Nlp in 2020

Tomasz Stanisławek[1,2]
(Thomas Wolf and Sebastian Ruder)

[1]Applica.ai

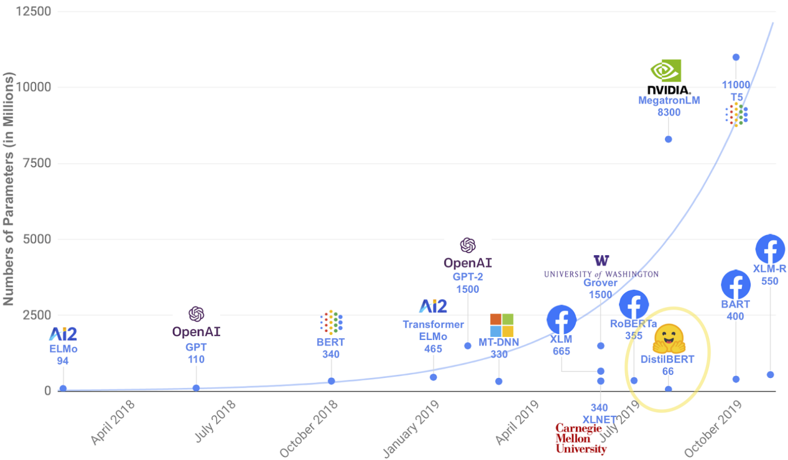
[2]Faculty of Mathematics and Information Science, Warsaw University of Technology

Mi2 DataLab seminar, 26.10.2020

Presentation plan

- Big models
- Efficiency
- Multilingual models
- Multimodal models
- Datasets and evaluation
- Honorable mentions
- Industry

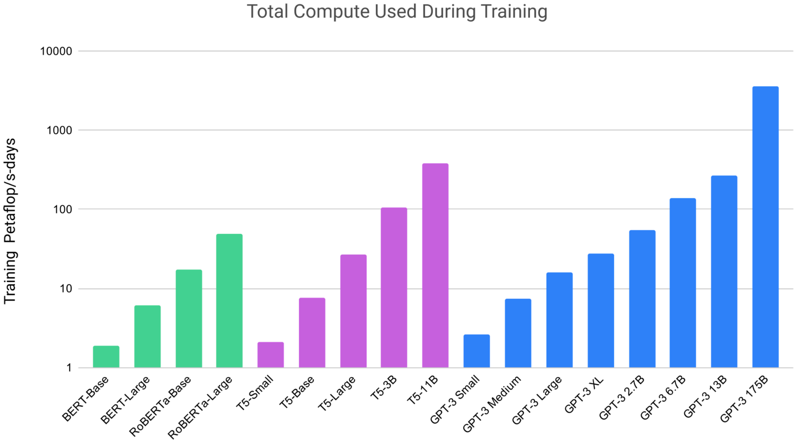
Big models



Numbers of Parameters (in Millions)
12500
10000
7500
5000
2500
0
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
Ai2
ELMo
94
OpenAI
GPT
110
BERT
340
Ai2
Transformer
ELMo
465
OpenAI
GPT-2
1500
MT-DNN
330
XLM
665
W
UNIVERSITY of WASHINGTON
Grover
1500
340
XLNET
Carnegie
Mellon
University
RoBERTa
355
DistilBERT
66
NVIDIA.
MegatronLM
8300
11000
T5
BART
400
XLM-R
550

Big models

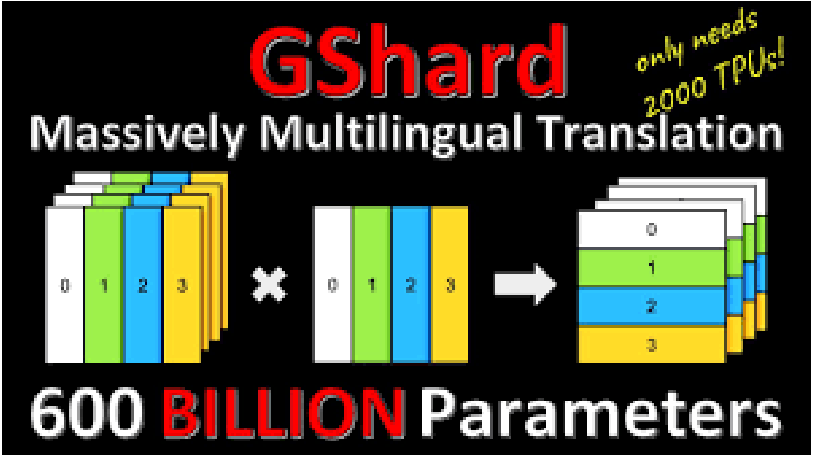
Language Models are Few-Shot Learners,
https://arxiv.org/pdf/2005.14165.pdf



"About 1 500 000 results" for "gpt-3" query (Google, 25.10.2020)

Big models

GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding, https://arxiv.org/abs/2006.16668



Model can efficiently be trained on 2048 TPU v3 in 4 days

Big models

https://twitter.com/fchollet/status/1122330598968705025

François Chollet
@fchollet

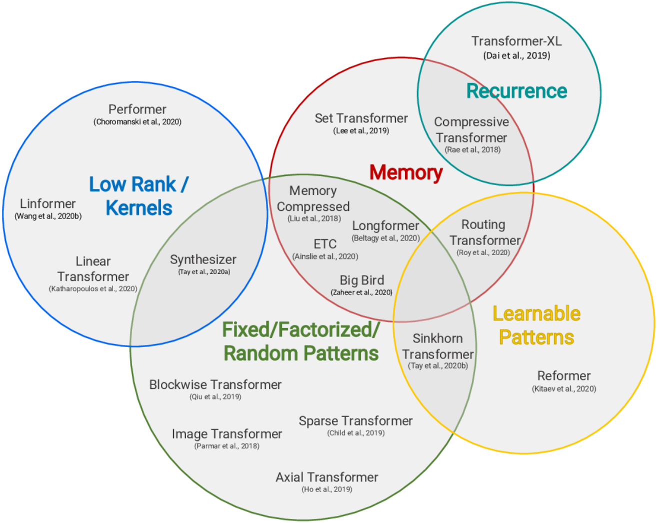
Training ever bigger convnets and LSTMs on ever bigger datasets gets us closer to Strong AI -- in the same sense that building taller towers gets us closer to the moon.

4:44 AM · Apr 28, 2019 · Twitter for Android

602 Retweets **42** Quote Tweets **2.3K** Likes

Efficient Transformers

Efficient Transformers: A Survey, https://arxiv.org/pdf/2009.06732.pdf



Efficiency-flavored "X-former" models

Efficient Transformers

Model size and Computational efficiency

Reducing the size of a pretrained model

Three main **techniques** currently investigated:

- ❑ **Distillation**

  $$L = -\sum_i t_i * log(s_i)$$

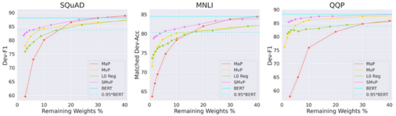
  With **t** the logits from the teacher and **s** the logits of the student

  DistilBert: 95% of Bert performances in a model 40% smaller and 60% faster

- ❑ **Pruning**

  **Movement Pruning:
  Adaptive Sparsity by Fine-Tuning**

  Victor Sanh[1], Thomas Wolf[1], Alexander M. Rush[1,2]
  [1]Hugging Face, [2]Cornell University
  {victor,thomas}@huggingface.co ; arush@cornell.edu



- ❑ **Quantization**

  From FP32 to INT8

$$Q(x, \text{scale}, \text{zero\_point}) = \text{round}\left(\frac{x}{\text{scale}} + \text{zero\_point}\right)$$

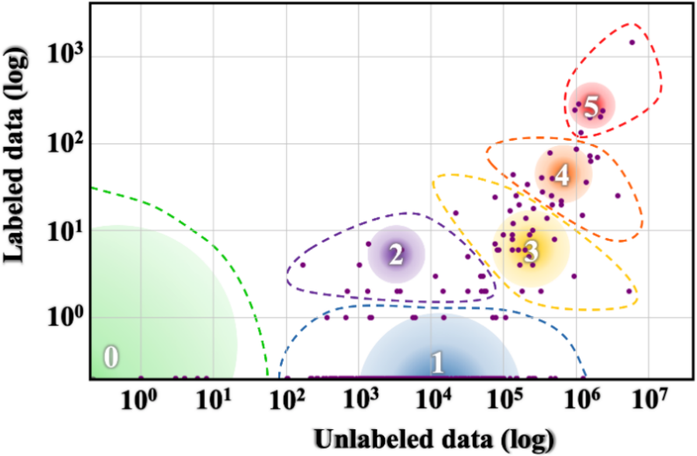
Prec	F1 score	Model Size	1 thread	4 threads
FP32	0.9019	438 MB	160 sec	85 sec
INT8	0.8953	181 MB	90 sec	46 sec

17

https://www.youtube.com/watch?v=8Hg2UtQg6G4

Multilingual models

The State and Fate of Linguistic Diversity and Inclusion in the NLP World, https://arxiv.org/pdf/2004.09095.pdf



The size and colour of a circle represent the number of languages and speakers respectively in each category. Colours (on the VIBGYOR spectrum) Violet–Indigo–Blue–Green–Yellow–Orange–Red) represent the total speaker population size from low (violet) to high (red).

Multilingual models

The State and Fate of Linguistic Diversity and Inclusion in the NLP World, https://arxiv.org/pdf/2004.09095.pdf

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.

More info: https://ruder.io/nlp-beyond-english/

Multilingual models

Unsupervised Cross-lingual Representation Learning at Scale, https://arxiv.org/pdf/1911.02116.pdf

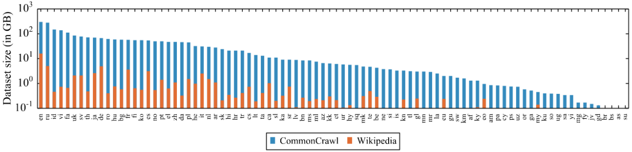


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

Model	train	#M	en	nl	es	de	Avg
Lample et al. (2016)	each	N	90.74	81.74	85.75	78.76	84.25
Akbik et al. (2018)	each	N	**93.18**	90.44	-	**88.27**	-
mBERT†	each	N	91.97	90.94	87.38	82.82	88.28
	en	1	91.97	77.57	74.96	69.56	78.52
XLM-R$_{Base}$	each	N	91.95	91.21	88.46	83.65	88.82
	en	1	91.95	77.83	76.24	69.70	78.93
	all	1	91.84	88.13	87.02	82.76	87.44
XLM-R	each	N	92.74	**93.25**	**89.04**	85.53	90.14
	en	1	92.74	81.00	76.44	72.27	80.61
	all	1	93.03	90.41	87.83	85.46	89.18

Table 2: **Results on named entity recognition** on CoNLL-2002 and CoNLL-2003 (F1 score). Results with † are from Wu and Dredze (2019). Note that mBERT and *XLM-R* do not use a linear-chain CRF, as opposed to Akbik et al. (2018) and Lample and Conneau (2019).

Multilingual models

mT5: A massively multilingual pre-trainedtext-to-text transformer, https://arxiv.org/pdf/2010.11934.pdf

Model	Pair sentence		Question answering		
	XNLI	PAWS-X	XQuAD	MLQA	TyDi QA-GoldP
Metrics	Acc.	Acc.	F1 / EM	F1 / EM	F1 / EM
Cross-lingual zero-shot transfer (models are trained on English data only)					
mBERT	65.4	81.9	64.5 / 49.4	61.4 / 44.2	59.7 / 43.9
XLM	69.1	80.9	59.8 / 44.3	48.5 / 32.6	43.6 / 29.1
InfoXLM	81.4	-	- / -	73.6 / 55.2	- / -
Phang et al. (2020)	80.4	87.7	77.2 / 61.3	72.3 / 53.5	76.0 / 59.5
XLM-R	79.2	86.4	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0
mT5-Small	67.5	82.4	58.1 / 42.5	54.6 / 37.1	34.9 / 23.9
mT5-Base	75.4	87.4	67.0 / 49.0	64.6 / 45.0	58.1 / 42.8
mT5-Large	81.1	89.6	77.8 / 61.5	71.2 / 51.7	57.8 / 41.1
mT5-XL	82.9	**90.2**	79.5 / 63.6	73.5 / 54.5	77.3 / 61.5
mT5-XXL (75% trained)	**84.8**	89.2	**81.9 / 65.7**	**75.5 / 56.9**	**80.8 / 66.3**
Translate-train (models are trained on English data plus translations in all target languages)					
XLM-R	82.6	90.4	80.2 / 65.9	72.8 / 54.3	66.5 / 47.7
Filter + Self-Teaching	83.9	91.4	82.4 / 68.0	76.2 / **57.7**	68.3 / 50.9
mT5-Small	64.7	87.8	64.3 / 49.5	60.2 / 41.1	48.2 / 34.0
mT5-Base	75.9	90.2	75.3 / 59.7	68.6 / 49.1	64.0 / 47.7
mT5-Large	81.8	91.3	81.2 / 65.9	73.3 / 54.2	71.1 / 54.9
mT5-XL	84.8	91.3	82.7 / 68.1	74.6 / 55.2	79.9 / 65.3
mT5-XXL (75% trained)	**87.2**	**92.0**	**85.0 / 70.8**	**76.3** / 56.8	**82.0 / 67.9**

Table 2: Results on XTREME sentence-pair classification and question answering tasks. Apart from mT5 (ours), all metrics are from Fang et al. (2020). Note, InfoXLM benefits from parallel training data, while Phang et al. (2020) leverages additional labeled data from related tasks. For the “translate-train” setting, we include English training data, so as to be comparable with Fang et al. (2020). This differs from XTREME “translate-train” setup of Hu et al. (2020). Full results for all languages in all tasks are provided in tables 6 to 10 (appendix).

Multimodal models

Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modalPre-training, https://arxiv.org/pdf/1908.06066.pdf

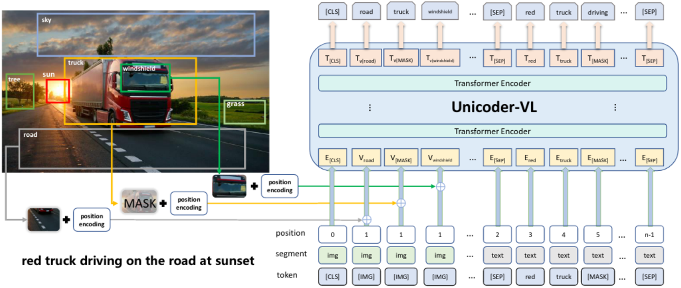
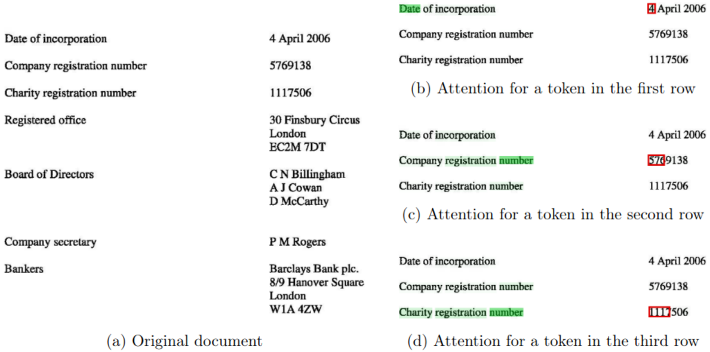


Figure 1: Illustration of Unicoder-VL in the context of an object and text masked token prediction, or *cloze*, task. Unicoder-VL contains multiple Transformer encoders which are used to learn visual and linguistic representation jointly.

Multimodal models

LAMBERT: Layout-Aware (Language) Modeling using BERTfor information extraction, https://arxiv.org/pdf/2002.08087.pdf



(a) Original document

(b) Attention for a token in the first row

(c) Attention for a token in the second row

(d) Attention for a token in the third row

Datasets and evaluation

UnifiedQA: Crossing Format Boundaries With a Single QA System, https://arxiv.org/abs/2005.00700

Extractive [SQuAD]

Question: At what speed did the turbine operate?
Context: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Gold answer: 16,000 rpm

Abstractive [NarrativeQA]

Question: What does a drink from narcissus's spring cause the drinker to do?
Context: Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...
Gold answer: fall in love with themselves

Multiple-Choice [ARC-challenge]

Question: What does photosynthesis produce that helps plants grow?
Candidate Answers: (A) water (B) oxygen (C) protein (D) sugar
Gold answer: sugar

Yes/No [BoolQ]

Question: Was America the first country to have a president?
Context: (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
Gold answer: no

Figure 1: Four formats (color-coded throughout the paper) commonly used for posing questions and answering them: Extractive (EX), Abstractive (AB), Multiple-Choice (MC), and Yes/No (YN). Sample dataset names are shown in square brackets. We study generalization and transfer across these formats.

Datasets and evaluation

BERTs of a feather do not generalize together: Large variability ingeneralization across models with similar test set performance, https://arxiv.org/pdf/1911.02969.pdf

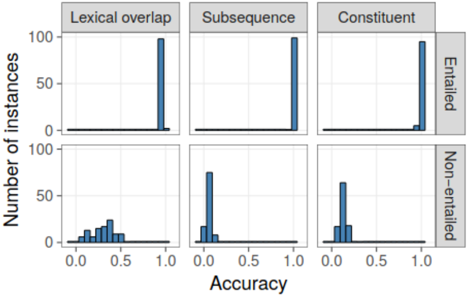
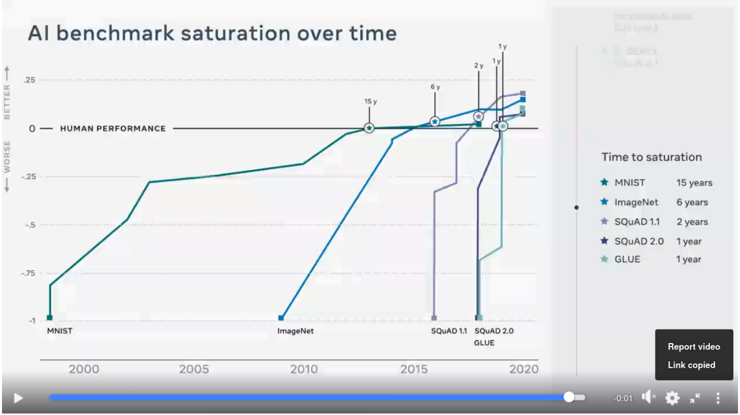


Figure 3: Out-of-distribution generalization: Performance on the HANS evaluation set, broken down into six categories of examples based on which syntactic heuristic each example targets and whether the correct label is *entailment* or *non-entailment*. The non-entailed lexical overlap cases (lower left plot) display a large degree of variability across instances.

Datasets and evaluation

https://dynabench.org



Honorable mentions

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, https://arxiv.org/abs/2004.10964

Domain	Task	RoBERTa	Additional Pretraining Phases: DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	**84.4**$_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	**87.8**$_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	**75.6**$_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	**81.3**$_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	**90.4**$_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	**94.6**$_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	**68.7**$_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	**95.6**$_{0.1}$

Table 5: Results on different phases of adaptive pretraining compared to the baseline RoBERTa (col. 1). Our approaches are DAPT (col. 2, §3), TAPT (col. 3, §4), and a combination of both (col. 4). Reported results follow the same format as Table 3. State-of-the-art results we can compare to: ChemProt (84.6), RCT (92.9), ACL-ARC (71.0), SciERC (81.8), HyperPartisan (94.8), AGNews (95.5), IMDB (96.2); references in §A.2.

Honorable mentions

- A Primer in BERTology: What we know about how BERT works, https://arxiv.org/pdf/2002.12327.pdf
- Explaining Deep Neural Networks, https://arxiv.org/pdf/2010.01496v1.pdf
- Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, https://arxiv.org/abs/1910.10683
- Meta-Learning in Neural Networks: A Survey, https://arxiv.org/abs/2004.05439
- Current Limitations of Language Models: What You Need is Retrieval, https://arxiv.org/abs/2009.06857v1

Industry

https://github.com/huggingface

Hugging Face

Solving NLP, one commit at a time!

NYC + Paris https://huggingface.co/ Verified

Repositories 36 Packages People 19 Projects

Pinned repositories

transformers
🤗Transformers: State-of-the-art Natural Language Processing for Pytorch and TensorFlow 2.0.
Python 35.6k 8.6k

tokenizers
💥Fast State-of-the-Art Tokenizers optimized for Research and Production
Rust 3.9k 273

datasets
🤗 Fast, efficient, open-access datasets and evaluation metrics for Natural Language Processing and more in PyTorch, TensorFlow, NumPy and Pandas
Python 4.4k 343

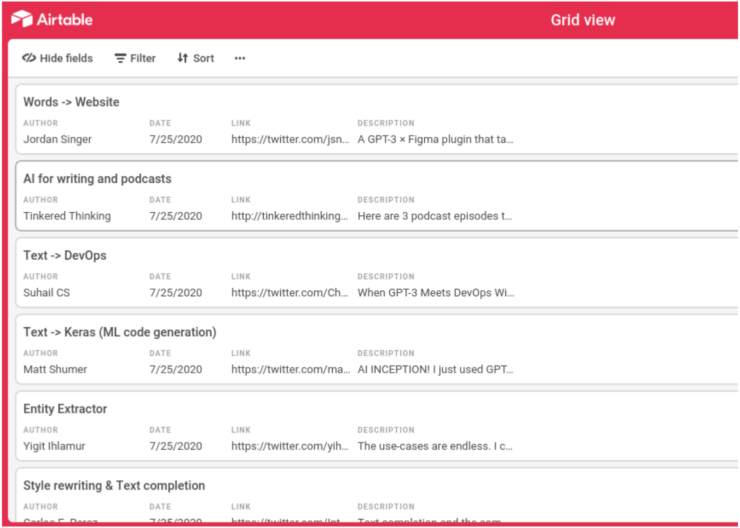
awesome-papers
Papers & presentation materials from Hugging Face's internal science day
1.6k 77

swift-coreml-transformers
Swift Core ML 3 implementations of GPT-2, DistilGPT-2, BERT, and DistilBERT for Question answering. Other Transformers coming soon!
Swift 963 109

knockknock
🚪✊Knock Knock: Get notified when your training ends with only two additional lines of code
Python 1.8k 158

Industry

https://gpt3examples.com/
https://medium.com/towards-artificial-intelligence/crazy-gpt-3-use-cases-232c22142044



Industry

COLING 2020 Industry Track Call for Papers

- Challenges of doing applied research at scale
- Noisy and/or unpredictable data (real world v. contrived)
- Negative results related to industry applications
- Analysis, modeling, and dataset construction under the constraint of respecting data privacy
- Algorithmic ethics and responsibility
- Evaluation methodologies, particularly for monitoring performance after deployment
- Trade-offs between resources (environmental and production) and performance; data size and modeling improvements
- Towards replicability in deep learning: experimental procedures necessary to develop successful models

Thank you!