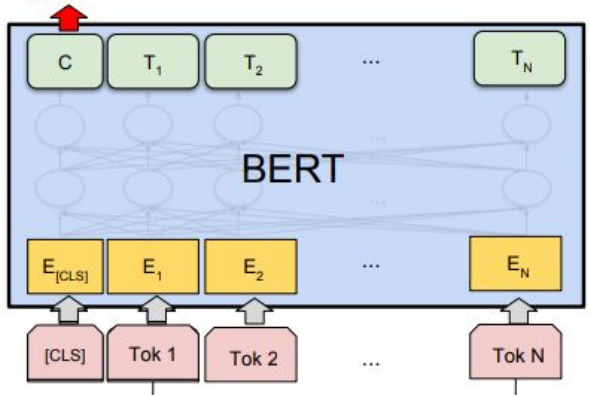
Evaluating Czech ALBERT

Petr Zelina
469366

What is ALBERT

- language model published by Google
- leaner version of BERT
- based on transformers
- creates bidirectional contextual embeddings of words
- pretrained on two self-supervised tasks:
  - Masked language modeling
  - Sentence order prediction
- can be adapted for many NLP tasks



BERT paper - https://arxiv.org/pdf/1810.04805.pdf

Czech ALBERT

- pretrained on TenTen dataset
- using SentencePiece tokenization, 30 000 tokens
- model parameters:
  - input tokens: 256
  - output embedding size: 512
  - attention heads: 8
  - intermediate size: 2084
- model size: 7 mil weights, approx. 100 MB
- speed:
  - GTX 1060, 6GB: 30 ms / prediction (batch size 16)
  - nVidia Tesla T4, 16GB: 20 ms / prediction (batch size 32)

SentencePiece - https://github.com/google/sentencepiece

Pretraining results

pretraining of csalbert is insufficient

Syn2015 model has better pretraining results because of less diverse dataset and more epochs

Pretraining of csalbert took 4 days, used less than ⅛ of TenTen dataset.

To achieve same quality as Google, it would take 5 months of training on single GPU (nVidia Tesla T4, 16GB)

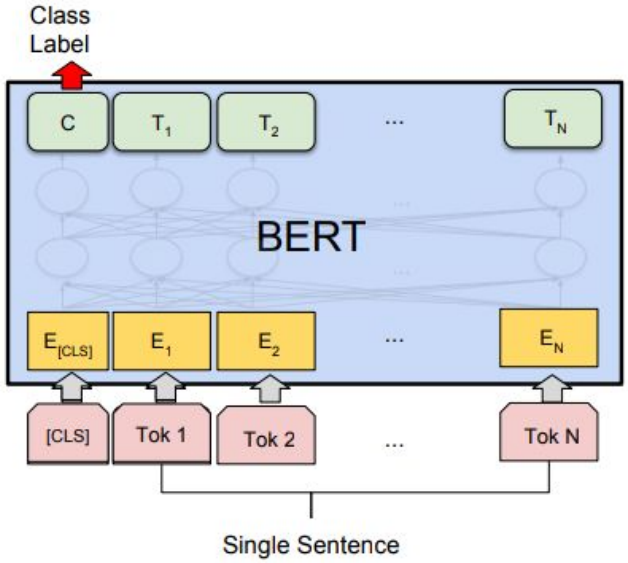
TenTen corpus is big enough.

	batch size	steps	total examples
Google	4096	125 000	512,0 M
csalbert	32	380 000	12,6 M

	MLM	SOP
Google	54 %	86 %
Syn2015 albert	28 %	75 %
csalbert	21 %	71 %

Text classification

Making a classifier from (AL)BERT model



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

- the first input token is always [CLS]
- the rest of input is filled with the tokenized text
- the output C is the contextual embedding of the [CLS] token
- we can connect additional layers to this embedding

- using embeddings of the other tokens leads to a bigger model that overfits faster (they are used for the masked modeling)

BERT paper - https://arxiv.org/pdf/1810.04805.pdf

Classification – Novinky

Dataset

- articles from novinky.cz from 8 categories
- 4096 articles per category

Results

- fasttext model for comparison
  - trained on Syn2015, embedding size 300,
  - 100 000 words in dictionary, 6 GB

model	accuracy
fasttetx builtin	81 %
fasttext + LSTM	87 %
csalbert	90 %

fasttext builtin: https://arxiv.org/pdf/1607.01759.pdf

Classification – Propaganda – Dataset

Dataset

- collection of news articles with manually labeled manipulative techniques
- 17 classification tasks (binary and multi-class)
  - Blaming, Emotions, Fear mongering, Location, Topic, …

The dataset does not have unified final form, I had to extract the answers from sql records of people who labeled it and select the most common answer.

Propaganda - https://www.aclweb.org/anthology/R19-1010.pdf

Classification – Propaganda – Dataset

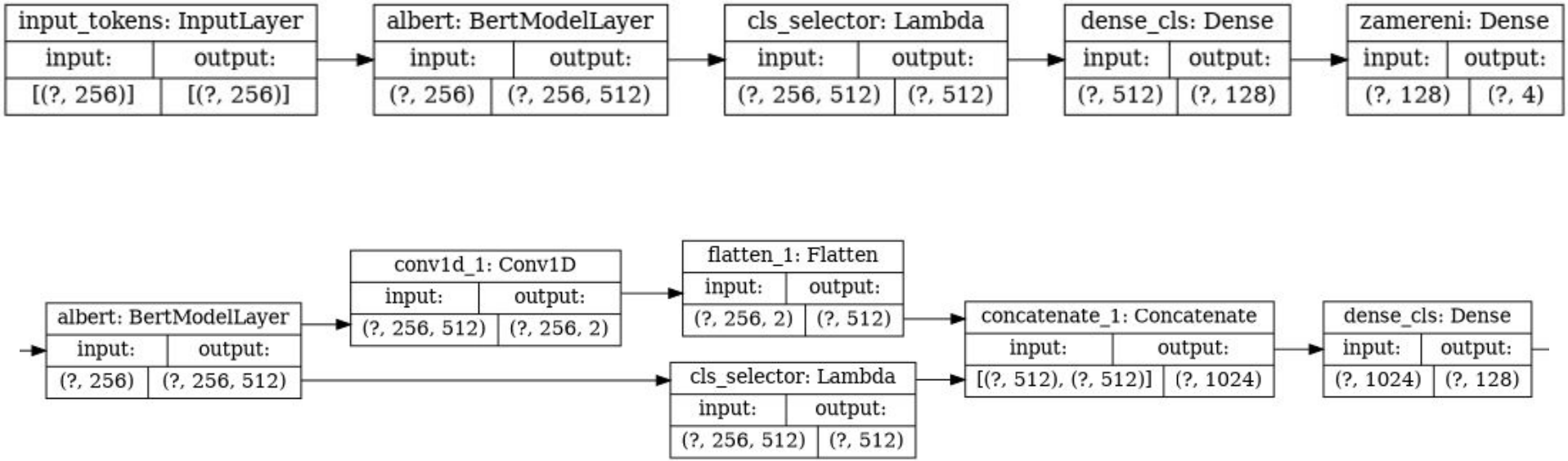
task	Blaming	Labelling	Argumentation	Emotions	Demonizing	Relativizing	Fear mongering	Fabulation	Opinion	Location	Source	Russia	Expert	Topic	Genre	Focus	Overall sentiment
classes	2	2	2	6	2	2	2	2	2	8	2	6	2	13	3	4	3
biggest class	0,68	0,85	0,58	0,85	0,97	0,93	0,93	0,80	0,87	0,36	0,60	0,66	0,57	0,29	0,91	0,58	0,80
delta	0,04	-0,03	0,17	-0,04	-0,01	0,01	-0,01	-0,01	0,02	0,40	0,10	0,16	0,24	0,42	0,05	0,29	-0,02
soa accuracy	0,72	0,82	0,75	0,81	0,96	0,94	0,92	0,79	0,89	0,76	0,70	0,82	0,81	0,71	0,96	0,87	0,78

Many tasks have unbalanced classes and their SOA classifier is is probably mostly guessing the majority class.

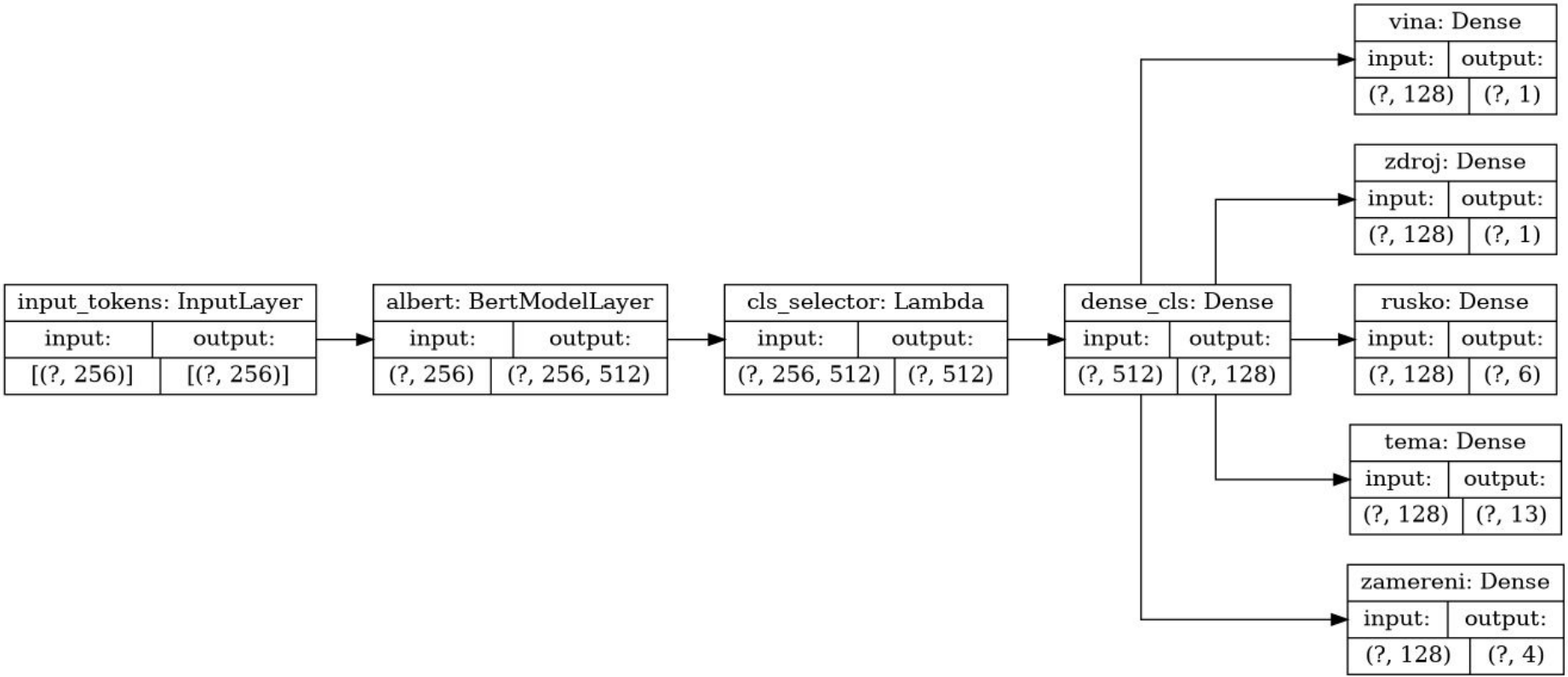
Because ALBERT is a big model, I chose to classify only a subset of the tasks.

Propaganda - https://www.aclweb.org/anthology/R19-1010.pdf

Classification – Propaganda – Architecture



Classification – Propaganda – Architecture



Classification – Propaganda – Class weights

How to deal with unbalanced classes?

- remove surplus
- duplicate scarce
- class weights
  - during backpropagation, make bigger steps when propagating minority class examples
  - increases the number of epochs before model starts to overfit
  - formula ensures normalization of learning rate

$$\text{class weight} = \frac{\text{dataset size}}{\text{class size} \cdot \text{number of classes}}$$

  - best results were achieved when the weights were brought closer to 1 every few epochs (model can learn real world frequency of classes), sqrt function

Classification – Propaganda – Results

task	Blaming	Argumentation	Location	Source	Russia	Expert	Topic	Focus
classes	2	2	8	2	6	2	13	4
biggest class	0,68	0,58	0,36	0,60	0,66	0,57	0,29	0,58
soa accuracy	0,72	0,75	0,76	0,70	0,82	0,81	0,71	0,87
csalbert accuracy	0,72	0,68	0,70	0,68	0,82	0,71	0,64	0,86

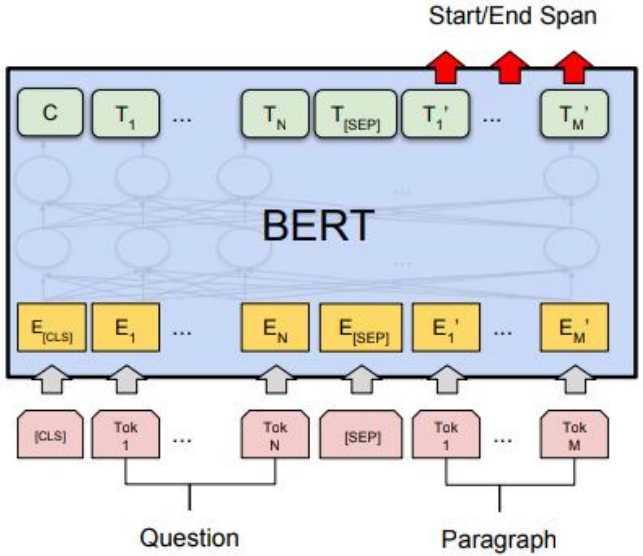
Even with insufficient pretraining, the model reaches similar accuracy as the SOA models in some tasks

Classification – Propaganda – Improvements

Many tasks are determined only by a small part of the article. The ALBERT model cannot see the whole article at once. These ranges are labeled only in a small subset of the dataset.

Question Answering

Making a question answering model from (AL)BERT



(c) Question Answering Tasks: SQuAD v1.1

Tokenized question and paragraph as input separated by [SEP] token

The model predicts on which token in the paragraph the answer starts and ends

BERT paper - https://arxiv.org/pdf/1810.04805.pdf

Czech SQAD – Dataset

ALBERT does not use lemmatized text.

Default architecture cannot answer yes/no questions (16.8 % of SQAD)

60 % of answers are in the first sentence.

Many questions are from wikipedia (date of birth, ...)

English SQuAD

- questions directly tied to paragraph (not whole article)
- 10x as many questions per paragraph

SQAD v3 - https://nlp.fi.muni.cz/raslan/2019/paper14-medved.pdf

Czech SQAD – Example

Jaké je hlavní město Polska?

Varšava

---

varšava (polsky warszawa, výslovnost [var'] ipa) je hlavní (od roku 1596) a největší město polska. v roce 2014 měla 1 726 581 obyvatel a s okolní aglomerací 3 003 000 obyvatel. varšava leží ve středním polsku v historickém mazovsku na středním toku visly ve varšavské kotlině v průměrné výšce 100 metrů n. m., 520 km východně od berlína a 250 km jižně od pobřeží baltského moře. varšava je také hlavním městem mazovského vojvodství. v metropoli je rozvinutý průmysl, zvláště zpracovatelský, ocelářský, elektrotechnický a automobilový. sídlí zde více než 60 vzdělávacích institucí, především varšavská univerzita (uniwersytet warszawski), univerzita kardinála stefana wyszyńského, varšavská technická univerzita (politechnika warszawska), varšavská ekonomická škola (szkoła główna handlowa) a další. je tu přes 30 divadel, včetně národního divadla a opery a má tu sídlo národní filharmonický orchestr. související informace naleznete také v článku dějiny varšavy. první opevněné osídlení na území dnešní varšavy byla osada bródno v 9. a 10. století a jazdów ve 12. a 13. století. po tom, co vévoda z płocku, boleslav ii. mazovský zaútočil v roce 1281 na jazdów, byla založeno nové sídlo v místě malé rybářské vesničky warzsowa.

Varšava
prob: 0.23 | slp: -1.20 | elp: -1.27

Varšava (polsky Wars
prob: 0.21 | slp: -1.20 | elp: -1.33

(polsky Warszawa
prob: 0.13 | slp: -2.18 | elp: -0.81

Varšava (polsky Warszawa
prob: 0.11 | slp: -1.20 | elp: -1.99

Varšava (polsky Warsza
prob: 0.10 | slp: -1.20 | elp: -2.11

Varšava leží ve středním Polsku
prob: 0.08 | slp: -2.82 | elp: -0.65

Varšava (polsky Warszawa, výslovnost [var'] IPA
prob: 0.05 | slp: -1.20 | elp: -2.78

Czech SQAD – Results

evaluation still in progress

due to sentencepiece tokenization “instability” and the current way of tokenization, some correct answers are marked as incorrect

exact match of answer extraction: 10 %
f1 of answer coverage: 17.7

SOA of answer sentence selection: 79 %

Czech SQAD – Improvements

Currently training model more suitable for wikipedia with improved tokenization

implement exact match for in N best predictions metric

Conclusion

Results are not great, but considering insufficient pretraining,
the model has potential.

For classification, more than 1000 examples per class is needed.

Future plans

Downstream tasks comparison with multilingual / slavic BERT

TF 2.1 implementation

- multiple GPUs
- half precision
- direct parallelized pipeline

Literature

BERT paper - https://arxiv.org/pdf/1810.04805.pdf
ALBERT paper - https://arxiv.org/pdf/1909.11942.pdf
Propaganda dataset - https://www.aclweb.org/anthology/R19-1010.pdf
Czech SQAD v3 - https://nlp.fi.muni.cz/raslan/2019/paper14-medved.pdf

Thanks to MetaCentrum