بررسی مدل‌های تولید متن

Text Generation Models

فهرست

فازبندی پروژه تولید متن

1. فاز اول: **Fine tune GPT2**

در این فاز بااستفاده از مدل اصلی GPT2 و داده های آموزشی (توییتر انگلیسی) finetune جدید این مدل انجام شد.

اگر مورد پذیرش واقع شود باید با داده های اموزشی عربی و فارسی نیز فاین تیون شود برای استفاده در فاز دوم و سوم

2. فاز دوم: **تولید متن براساس انتخاب از احساس**

در این فاز بااستفاده از یکی از مدل های GPT2 (برای تست مدل زبان المانی)و داده های آموزشی برچسب خورده، مدل آموزش دیده شد. و در نهایت با فیلتر احساس مثبت و یا احساس منفی، متن تولید خواهد شد

توضیح: اگر فاز اول موردتایید واقع شود از مدل ساخته شده در فاز دوم استفاده میشود.

و در غیراین صورت از مدل های از قبل تولید شده GPT2 (عربی و فارسی) استفاده خواهیم کرد.

3. فاز سوم: **تولید متن بر اساس انتخاب موضوع**

مقایسه انواع مدل‌های GPT2 , GPT3 در سه زبان

	English	Arabic	Persian
GPT2 Small (124M)	✓	✓ 128 میلیون پارامتر	
Gpt2 Medium (355M)	✓	✓ 355 میلیون پارامتر	✓
GPT2 Large (774M)	✓	✓ 774 میلیون پارامتر	
GPT2 Mega(XL (1.5B))	✓	✓ 1.5میلیارد پارامتر	
GPT3	✓	✓ 10میلیارد پارامتر	

- **مدل انگلیسی GPT2 :**

✓ این مدل را توانستیم در CPU , GPU با داده و تنظیمات جدید fine tuneکنیم.
تنها از مدل SMALL استفاده کنیم که مدلی به نسبت قابل قبول است. سایر مدل ها در GPU احتیاج به رم بالایی دارد.

- **مدل عربی GPT2 (AraGPT)**

  - براساس بررسی‌های انجام شده در مدل عربی نیز با توجه به توضیحات و تحقیات، مناسب است از مدل SMALL استفاده شود.

- **مدل عربی GPT3 :**

  - موسسه نوآوری فناوریTII یک مرکز تحقیقاتی مستقر در ابوظبی، اخیراً توسعه NOOR، بزرگترین مدل پردازش زبان طبیعی زبان عربی را تا به امروز اعلام کرده است.

  - **با ۱۰ میلیارد پارامتر**، محققان TIIمعتقدند که NOORبه "مدل اکتشافی در زبان عربی" تبدیل خواهد شد.
  - ابتسام المازروعی، مدیر واحد مرکز هوش مصنوعی TIIمی‌گوید: «مدل‌های زبانی بزرگ، دنیای پردازش زبان طبیعی را طوفانی کرده‌اند. مجموعه داده‌های منحصر به فرد عربی که برای آموزش مدل جمع‌آوری شده است، نتیجه ماه‌ها کار است که شامل مدیریت، حذف و فیلتر کردن منابع مختلف می‌شود.»

  - به گزارش نشنال نیوز، **این مدل به طور قابل توجهی بزرگتر از AraGPTاست** که قبلاً به عنوان بزرگترین مدل NLPعربی زبان در نظر گرفته می شد. AraGPT**حدود ۱.۵ میلیارد پارامتر داشت** – اندازه افزایش یافته NOORبه آن اجازه می دهد تا وظایف پیچیده تری را انجام دهد. به منظور توسعه این مدل، محققان هوش مصنوعی در TIIآن را با استفاده از متون مختلف از جمله متون فنی، شعر و روزنامه آموزش دادند.

  - این مدل تا حدودی شبیه مدل GPT-3است و می تواند کارهای مختلفی از خلاصه کردن متون گرفته تا توسعه ربات چت و ارزیابی زبان را انجام دهد. TIIبه تلاش‌های خود در زمینه هوش مصنوعی ادامه خواهد داد. NOOR اولین گام مؤسسه به سمت مشارکت در استراتژی امارات متحده عربی برای هوش مصنوعی است، ابتکاری برای تقویت مشخصات فنی این کشور در سراسر جهان.

- **مدل فارسی**

  - نیز یک مدل زبان gpt2است که با پارامترهای هایپر مشابه **gpt2-medium**استاندارد آموزش داده شده است.

مدل عربی GPT2 (Ara GPT)

AraGPT2

- GPT2-base and medium
  - ✓ از معماری gpt2 پیروی می کنند و کاملاً با کتابخانه ترانسفورماتور سازگار است.
  - ✓ در مدل small از ۱۲۸ میلیون پارامتر و در مدل medium از ۳۵۵ میلیون پارامتر استفاده شده است.
  - ✓ برای fine tune مدل ۱۲۸ می‌توان از GPU استفاده کرد.
- GPT2-large and GPT2-mega
  - ✓ در مدل mega از ۱.۵ میلیارد پارامتر و در مدل large از ۷۷۴ میلیون پارامتر استفاده شده است.

- AraGPT2بر روی همان مجموعه داده بزرگ عربی مانند AraBERTv2 آموزش دیده است.

1.https://github.com/minimaxir/gpt-2-simple

2.https://github.com/imcaspar/gpt2-ml/

مدل فارسی GPT2

- مدل فارسی یک مدل زبان gpt2است که با پارامترهای هایپر مشابه **gpt2-medium**استاندارد با تفاوت های زیر آموزش داده شده است:

- اندازه متن از ۱۰۲۴ به ۲۵۶ زیر کلمه کاهش یافته است تا آموزش مقرون به صرفه باشد.
- به جای BPE، google sentence piece tokenizorبرای توکن سازی استفاده می شود.
- مجموعه داده آموزشی فقط شامل متن فارسی است.

gpt-persian با هدف تحقیق در شعر فارسی آموزش داده شده است. به همین دلیل تمام کلمات و اعداد انگلیسی با نشانه های مخصوص جایگزین می شوند و تنها از الفبای استاندارد فارسی به عنوان بخشی از متن ورودی استفاده می شود.

به نظر می رسد با استفاده از داده های زیاد فارسی و پریترین این مدل بتوان نتیجه مطلوبی دریافت کرد.

Transferred model accuracy

دقت مدل GPT2

Trained on ↓ \ Tested on →	Small (124M)	Medium (355M)	Large (774M)	XL (1.5B)
Small (124M)	99.3%	96.6%	90.9%	79.3%
Medium (355M)	99.0%	98.5%	96.9%	91.8%
Large (774M)	98.4%	97.9%	97.9%	95.7%
XL (1.5B)	96.9%	96.7%	96.6%	96.0%

به طور کلی در مدلGPT2 دقت مدل در پاراکترهای SMALL قابل قبول تر است.
از آنجا که برای Fine tune مدل نیز با GPU در دسترس، میتوان تنها از مدل SMALL استفاده کرد، در جدول فوق هم دقت این مدل در آموزش و تست نیز قابل قبول است.

Model	Hardware	num of examples (seq len = 1024)	Batch Size	Num of Steps	Time (in days)
AraGPT2-base	TPUv3-128	9.7M	1792	125K	1.5
AraGPT2-medium	TPUv3-8	9.7M	80	1M	15
AraGPT2-large	TPUv3-128	9.7M	256	220k	3
AraGPT2-mega	TPUv3-128	9.7M	256	780K	9

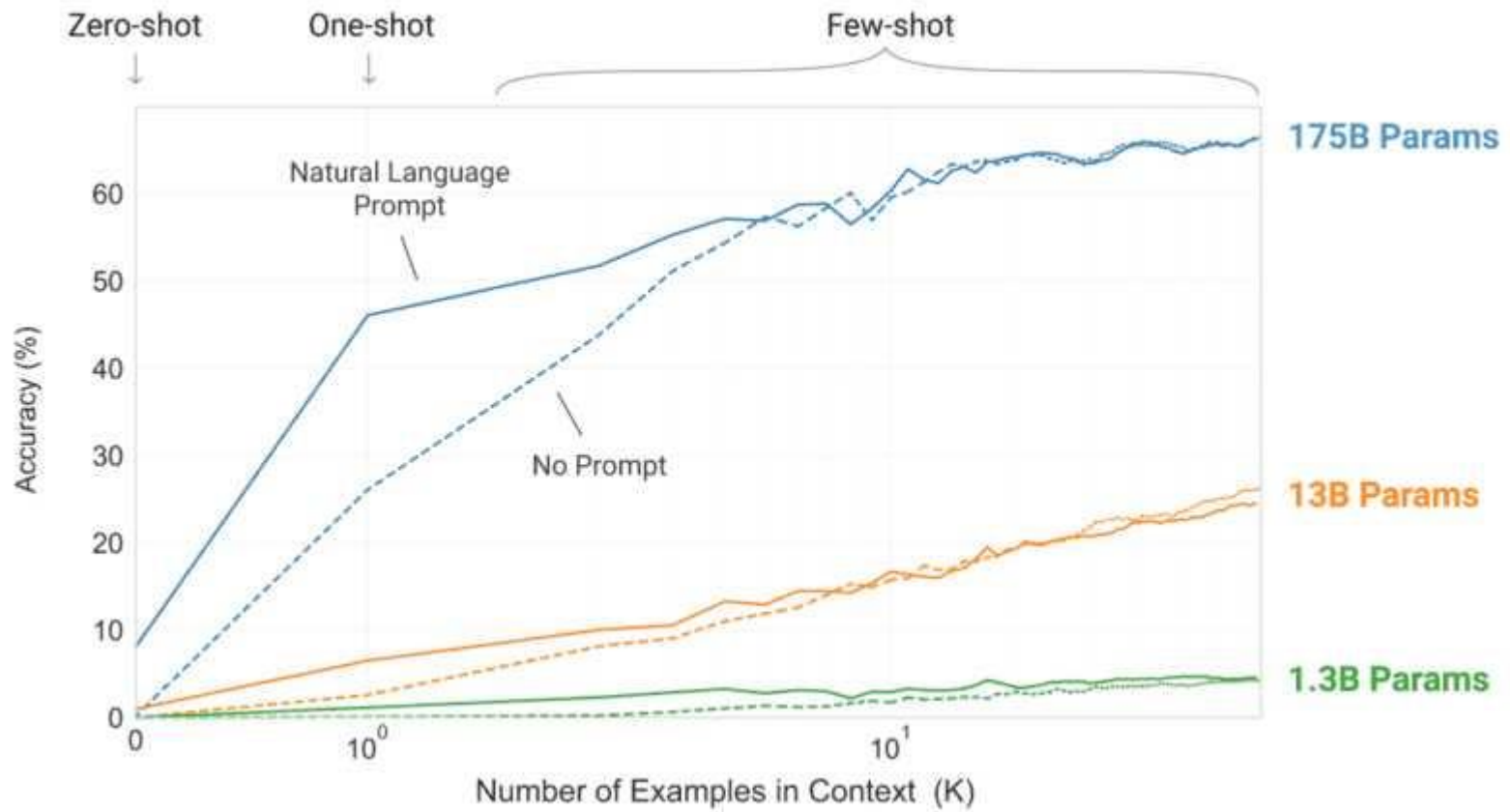
چرا مدل GPT2 را انتخاب می‌کنیم؟

- ✓ به طور کلی، GPT-2در حفظ متن در طول نسل خود بهتر است و **برای تولید متن محاوره ای** مناسب است.
- ✓ متن نیز به طور کلی از نظر گرامری صحیح است، با حروف بزرگ و غلط املایی کمی.
- ✓ مدل اصلی GPT-2بر روی **منابع بسیار متنوعی آموزش داده شد**، که به مدل اجازه می‌دهد اصطلاحاتی را که در متن ورودی دیده نمی‌شوند ترکیب کند.
- ✓ GPT-2تنها می تواند حداکثر ۱۰۲۴ توکن در هر درخواست تولید کند (حدود ۳–۴ پاراگراف متن انگلیسی).

جزئیات انواع مدل GPT-3

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

مقایسه دقت (*accuracy*) در پارامترهای مختلف GPT-3



Zero / One / Few-Shot based task accuracy comparison for models of different sizes

MODEL: generative adversarial network(GANs)

generative adversial networks چیست؟

- ✓ GANها یک مدل مولد مبتنی بر یادگیری عمیق هستند.
- ✓ به طور کلی تر، GANها یک معماری مدل برای آموزش یک مدل مولد هستند و استفاده از مدل های یادگیری عمیق در این معماری رایج ترین است.
- ✓ معماری GANبرای اولین بار در مقاله ۲۰۱۴ توسط Ian Goodfellow و همکارانش توضیح داده شد.
- ✓ ترکیبی از دو مدل از پیش آموزش‌دیده، BERT و GPT-2 است
- ✓ بیشتر برای داده‌های تصویری استفاده می‌شوند
- ✓ معماری GAN مبتنی بر CNN که برای تصاویر استفاده می‌شود، می‌تواند در تولید متن نیز برای تولید نتایج قابل مقایسه با مدل‌های RNNاستفاده شود.
- ✓ معماری مدل GANشامل دو مدل است:

a generator model برای تولید نمونه‌های جدید

a discriminator model یک مدل تفکیک‌کننده برای طبقه‌بندی اینکه آیا نمونه‌های تولید شده واقعی هستند، از دامنه یا جعلی هستند

Model: PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization)

در مدل PEGASUS، جملات مهم از یک سند ورودی حذف/ماسک می‌شوند و با هم به عنوان یک دنباله خروجی از جملات باقی‌مانده، شبیه به یک خلاصه استخراجی، تولید می‌شوند.

- مدل PEGASUS از معماری seq2seqاستفاده میکند.
- نوآوری این مدل در معیار پیش‌آموزش خودنظارتی نهفته است.
- یادگیری خودنظارتی self-supervision ابزار جدید و کارآمدی در **یادگیری عمیق** است.
- این نوع یادگیری ما را از وابستگی داده‌ها به نمونه‌های برچسب‌دار بی‌نیاز می‌کند و باعث می‌شود حجم قابل ملاحظه‌ای از داده‌های بدون برچسب در فرایند آموزش در دسترس قرار گیرد.
- ترکیب مدل‌های مبتنی بر Transformerبا روش پیش‌آموزش خودنظارتی Self-monitoring pre-training (مثل BERT، GPT-2، XLNet، ALBERT، T5و ELECTRA) در مدل‌سازی زبان تاثیر بسزایی بر جای گذاشته است.

ایده اصلی این روش این است که هر قدر روش پیش‌آموزش خودنظارتی به هدف و وظیفه اصلی نزدیکتر باشد، تنظیم دقیق به شکل بهتری انجام خواهد شد.

در مدل PEGASUS، جملات کامل از سند حذف می‌شوند و مدل برای پیش‌بینی این جملات آموزش داده می‌شود

PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

نحوه عملکرد مدل

در PEGASUS ، چندین جمله کامل از اسناد در حین پیش آموزش برداشته می شوند و مدل وظیفه بازیابی آنها را بر عهده دارد.

ورودی

سندی با جملات گم شده (missing) است،

خروجی

جملات مفقود شده ای است که با هم ترکیب (Concatenated) شده‌اند.

مزیت

مزیت این خود نظارتی self-supervision این است که می توانید هر تعداد اسناد و مدارک بدون هیچ‌گونه دخالت انسانی ایجاد کنید

Model: Google T5 (Text-to-Text Transfer Transformer)

- Google T5یک مدل زبان مبتنی بر ترانسفورماتور جدیدتر از BERT که در فوریه ۲۰۲۰ منتشر شد، تنها ۱۱ میلیارد پارامتر داشت.

- Google's T5 یک تبدیل متن به متن Text-To-Text است که یک چارچوب مشترک NLPاست که در آن تمام وظایف NLP در قالب یک متن به متن یکپارچه که در آن ورودی و خروجی همیشه رشته‌های متنی هستند،
- چارچوب متن به متن ما به ما اجازه می‌دهد تا از همان مدل، تابع زیان، و پارامترهای بیش از حد در هر وظیفه NLPاستفاده کنیم، از جمله ترجمه ماشینی، خلاصه‌سازی اسناد، پاسخ به سوال، و وظایف طبقه‌بندی (به عنوان مثال، تحلیل احساسات). ما حتی می‌توانیم T5 را برای کارهای رگرسیون با آموزش آن برای پیش‌بینی نمایش رشته یک عدد به جای خود عدد به کار ببریم.
- این کاملاً متفاوت از مدل‌های سبک BERTاست که فقط می‌توانند یک برچسب کلاس یا یک دامنه از ورودی را تولید کنند. T5به ما این امکان را می دهد که از همان مدل به همراه تابع ضرر و فراپارامترها در هر کار NLPاستفاده کنیم.

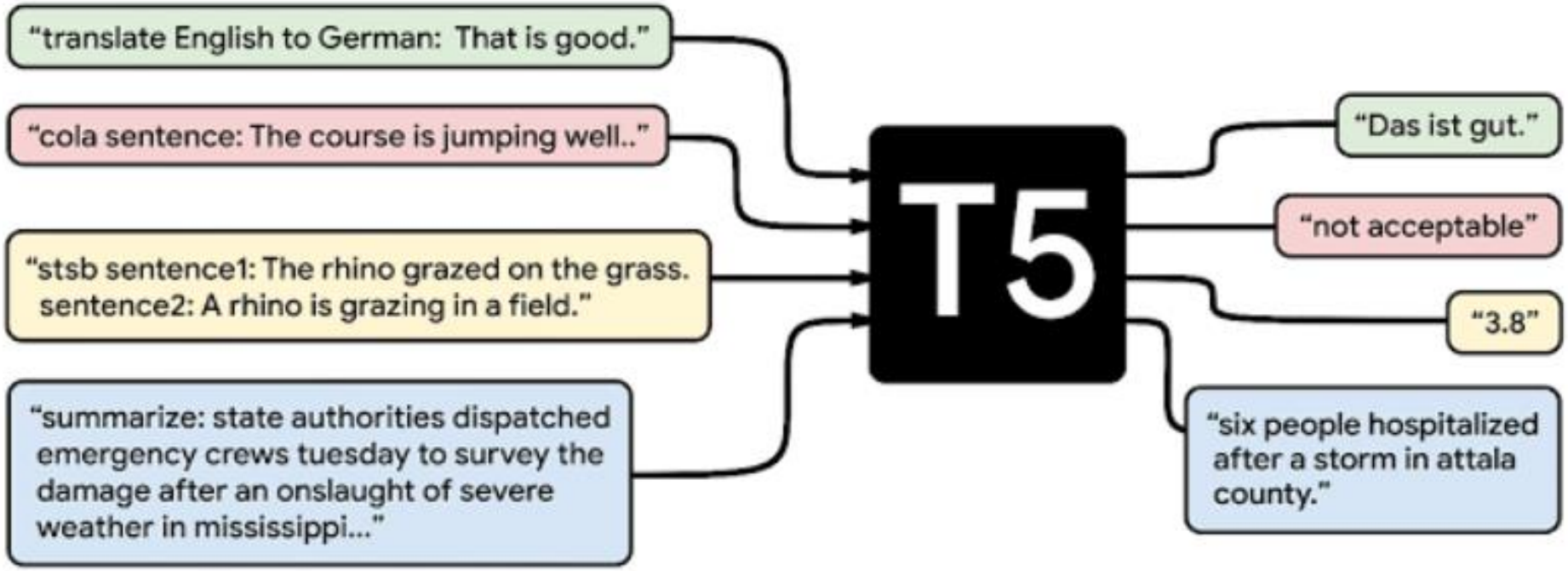
Model: Google T5 (Text-to-Text Transfer Transformer)

ترجمه

وظایف طبقه‌بندی

شباهت جمله

خلاصه کردن اسناد



Model: Google T5 (Text-to-Text Transfer Transformer) نسخه ها و سایزهای

- T5 comes in different sizes:
  - **t5-small**
  - **t5-base**
  - **t5-large**
  - **t5-3b**
  - **t5-11b.**

- ✓ T5v1.1 نسخه بهبود یافته T5با برخی تغییرات معماری است و فقط در C4بدون اختلاط در وظایف نظارت شده از قبل آموزش داده شده است.
- ✓ mT5 یک مدل T5چند زبانه است. این از قبل در مجموعه mC4که شامل ۱۰۱ زبان است، آموزش دیده است
- ✓ byT5 یک مدل T5است که از قبل بر روی توالی بایت ها به جای توکن های زیرکلمه SentencePieceآموزش دیده است.

- ✓ مدل بزرگتر نتایج بهتری می دهد، اما همچنین به قدرت محاسباتی بیشتری نیاز دارد و زمان زیادی را برای آموزش نیاز دارد. اما این یک فرآیند یکباره است. هنگامی که یک مدل تولید پارافراسی با کیفیت خوب و آموزش داده شده بر روی یک مجموعه داده مناسب دارید، می توان از آن برای تقویت داده در چندین کار NLPاستفاده کرد.

The Model: BART

- ✓ از معماری استاندارد ترجمه seq2seq/machine با رمزگذار دو طرفه مانند BART و رمزگشای چپ به راست مانند GPTاستفاده می کند.
- ✓ BARTبه ویژه هنگامی که برای تولید متن به خوبی تنظیم شده باشد مؤثر است، اما برای کارهای درک مطلب نیز به خوبی کار می کند. این عملکرد RoBERTaرا با منابع آموزشی مشابه در GLUEو SQuADمطابقت می‌دهد، به نتایج پیشرفته‌ای جدید در طیف وسیعی از وظایف گفتگوی انتزاعی، پاسخ‌گویی به سؤال، و خلاصه‌سازی با حداکثر ۶ ROUGEدست می‌یابد.
- ✓ BART با دو عمل اصلی آموزش داده می‌شود:
- ✓ تخریب متن با تابع نویز دلخواه (corrupting text with an arbitrary noising function)
- ✓ یادگیری مدلی برای بازسازی متن اصلی (learning a model to reconstruct the original text)
- ✓ BARTاز یک معماری ترانسفورماتور استاندارد رمزگذار–رمزگشایی مانند مدل اصلی ترانسفورماتور که برای ترجمه ماشین عصبی استفاده می شود، بهره می‌برد. اما ترکیبی از برخی تغییرات BART و GPT می‌باشد و در واقع از مدل رمزگذار BART و مدل رمزگشایی GPTاستفاده می‌کند.

The Model: BART

مدل‌های مختلف:BART

BARTweet

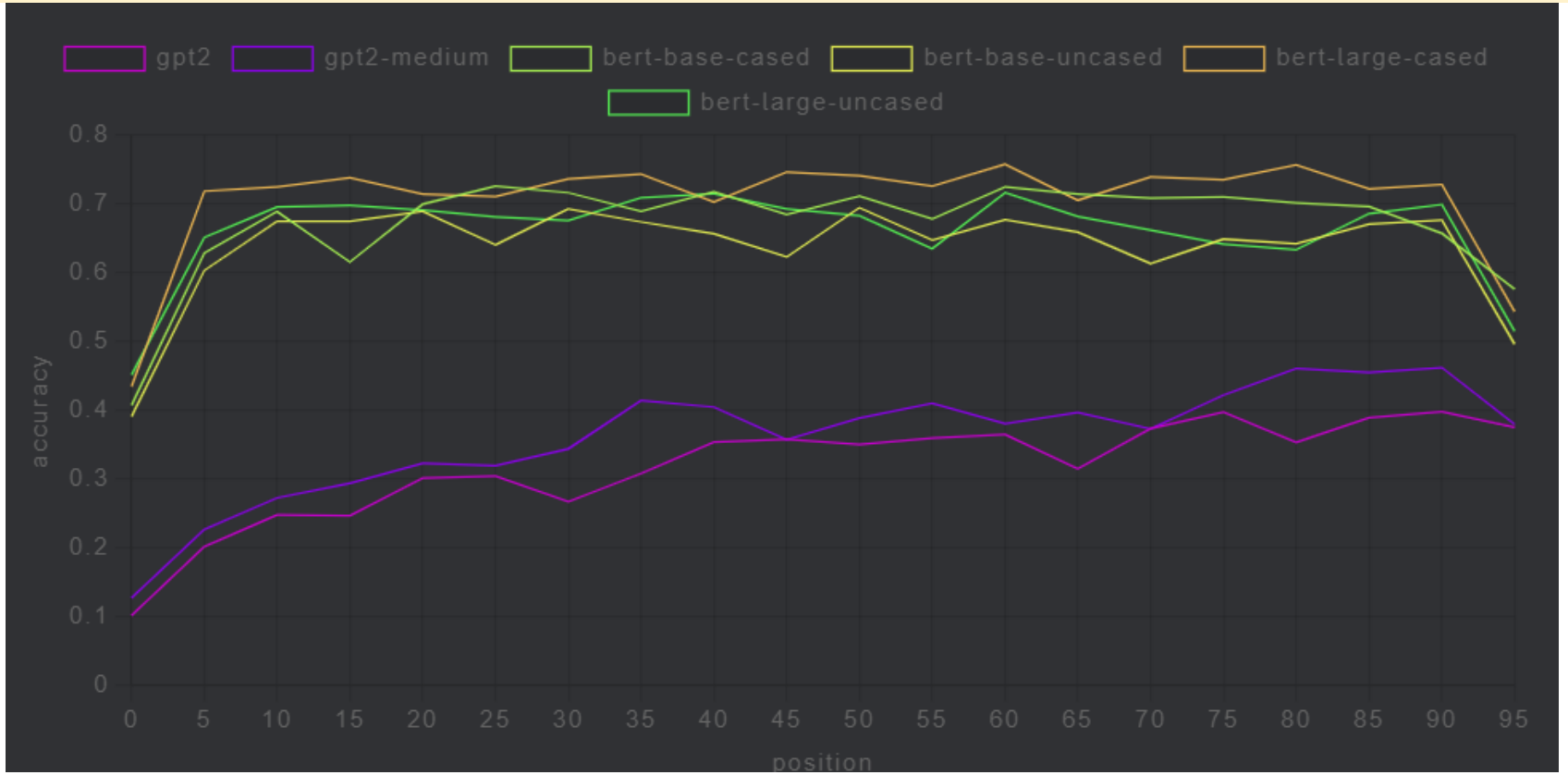
مدل BARTweet در BARTweet پیشنهاد شد: یک مدل زبان از پیش آموزش‌دیده برای توییت‌های انگلیسی است

BartGeneration

مدل BartGeneration یک مدل BARTاست که می تواند برای کارهای ترتیب به دنباله با استفاده کرد

مقایسه دو مدل BART VS GPT2

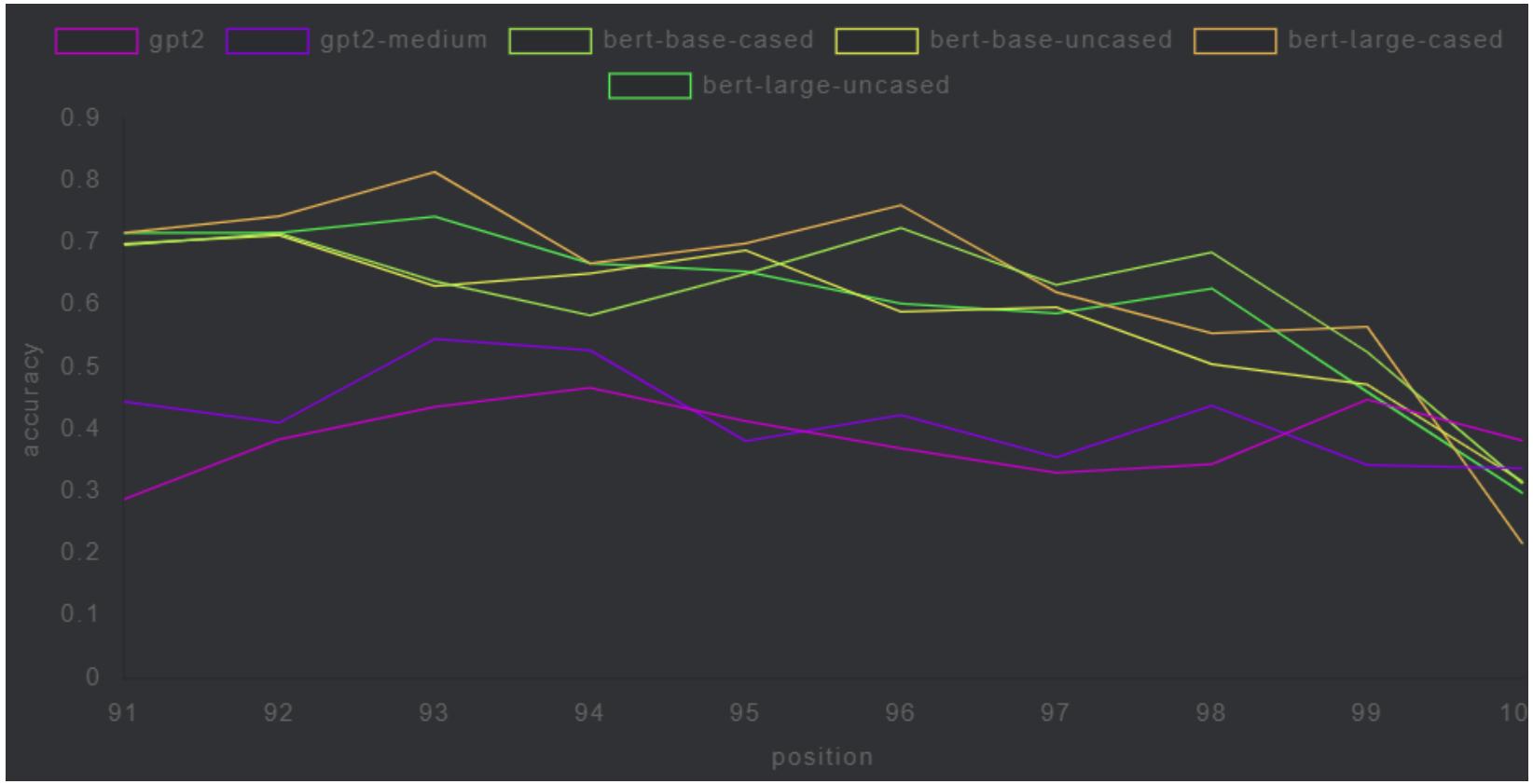
در شکل زیر می توانیم عملکرد هر ۶ مدل را روی این وظایف مشاهده کنیم. داده ها به گروه های ۵ positions به طور همزمان (یعنی موقعیت های ۰–۴، ۵–۹ و غیره) هموار شده اند. می‌توانید ببینید که عملکرد GPT-2 همچنان به افزایش خود ادامه می‌دهد، در حالی که مدل‌های BART پس از ارائه حدود ۵ نشانه زمینه نسبتاً پایدار هستند. جالب توجه است که عملکرد BART نسبت به موقعیت های ۵–۱۰ توکن اخیر به شدت کاهش می یابد.



https://lukesalamone.github.io/posts/bert-vs-gpt2/#:~:text=BERT%20and%20GPT%2D2%20perform,improve%20as%20context%20length%20increases.

مقایسه دو مدل BART VS GPT2

وقتی روی ۱۰ موقعیت نهایی زوم می کنیم، خواهیم دید که هر دو نوع GPT-2 در واقع همه مدل های BART را در نهایت شکست می دهند.

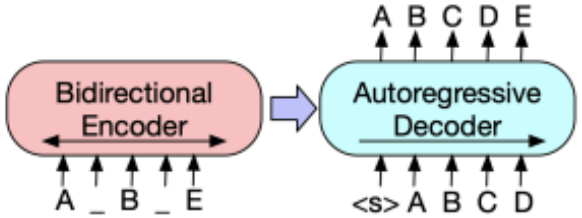


مقایسه دو مدل BART VS GPT2

BART و GPT-2 بسته به موقعیت توکن پیش‌بینی‌شده، در کار پیش‌بینی توکن کاملاً متفاوت عمل می‌کنند. برای طول توالی ثابت ۱۰۰ توکن، BERT زمانی بهترین عملکرد را دارد که توکن ماسک‌دار بین موقعیت‌های ۵ و ۹۵ باشد، در حالی که GPT-2 به طور مداوم با افزایش طول زمینه بهبود می‌یابد. جالب اینجاست که وقتی نشانه نهایی در دنباله قرار است پیش بینی شود، عملکرد BERT به شدت کاهش می یابد، در حالی که عملکرد GPT-2 ثابت می ماند.

BART

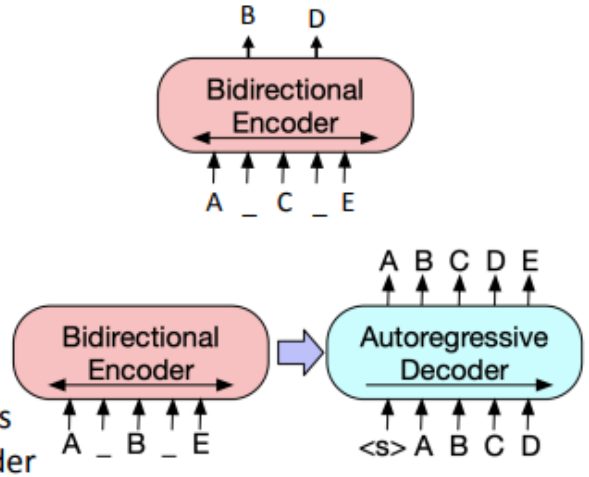
- BERT is good for "analysis" tasks, GPT is a good language model
- What to do for seq2seq tasks?
- Sequence-to-sequence BERT variant: permute/make/delete tokens, then predict full sequence autoregressively
- Uses the transformer encoder-decoder we discussed for MT (decoder attends to encoder)



Lewis et al. (2019)

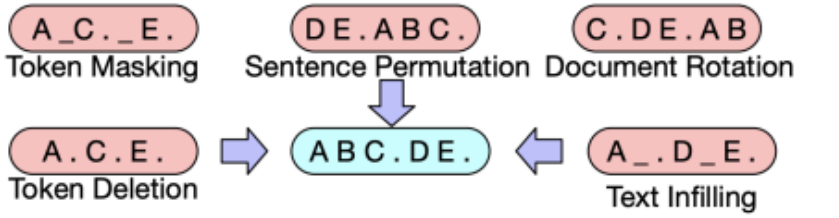
BERT vs. BART

- BERT: only parameters are an encoder, trained with masked language modeling objective
  - No way to do translation or left-to-right language modeling tasks
- BART: both an encoder and a decoder
  - Typically used for enc-dec tasks but also can just use the encoder as a replacement for BERT



Lewis et al. (2019)

BART



Infilling is longer spans than masking

- They try several strategies for generating training data. Infilling is a particularly helpful strategy for better performance

Lewis et al. (2019)

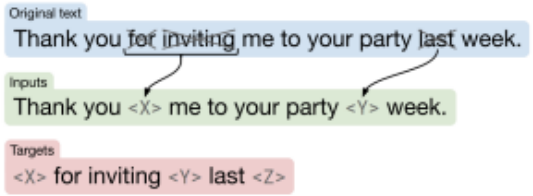
BART for Summarization

- **Pre-train** on the BART task: take random chunks of text, noise them according to the schemes described, and try to "decode" the clean text
- **Fine-tune** on a summarization dataset: a news article is the input and a summary of that article is the output (usually 1-3 sentences depending on the dataset)
- Can achieve good results even with **few summaries to fine-tune on**, compared to basic seq2seq models which require 100k+ examples to do well

Lewis et al. (2019)

T5

- Pre-training: similar denoising scheme to BART (they were released within a week of each other in fall 2019)
- Input: text with gaps. Output: a series of phrases to fill those gaps.



Raffel et al. (2019)

T5

Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full dataset	0	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
2^{29}	64	**82.87**	**19.19**	**80.97**	**72.03**	**26.83**	**39.74**	**27.63**
2^{27}	256	82.62	**19.20**	79.78	69.97	**27.02**	**39.71**	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

summarization — machine translation

- Colossal Cleaned Common Crawl: 750 GB of text
- We still haven't hit the limit of bigger data being useful for pre-training: here we see stronger MT results from the biggest data

Raffel et al. (2019)

Successes of T5

- How can we handle a task like QA by framing it as a seq2seq problem?
- Need to have text input and text output

Dataset	SQuAD 1.1
Input	At what speed did the turbine operate? \n (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Output	16,000 rpm

- Format: *Question \n Passage* —> *Answer*
  encoder decoder

Raffel et al. (2019)

OpenAI GPT/GPT2

- Very large language models using the Transformer architecture
- Straightforward left-to-right language model, trained on raw text
- GPT2: trained on 40GB of text collected from upvoted links from reddit
- 1.5B parameters — by far the largest of these models trained when it came out in March 2019

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

- Because it's a language model, we can **generate** from it

Radford et al. (2019)

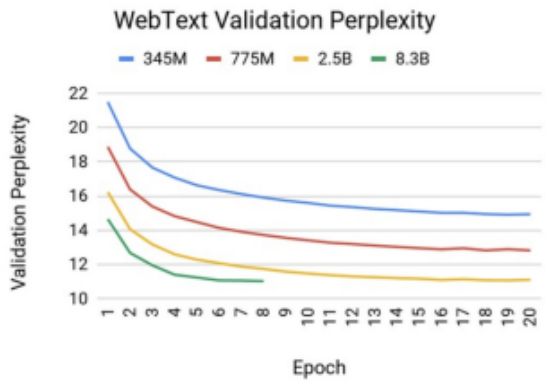
Pre-Training Cost (with Google/AWS)

- BERT: Base $500, Large $7000
- GPT-2 (as reported in other work): $25,000
- This is for a single pre-training run…developing new pre-training techniques may require many runs
- *Fine-tuning* these models can typically be done with a single GPU (but may take 1-3 days for medium-sized datasets)

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Pushing the Limits

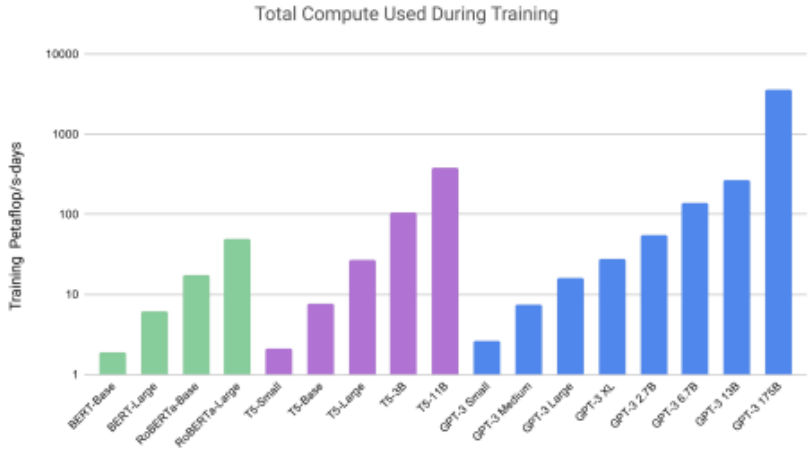
- NVIDIA: trained 8.3B parameter GPT model (5.6x the size of GPT-2)
- Arguable these models are still underfit: larger models still get better held-out perplexities



NVIDIA blog (Narasimhan, August 2019)

Pushing the Limits: GPT-3

- 175B parameter model: 96 layers, 96 heads, 12k-dim vectors
- Trained on Microsoft Azure, estimated to cost roughly $10M



Brown et al. (2020)

Pre-GPT-3: Fine-tuning

- Fine-tuning: this is the "normal way" of doing learning in models like GPT-2
- Requires computing the gradient and applying a parameter update on every example
- **This is super expensive with 175B parameters**

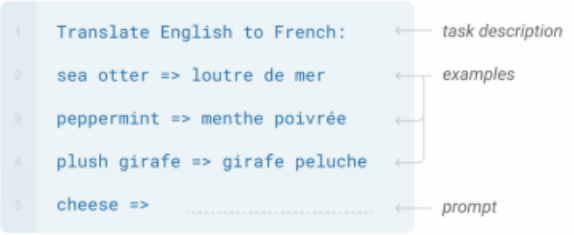
Brown et al. (2020)

GPT-3: Few-shot Learning

- GPT-3 proposes an alternative: **in-context learning.** Just uses the off-the-shelf model, no gradient updates
- This procedure depends heavily on the examples you pick as well as the prompt ("*Translate English to French*")

Few-shot

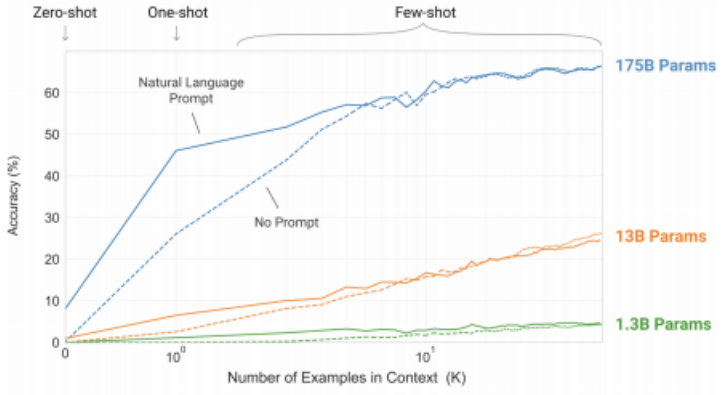
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Brown et al. (2020)

GPT-3

- **Key observation:** few-shot learning only works with the very largest models!



Brown et al. (2020)

GPT-3

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

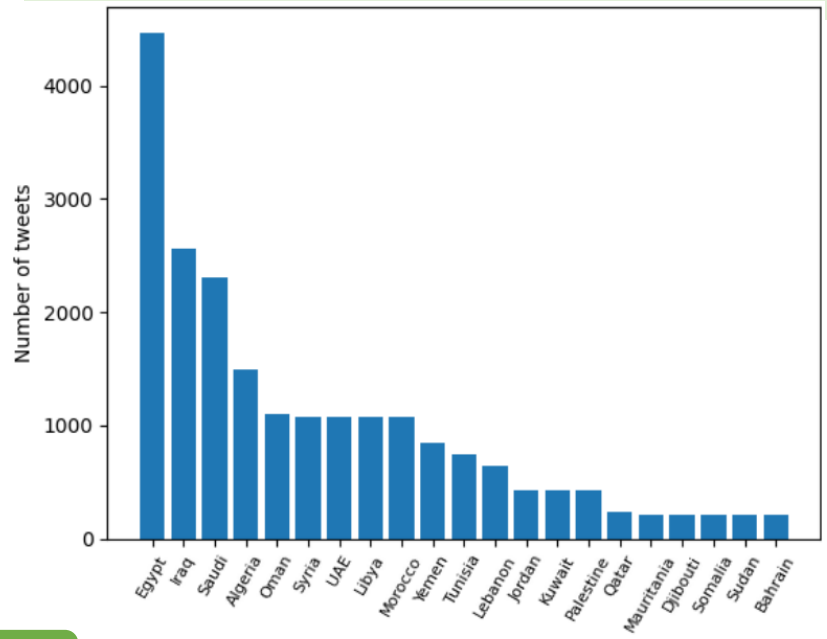
- Sometimes very impressive, (MultiRC, ReCoRD), sometimes very bad
- Results on other datasets are equally mixed — but still strong for a few-shot model!

Brown et al. (2020)

در این بخش به بررسی مقاله ای در مورد "شناسایی گویش عربی با استفاده از BERT Fine-Tuning" می پردازیم:

- **Data :**

✓ مجموعه داده های آموزشی ارائه شده شامل۲۱۰۰۰ توییت با ۲۱ برچسب مختلف به نمایندگی از ۲۱ زبان(مصر، بحرین، عراق، عربستان سعودی، الجزایر، عمان، سوریه، امارات متحده عربی، لیبی، مراکش، یمن، تونس، لبنان، اردن، کویت، فلسطین، قطر، موریتانی، جیبوتی، سومالی، سودان).



✓ ۴۹۵۷ توییت با برچسب برای توسعه، ۵۰۰۰ برای آزمایش . ما داده های آموزشی را به ۱۹۹۵۰ توییت برای آموزش و ۱۰۵۰ توییت برای اعتبارسنجی تقسیم کردیم تا بتوانیم از بیشتر داده های داده شده استفاده کنیم.

✓ یک تابع پیش پردازش ساده برای حذف پیوندها، علائم نگارشی و کاراکترهای تکراری استفاده شد. ماهمچنین سعی کردم دایرکت ها را حذف کنم اما عملکرد بدون حذف آنها کمی بهتر بود.

✓ با نگاهی به توزیع مجموعه داده ها در شکل، می توانیم متوجه شویم که داده ها کاملاً نامتعادل هستند. به عنوان مثال، توییت های مصر ۲۱.۳ درصد از داده ها را نشان می دهد، اما بحرین ۰.۰۱ درصد از مجموعه داده ها را نشان می دهد

https://aclanthology.org/2020.wanlp-1.33.pdf 28

BERT System

تفاوت میان BERT و mBERT این است که BERT یک مدل زبان انگلیسی از قبل آموزش دیده است، اما mBERT وزن های از پیش آموزش دیده را برای عربی و ۱۰۳ زبان دیگر ارائه می دهد.
در این مقاله از mBERT استفاده شده است.
ما از یک مدل از پیش آموزش دیده mBERT استفاده کرده بودیم و حدود ۷.۲ میلیارد آن از قبل آموزش داده شده بود.
روند آموزش ما به دو مرحله تقسیم شد: **tuning the language model** تنظیم مدل زبان با استفاده از ۱۰ میلیون توییت و سپس با استفاده از ۱۹۹۵۰ توییت برای کار طبقه بندی **classification task** تنظیم کنید.

Language Model Training:

ما مدل را برای یک دوره در تمام توییت‌های بدون برچسب آموزش داده‌ایم.

در ابتدا، ما توییت های طولانی را کوتاه کرده ایم. توییت‌ها و موارد کوتاه‌تر را اضافه کرد تا طول ۱۰ کلمه برای همه توییت‌ها به دست آید.

ما مدل را در ۱۰ جلسه آموزش داده ایم و در هر جلسه با ۱ میلیون توییت آموزش داده شده است.

برای یک دوره زمان تمرین برای یک جلسه حدود ۱۲ ساعت طول می کشد زیرا بهینه ساز مورد استفاده برای تمرین، الگوریتم Adam با کاهش وزن (AdamW با نرخ یادگیری=۲e-5، epsilon=1e-5، کاهش وزن=۰.۰۰۱ و بتا (۰.۹،۰.۹۹۹) است.). تمام زمان آموزش حدود ۱۲۰ ساعت در حال اجرا در مجازی Azure بود

دستگاه با پردازنده گرافیکی Tesla M60 GPU

Classification Training:

همانطور که در جدول 1 توضیح داده شده است، باید یک نرمال سازی دسته ای، dropout و نگاشت لایه خطی به کلاس ها ما از همان optimizer, Adamاستفاده کردیم.

ما مدل را برای 4 epoch آموزش دادیم. در دوره‌های مختلف، حدود 1 ساعت آموزش روی GPU GeForce RTX 2060طول می‌کشد.

language model	classification
BERT Base model	BERT language model
Dropout layer with 0.1 drop_prop	Batch normalization with tanh activation function
linear layer	Dropout layer with 0.1 drop_prop
	Linear layer

Table 1: Modified BERT Model architecture

Traditional Machine Learning

استخراج ویژگی»» : از ترکیبی از کاراکترهای n-gramو کلمات unigram TF-IDFاستفاده کرده ایم:

- TF-IDF for unigram with max df=0.05,min df=0.0001 after removing stop words contained in natural language processing toolkit library in python(NLTK)
- • (2, 9) character n-grams with respect to boundaries, using sublinear transformation and maximum number of feature =40,000 (twice)

Conclusion

توییت ها آزمایش‌های ما نشان می‌دهد که مدل BERT بهترین امتیاز F1 را کسب کرده است

مدل های یادگیری ماشینی استفاده از توییت های بدون برچسب برای تنظیم دقیق مدل زبان یک راه عالی است

Figure 2: Confusion matrix of development set predictions

Models	Features		F1 score
	char_wb	word	
Non-linear SVM	(2,9)		17.1
Non-linear SVM		unigram	13.1
Non-linear SVM	(2,9)	unigram	17.2
Linear SVM	(2,9)	unigram	16.3
Bernolli NB	(2,9)	unigram	16.0
Voting classifier	(2,9)	unigram	**18.1**
	Deep Learning		
BERT	Without fine-tuning of language model		**23.5**
BERT	With fine-tuning of language model		**24.05**

Model		ویژگی	معماری
GANs	ترکیبی از دو مدل از پیش آموزش‌دیده، BERT و GPT-2	بیشتر برای داده‌های تصویری استفاده می‌شوند	CNNمبتنی بر GANمعماری
PEGASUS		self-supervision یادگیری خودنظارتی	seq2seq
Google T5	۱۱ میلیارد پارامتر	ترجمه ماشینی، خلاصه‌سازی اسناد، پاسخ به سوال، و وظایف طبقه‌بندی	Text-to-Text Transformer
BERT			seq2seq/machine معماری ترانسفورماتور استاندارد ترکیبی از برخی تغییرات GPTو BERT

جمع‌بندی

در مدل انگلیسی همان طور که fine tune نیز انجام شده است و براساس ارزیابی های مدل اصلی GPT2 ،مدل قابل قبول GPT2-128 است.

در مدل عربی نیز دو مدل GPT2 و GPT3(noor) مورد بررسی قرار گرفت .اگر چه این مدل به صورت پایلوت در پژوهشگاه ران نشده است، اما براساس ارزیابی ها مدل GPT2 عربی می‌تواند قابل قبول باشد .مدل noorنیز ساخته یک پژوشگاه عربی مختص زبان عربی و برگرفته ازGPT3 است.

در مدل فارسی با بررسی های انجام شده در تنها مدل آماده GPT2 به نام بلبل زبان، احتمال این می رود که با دردست داشتن دیتاست بالا بتوان ریترین این مدل را نجام داد.دیگرمدل ها از جمله GAN نیز مورد بررسی قرار گرفت که از انجا که پایه این مدل CNNاست و سه کاربر اصلی آن تولید Image-to-image translationText-to-image synthesisVideo generation:این مدل پیشنهاد نمی‌شود.

کارهای آینده

- ✓ پیاده سازی تولید متن مبتنی بر احساس در زبان انگلیسی با استفاده از مدل GPT-2 SMALL
- ✓ پیاده سازی تولید متن مبتنی بر احساس در زبان عربی با استفاده از مدل GPT-2 SMALL
- ✓ آموزش دوباره مدل GPT2 فارسی
- ✓ تولید متن بر اساس انتخاب موضوع(سیاسی–اجتماعی–...)