German Language Model Training & Evaluation

•••

Electra and Beyond

About Us

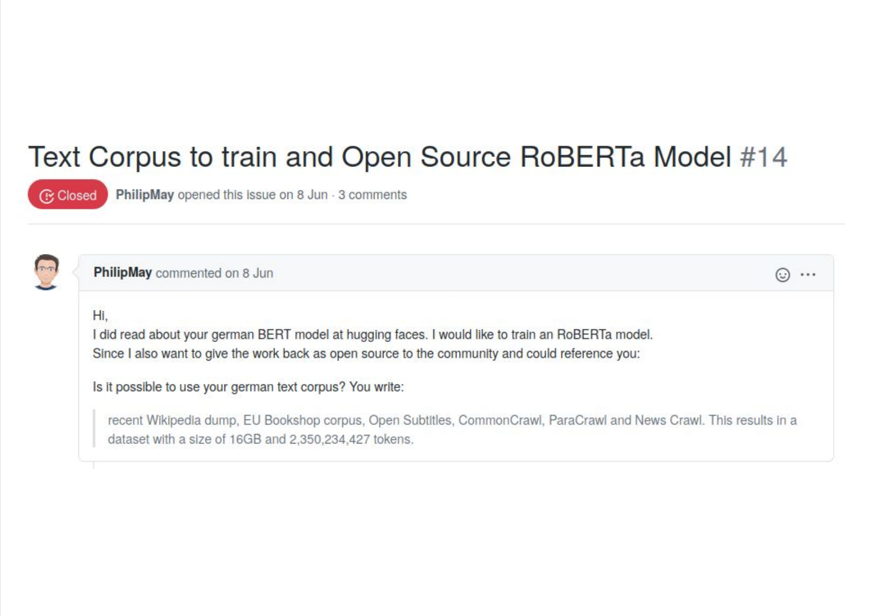
Philipp Reissel

ambeRoad

Philip May

T · · Systems ·

How it Started



Text Corpus to train and Open Source RoBERTa Model #14

Closed **PhilipMay** opened this issue on 8 Jun · 3 comments

PhilipMay commented on 8 Jun

Hi,
I did read about your german BERT model at hugging faces. I would like to train an RoBERTa model.
Since I also want to give the work back as open source to the community and could reference you:

Is it possible to use your german text corpus? You write:

> recent Wikipedia dump, EU Bookshop corpus, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl. This results in a dataset with a size of 16GB and 2,350,234,427 tokens.

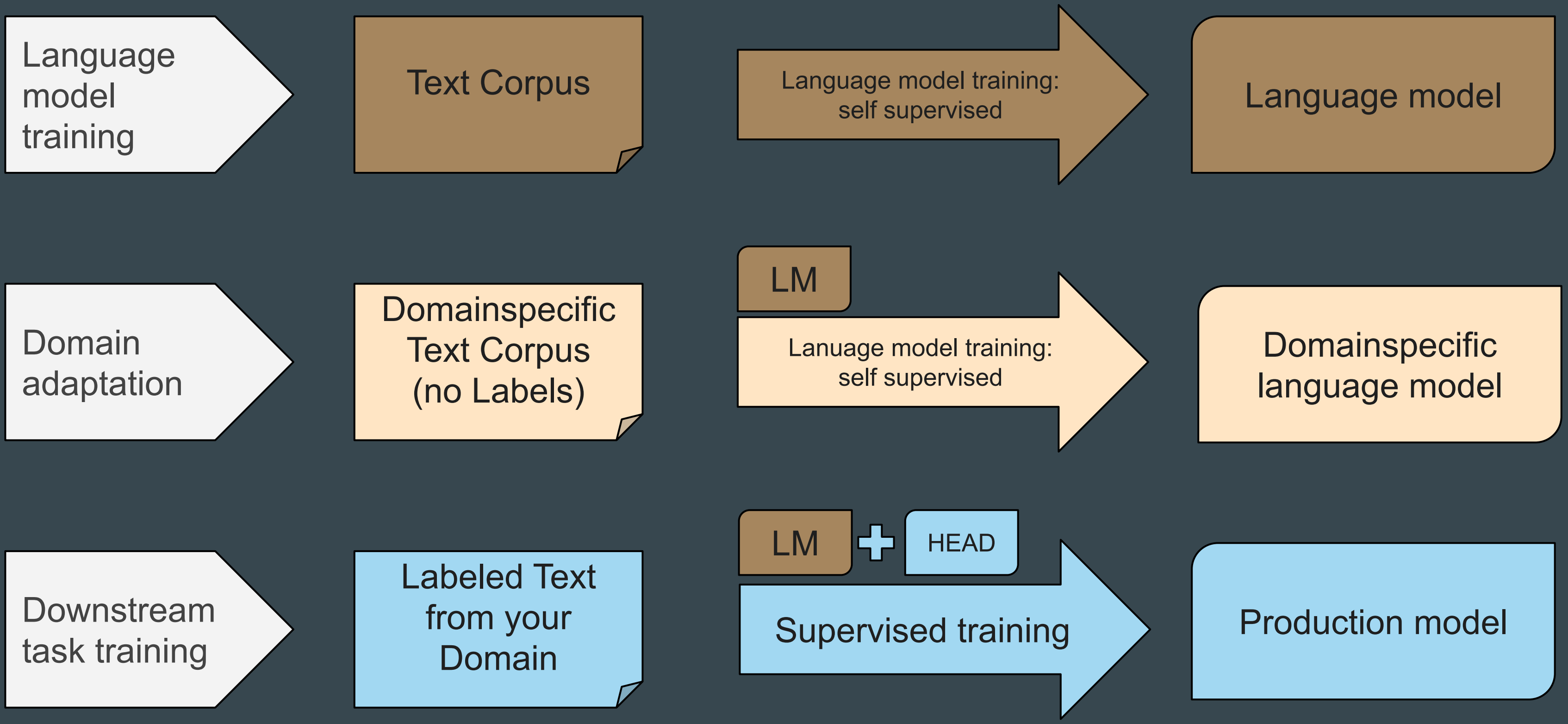
How it's going

Model card

German Electra Uncased

[1]

Language Model Training and Usage



Language model training

Text Corpus

Language model training: self supervised

Language model

Domain adaptation

Domainspecific Text Corpus (no Labels)

LM

Lanuage model training: self supervised

Domainspecific language model

Downstream task training

Labeled Text from your Domain

LM

HEAD

Supervised training

Production model

How hard can it be?

German Data Sources

Open Legal Data

Library Genesis^{2M}

German Vocab Specialities: “**Donaudampfschiffahrtsgesellschaftskapitän**”

=> donau ##dampf ##schiff ##ahrt ##sg ##es ##el ##ls ##cha ##ft ##ska ##pit ##än

Computational Resources (Pricing)

Ingredients to train a Language Model

SoMaJo

1	Text corpus	● one sentence per line ● blank line between documents	● large ● cover your domain
2	Vocabulary	● all tokens ● special tokens	● generated from text corpus
3	Tokenizer config	● case ● accent handling	● max length
4	Model config / architecture	● attention heads ● embedding size	● hidden layers ● etc.
5	Training config	● batch size ● learning rate	● train steps ● etc.

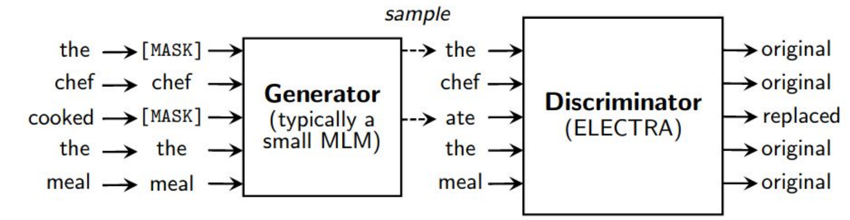
Where to build on?

BERT

- The Basis
- (Whole) Word Masking
- Next Sentence Prediction (NSP)

ELECTRA

- BERT + GAN like Structure
- More Efficient
- No NSP

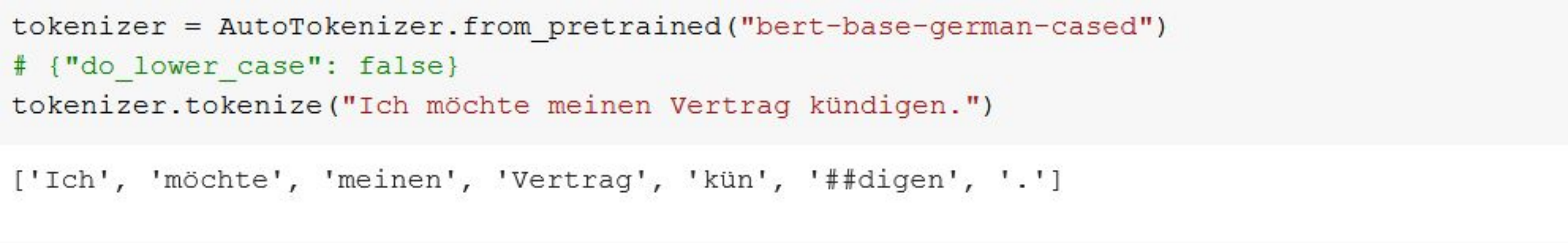


RoBERTa / XLM-RoBERTa

- BERT + Lots of GPUs
- No Next Sentence Prediction
- Built for Multilinguality on Purpose

The Tokenizer - What we have done differently

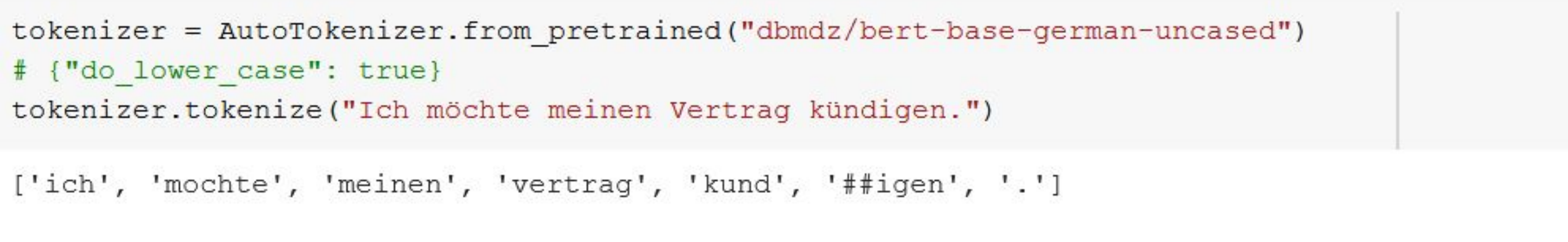
1



```
tokenizer = AutoTokenizer.from_pretrained("bert-base-german-cased")
# {"do_lower_case": false}
tokenizer.tokenize("Ich möchte meinen Vertrag kündigen.")
```

```
['Ich', 'möchte', 'meinen', 'Vertrag', 'kün', '##digen', '.']
```

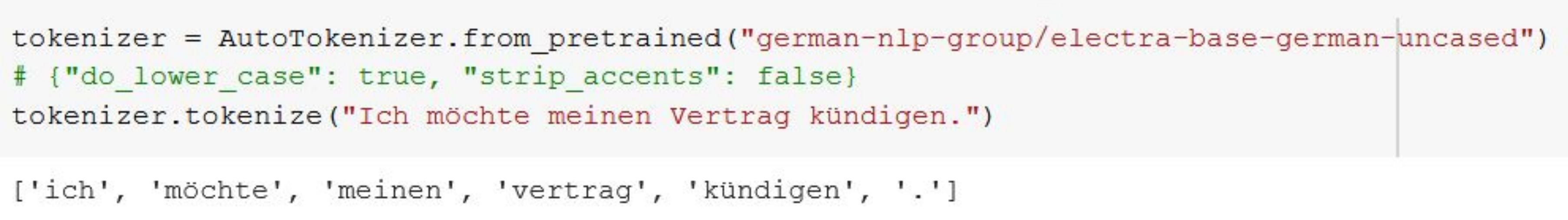
2



```
tokenizer = AutoTokenizer.from_pretrained("dbmdz/bert-base-german-uncased")
# {"do_lower_case": true}
tokenizer.tokenize("Ich möchte meinen Vertrag kündigen.")
```

```
['ich', 'mochte', 'meinen', 'vertrag', 'kund', '##igen', '.']
```

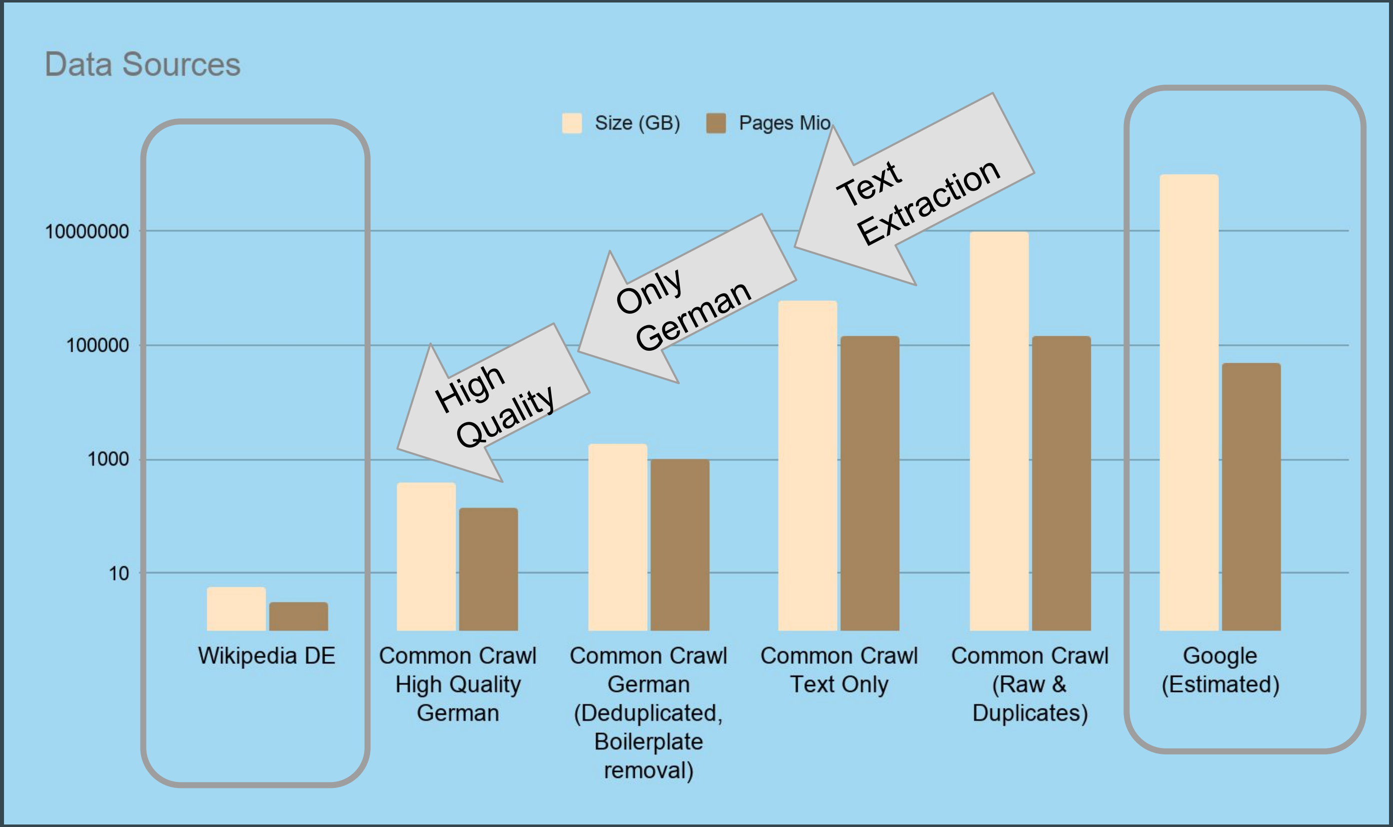
3



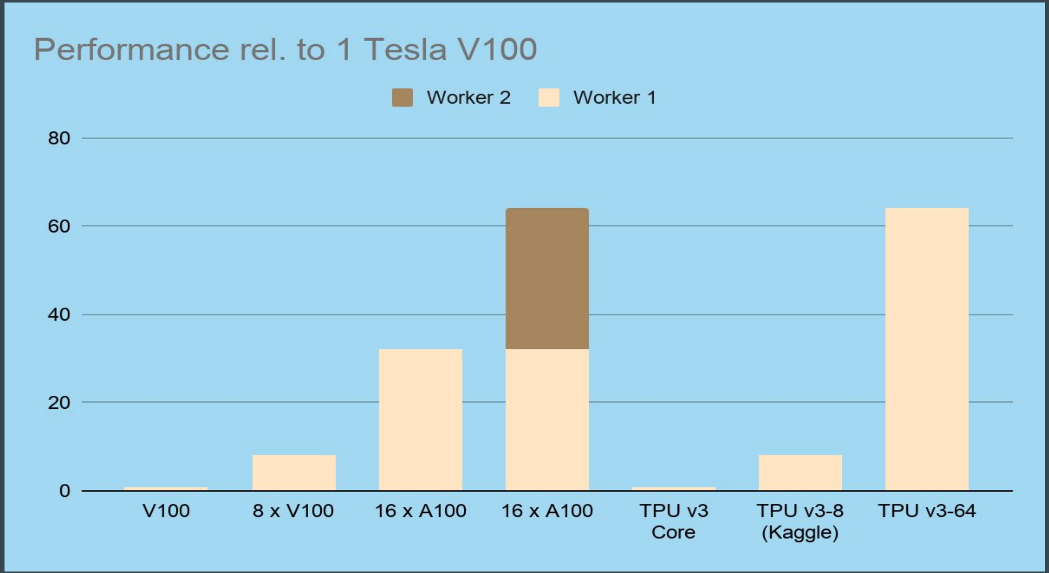
```
tokenizer = AutoTokenizer.from_pretrained("german-nlp-group/electra-base-german-uncased")
# {"do_lower_case": true, "strip_accents": false}
tokenizer.tokenize("Ich möchte meinen Vertrag kündigen.")
```

```
['ich', 'möchte', 'meinen', 'vertrag', 'kündigen', '.']
```

Size Matters ! - I want it all



Who is gonna pay for it?



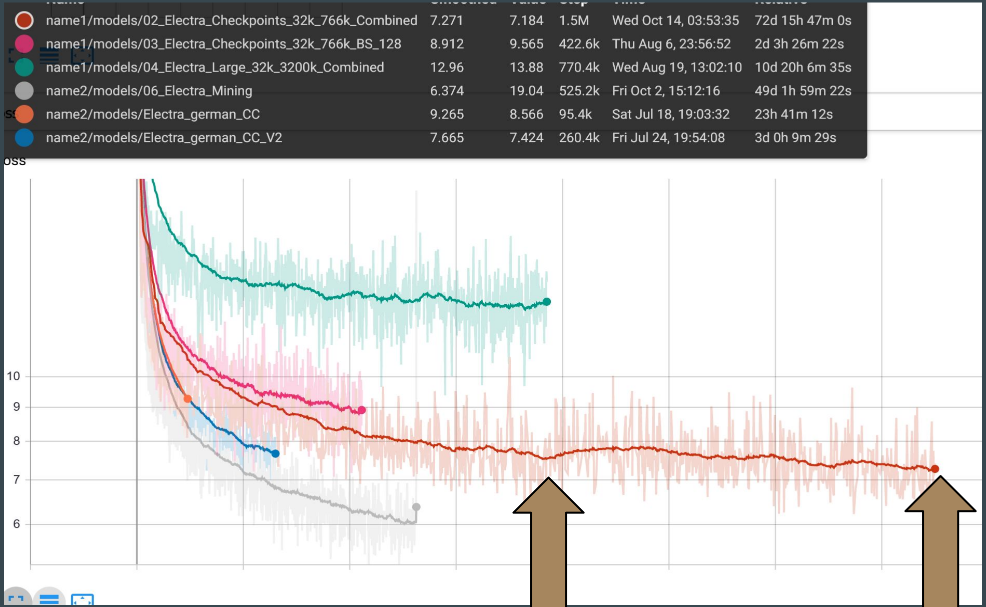
0.74 $ 5.92 $ 20 $ 40 $ 2.40 $ 20 $

Approx 1000 GPU hours (V100)

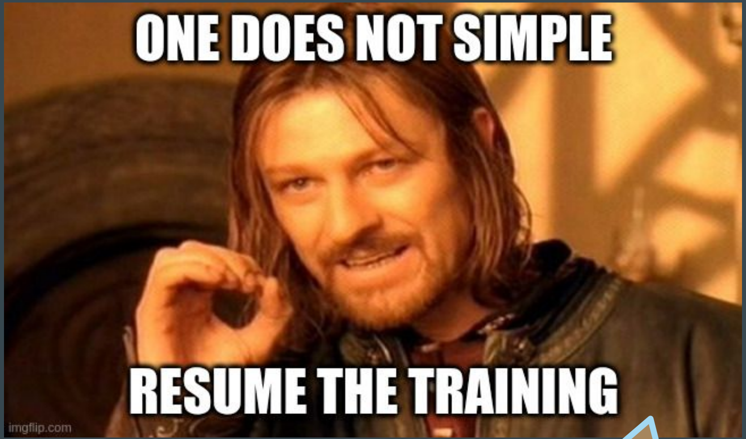
Language model

TPUs

The Training



	Name	Smoothed	Value	Step	Time	Relative
●	name1/models/02_Electra_Checkpoints_32k_766k_Combined	7.271	7.184	1.5M	Wed Oct 14, 03:53:35	72d 15h 47m 0s
●	name1/models/03_Electra_Checkpoints_32k_766k_BS_128	8.912	9.565	422.6k	Thu Aug 6, 23:56:52	2d 3h 26m 22s
●	name1/models/04_Electra_Large_32k_3200k_Combined	12.96	13.88	770.4k	Wed Aug 19, 13:02:10	10d 20h 6m 35s
●	name2/models/06_Electra_Mining	6.374	19.04	525.2k	Fri Oct 2, 15:12:16	49d 1h 59m 22s
●	name2/models/Electra_german_CC	9.265	8.566	95.4k	Sat Jul 18, 19:03:32	23h 41m 12s
●	name2/models/Electra_german_CC_V2	7.665	7.424	260.4k	Fri Jul 24, 19:54:08	3d 0h 9m 29s



V1 (Original)

V2 (Extended)

Learning Rate decay !!!

Evaluate and compare Language Models

On downstream Tasks

- GermEval 18
- Offensive Language
- F1 macro

With individual Hyperparameters

- Extensive automated Hyperparameter optimization (Optuna)
- per model
- avoid HPs that prefer one model and penalizes the other

Cross validation: avoid overfitting on validation set

Do multiple evaluations to show statistical significance

Compare with boxplot

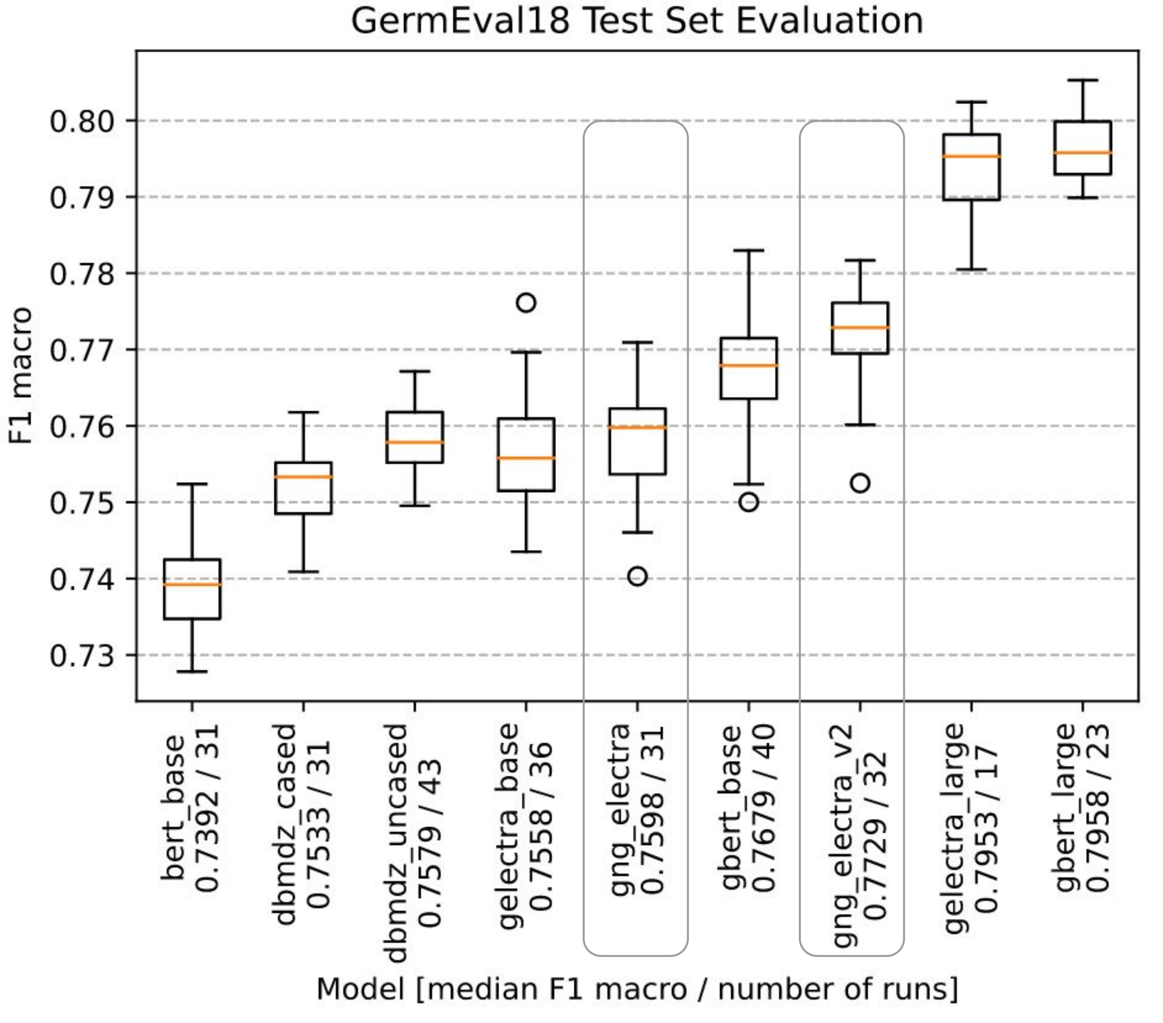
Version 2 of our Electra Model

many models suffer from undertraining

while others benefit from extensive training

from 766,000 to 1,500,000 steps

Results


GermEval18 Test Set Evaluation
F1 macro
0.80
0.79
0.78
0.77
0.76
0.75
0.74
0.73
bert_base 0.7392 / 31
dbmdz_cased 0.7533 / 31
dbmdz_uncased 0.7579 / 43
gelectra_base 0.7558 / 36
gng_electra 0.7598 / 31
gbert_base 0.7679 / 40
gng_electra_v2 0.7729 / 32
gelectra_large 0.7953 / 17
gbert_large 0.7958 / 23
Model [median F1 macro / number of runs]

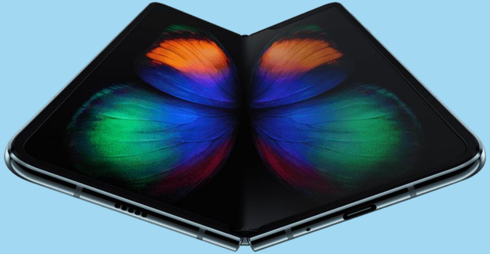
Discussion: Monolingual Models - the path to nowhere?

Monolingual:
Limited Knowledge from one Language

German Electra (V2):
$1.3 * 10^{20}$ Flops = 100 TPU Days

VS

Multilingual Models:
Knowledge from multiple Languages combined



mt5 11 Billion Params:
$3.3*10^{22}$ Flops = 33 000 TPU Days

You want to know the details?

- **Version 1 & 2 of german-nlp-group/electra-base-german-uncased:** german-nlp-group/electra-base-german-uncased · Hugging Face
- [2003.10555] ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators
- Electra PR for (keep accents): https://github.com/google-research/electra/pull/88
- SoMaJo: https://github.com/tsproisl/SoMaJo
- https://gitter.im/German-Transformer-Training/community
- Download and clean Common Crawl: https://github.com/facebookresearch/cc_net