن آموزش دیده، GPT-3 بر روی ۵۷۰ میلیارد گیگابایت متن
  - یادگیری با few-shot learning
  - OpenAIجزئیات کامل الگوریتم های آن را آشکار نکرده است!

روند صعودی تعداد پارامترهای روش های مبتنی بر ترنسفورمر