“Friends” dialogues’ classification challenge

by Andrew Argatkiny

* Image generated by Sber’s RuDalle via query “Герои сериала "Друзья" с вопросительным знаком над головами”

/ Basic EDA

Key insights:

- Each data point consists of two semantically different parts
- The quality of translation to Russian is not perfect
- Occurrences of characters' responses seem to be evenly distributed in train and validation sets.
- ~ 25000 train, 2800 val, 3100 test dataset sizes
- 94% of dialogues (other_speaker + friend_response parts) have length <= 256 words

Ага, послушай. Прежде чем вы сделаете что-нибудь в духе Джоуи, вы могли бы пробежаться по этому.

Танцевать каратэ?

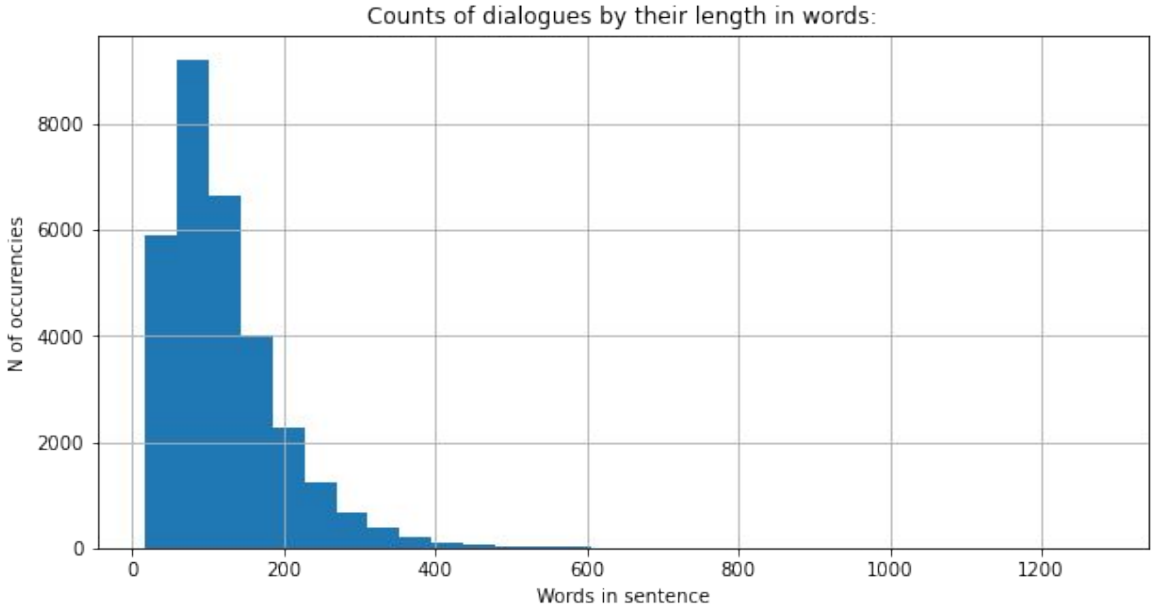
как насчет того, чтобы переехать к вам?

Ух ты! Ты выглядишь, перестань есть горячо! Это похоже на высшую степень жара!

Потому что я ее уже пригласил.

Знаю, да.

Ну что ж, отлично.



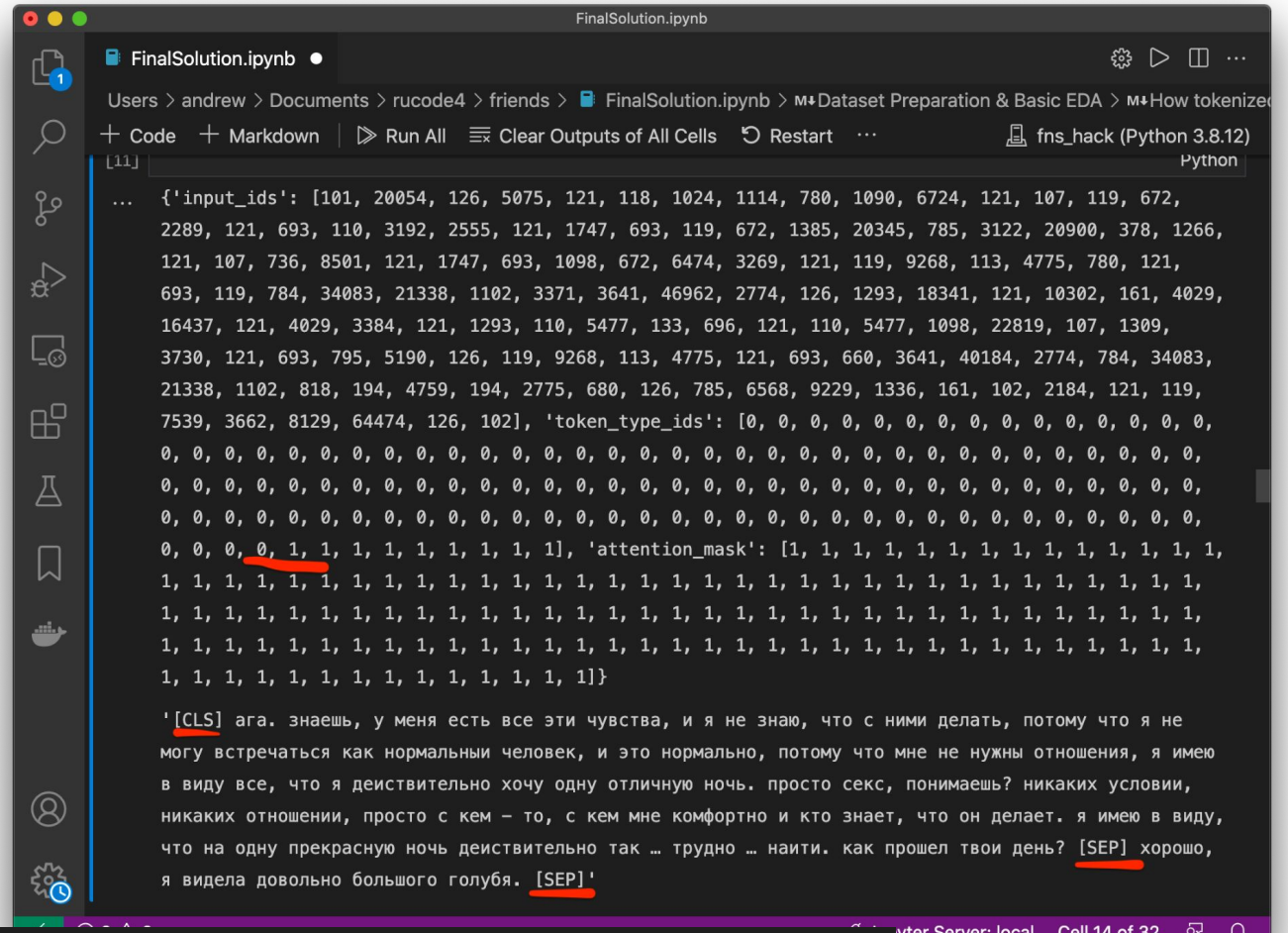

```
РОСС        0.176569
РЕЙЧЕЛ      0.176089
ЧЕНДЛЕР     0.170568
ДЖОУИ       0.166287
МОНИКА      0.160525
ФИБИ        0.149962
Name: label, dtype: float64

РОСС        0.176746
РЕЙЧЕЛ      0.176026
ЧЕНДЛЕР     0.170626
ДЖОУИ       0.166307
МОНИКА      0.160547
ФИБИ        0.149748
Name: label, dtype: float64
```

/ Solution: use transformer neural network model!

SBER DEVICES

Fine-tuning ruBert base model from Sber-AI via Hugging Face API.

- Why even bother with other approaches when you can use SOTA model pretrained on Russian language?
- BERT is pretrained on the NSP (next sentence prediction) task so it's naturally adopted for two-parts text inputs classification
- Just feed the tokenizer with pairs of texts.
- I set truncation and padding to 256 tokens (out of max possible 764 for bert-base), so little information was lost.
- But the trade-off is longer training time.



/ Solution outline

Architecture & parameters

- Dense multi-classification layer with softmax activation on top of BERT's final embedding layer with dropout (**p = 0.3**)

- **3** epochs
- learning rate **2e-5**
- **AdamW** optimizer
- weights decay **1e-2**
- input sequence length **256**
- batch size **16** (effective **32** due to backprop delay)
- constant scheduling
- use of mixed precision

- Validation set as provided by organizers
-
- Training took ~ 45 mins on GTX 1070

Tips & tricks

- Use mixed precision (fp16) to train faster and be able to use larger batches.

- Average gradients for n batches and backprop once in n iterations to allow n times larger batch sizes.

- Different regularization techniques do help.

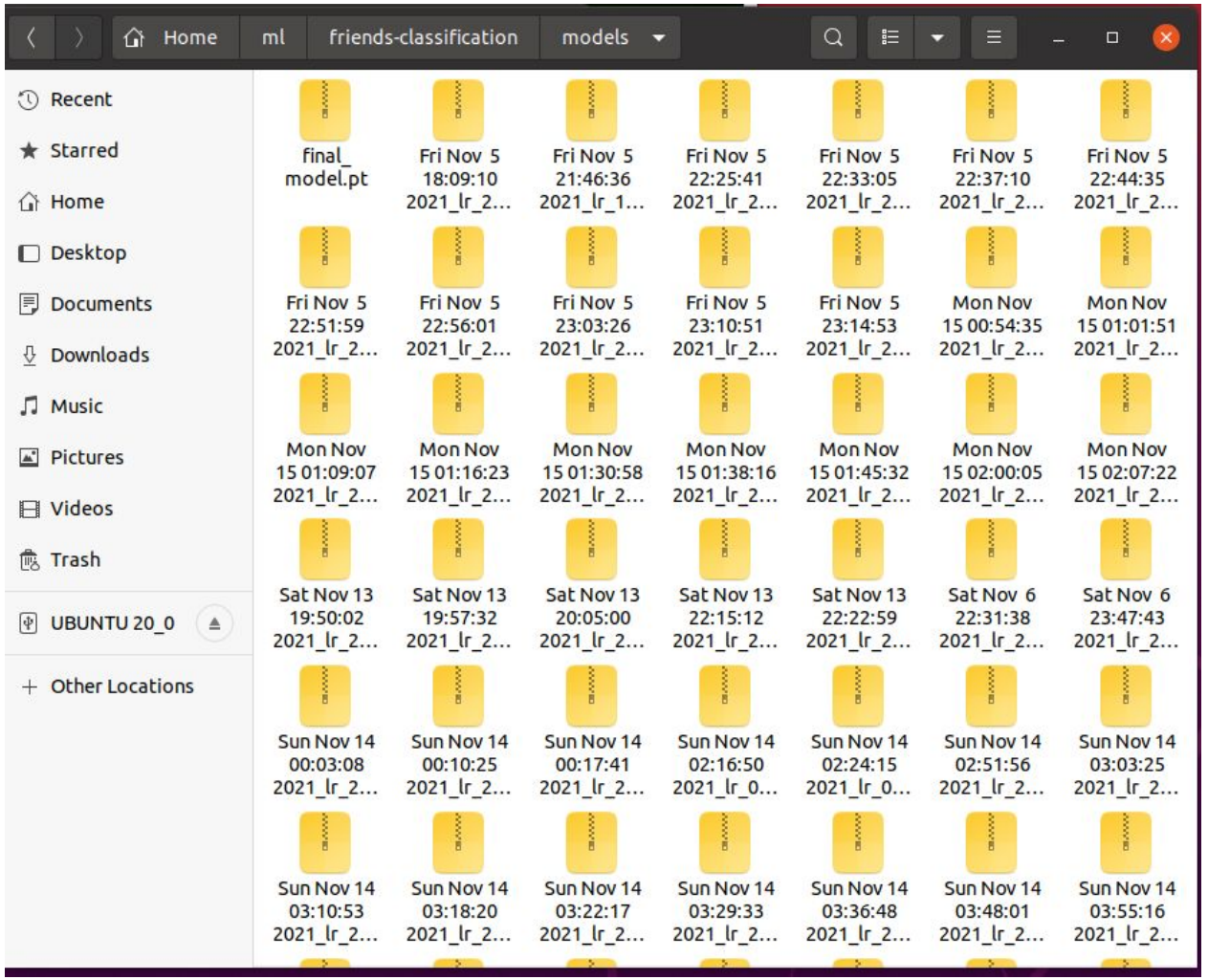
- Use all available data for final model.

- Consider local validation improvements, don't overfit to LB (I scored 7th place on public LB, but came 1st on private).

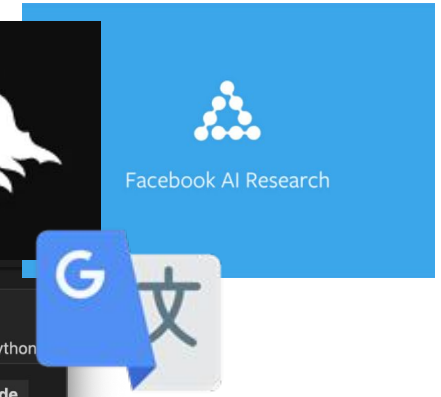
/ A single submission to LB

But it was far from one model that I had made and tested:

Best accuracy on validation set was **35.2-35.3%**



/ Failed attempts – backtranslation and back-backtranslation



RUS -> ENG

- Facebook AI Research's mBART-50 multilingual model (2020) https://arxiv.org/abs/2001.08210
- University of Helsinki's models based on MarianMT transformer architecture
- Google's NMT (well, sort of - googletrans, translators, gsheets)

Facebook model shows the best quality among 3 options (subjectively)

Back from ENG to RUS

- Facebook English translation back to Russian

FriendsPredictNLP2.ipynb

FriendsPredictNLP2.ipynb | FriendsPredictNLP3.ipynb | FriendsPredictNLP_roberta.ipyn

Users > andrew > Documents > rucode4 > friends > FriendsPredictNLP2.ipynb > train_bert(net, criterion, o

+ Code | + Markdown | Run All | Clear Outputs of All Cells | Outline | ...

[8]

```
df_train_3X.drop_duplicates()
```

Python

	Id	other_speaker	friend_response	label	label_code
0	0	All right! All right! Enough! Enough! Enough! ...	You know, I think you can take it.	ЧЕНДЛЕР	5
1	1	Oh, you want a good name, go with Joey. Joey i...	Hey, you know what, if you're going to do this...	ЧЕНДЛЕР	5
2	2	My God, what!?	As I said, I was thinking of taking Emma to th...	РОСС	3
3	3	Pregnancy has a severe effect on the body.	Hey, but even though you got this cool glow of...	ДЖОУИ	0
4	4	Well, no, that's the wrong decision. That's no...	Well, listen, yesterday I would have even thou...	РОСС	3
...	...	...	...	...	...
24988	24988	Nothing! Well, I had a blinding pain in my sto...	Looks like herriation. You must, you - go to t...	ЧЕНДЛЕР	5
24989	24989	Hi.	Very good handshake, good wrist movement.	ФИБИ	4
24990	24990	Joey, how did you spoil it?	Yes, it's easy. Yes, I can do it with anything...	ДЖОУИ	0
24991	24991	Hey! How was your date with Jake?	We could not take the eyes from each other all...	ФИБИ	4
24992	24992	Phoebe, 800 - free, 801 - Utah.	No, no, no, oh no no, no, it should be 801. Be...	ФИБИ	4

74198 rows × 5 columns

⊗ 0 ⚠ 13 | Jupyter Server: local | Cell 31 of 47

- Up to **3X** size English dataset as compared to original
- Up to **2X** size Russian dataset

/ Failed attempts – Other augmentation methods

Mechanical augmentations

Based on https://www.kaggle.com/shonenkov/nlp-albumentations

- Random shuffling of sentences in multi-sentence lines.
- Neighboring words swapping for all words in sequence with some probability.
- Apply each augmentation n times to underlying dataset.

Largest dataset with all augmentations and translation versions had ~ 370 000 observations.

Replacements by synonyms with Roberta's contextualized embeddings

	Id	other_speaker
90	90	It is somehow sad.
91	91	Umm, listen, I think you should know something.
92	92	I know. I know. Oh my God. No cloth! Can you b...
93	93	A?
94	94	You know what, give me a second, and I leave y...
95	95	Sorry!
96	96	I am Kristen.
97	97	No, got into your food!
98	98	In fact, it is not true, in an incredible Halk...
99	99	Good. Hi, Mont, why did you tell the guys that...

Id		
90	It is somehow forbidden.	Ye
91	Umm, listen, I read taxpayers should see somet...	
92	I do. Si know. Oh my Zeus. No ID! Can you brin...	
93	A?	
94	You know what, read me yourself second, hell o...	
95	Sorry!	
96	I am ill.	Y
97	No, got level your game!	
98	In 1998, it it fairly easy, in an incredible H...	S
99	Good. Hi, Marco, everything did you tell the g...	

- nlpaug library: https://github.com/makcedward/nlpaug

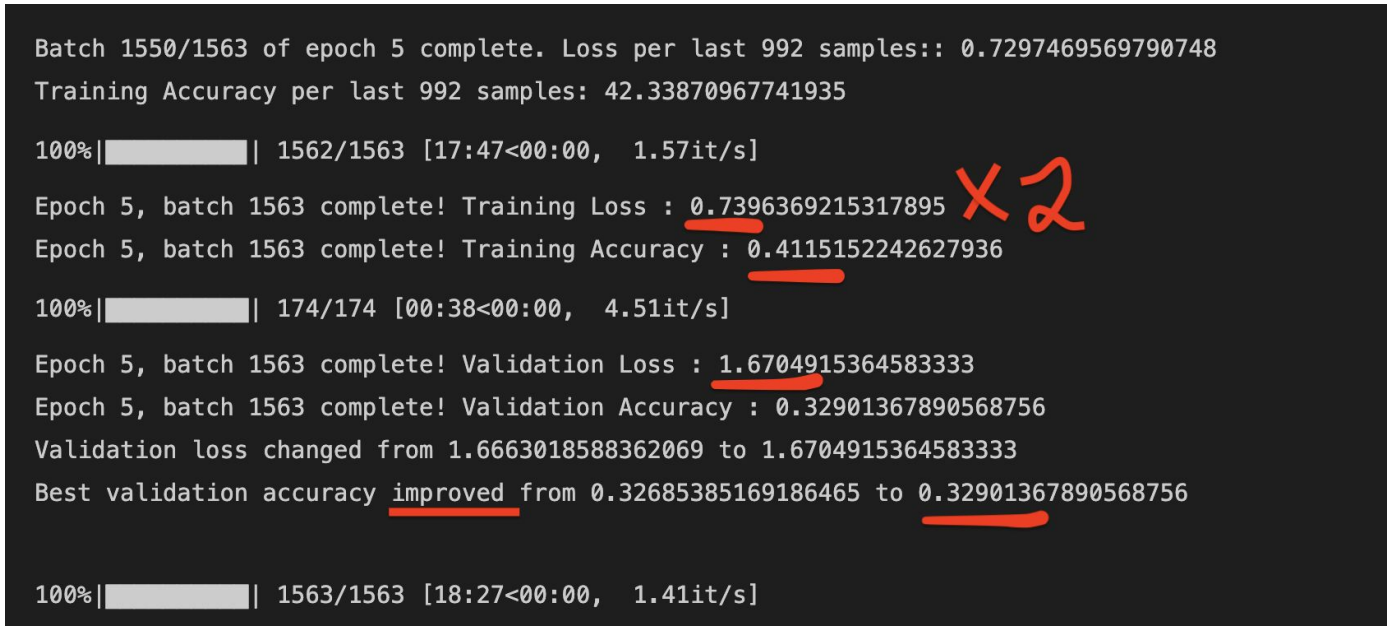
BERT-like transformers have yet a long way to go to excel at paraphrasing!

/ Other approaches

- Additional dense layer with dropout before softmax
- Freezing some or all BERT's layers during some or all epochs
- Separate training on friend_response/ other_speaker
- Linear schedule with warm up
- different Russian and English models (sberbank-ai/sbert_large_mt_nlu_ru, bert-base uncased and cased, roberta and others)
- Hyperparameters tuning
- Best results were achieved when I trained for first 1-2 epochs with **lr=2e-5** and then decreased to **lr=2e-6.**

Interesting observation:

Validation accuracy keeps growing quite for some time after training and validation metrics divergence and deteoritation of validation loss

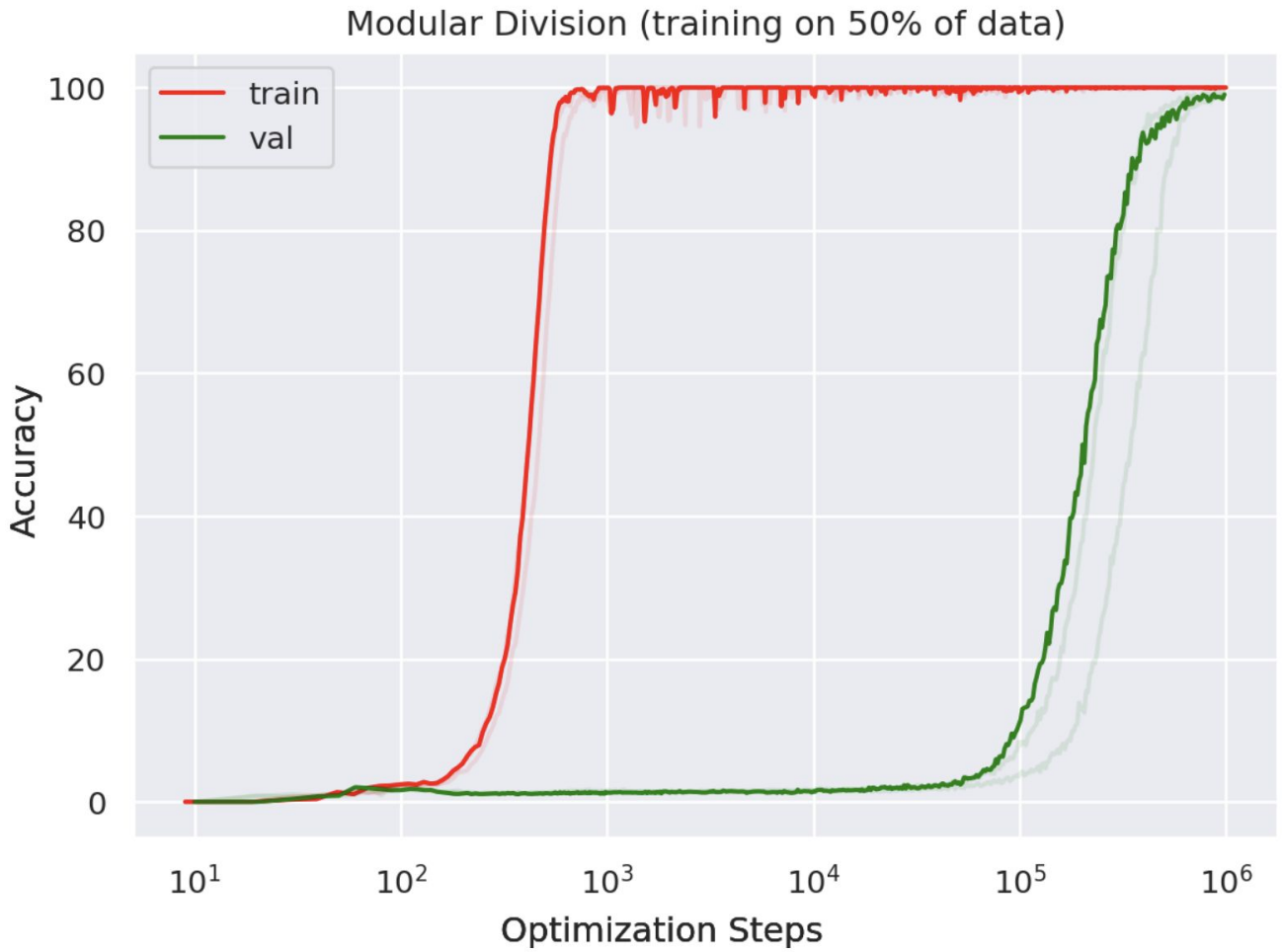


So it sparks one interesting idea to try...

/ Possible approaches to explore

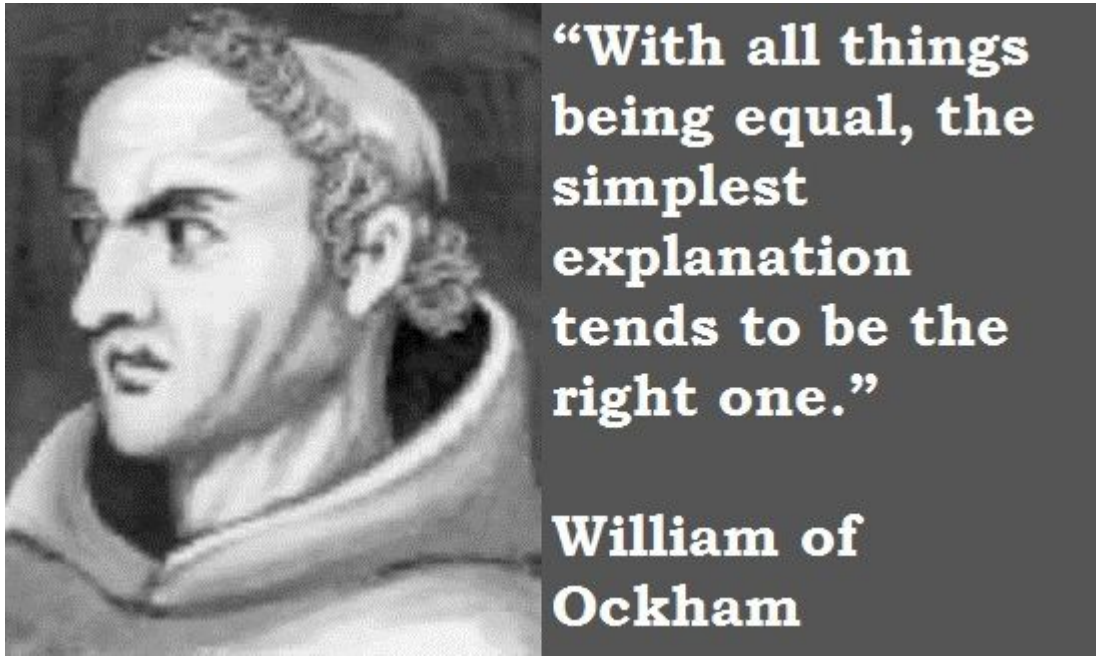
- Despite the utter overfitting after first few epochs, keep training for many-many iterations (batches), and the model will possibly generalize
- This actually can happen for some datasets according to recent OpenAI's research paper **GROKKING: GENERALIZATION BEYOND OVERFITTING ON SMALL ALGORITHMIC DATASETS**
- https://mathai-iclr.github.io/papers/papers/MATHAI_29_paper.pdf

- More down-to-earth approach: try increasing input sequence size to maximum and see if it brings marginal improvements



/ Morale – oftentimes, simpler models are better ones.

- Okkam’s razor in action
- ~~Attention~~ baseline is all you need!



/ Thanks for your attention!

Andrew Argatkiny

BA/DA/DS at WAYGROUP

https://github.com/andrewargatkiny

tg: @andrewargatkiny