# Preprocessing Sequential Data for Machine Learning Facilitation using Curriculum Learning

Project Proposal

Maxim K. Surkov, group BPM171

Research Advisors: Ivan P. Yamshchikov

Linguistic Advisor: Department of Foreign Languages

Saint Petersburg School of Physics, Mathematics, and Computer Science
Department of Computer Science

6 april 2021

# Outline

- Motivation and definitions
- Literature Review
- Methodology
- Results

# Motivation

- social networks
- voice assistants
- translators
- chatbots

- classification
- machine translation
- natural language understanding

- tiny language models
- GPT-3
  - extremely large
- **BERT**
  - high quality

# Motivation

- pre-training
  - required time: from 1-2 days to **1-2 weeks**
  - world record: 47 minutes using **1472** GPUs

| Dataset | Samples |
|---|---|
| Wikipedia | 3-600M |
| BooksCorpus | 74M |

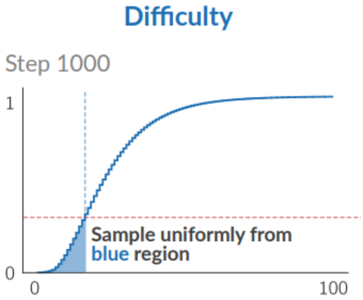
- fine-tuning
  - required time: 1-2 days

| Dataset | Samples |
|---|---|
| HND | 600k-2M |
| s140 | 1.6M |
| IWSLT | 200-230k |
| QQP | 364k |
| MNLI | 393k |

# Curriculum Learning. Definition

- task: machine translation
- Models: BERT, LSTM
- Datasets: IWSLT'15, IWSLT'16, WMT'16
- Algorithm:

1. sort samples by text complexity (length, log-likelihood)
2. for $T$ steps (consider $t$-th step)
    - calculate $c(t) \in [0, 1]$
    - form the batch from $c(t)$ **easiest** samples
    - do training step



# Research Field

| metric | classification | MT | pre-training | NLU |
|---|---|---|---|---|
| length | | ✓ | | |
| *language features*[1] | | | | |
| entropy | | | | |
| model-based | | | | ✓ |
| word frequency based | | ✓ | | |
| log-likelihood | | ✓ | | |
| ? | | | | |

- length is the best metric now
- there is no universal approach
- classification and pre-training is not investigated

[1] van der Sluis et al. (2010) showed that there is poor correlation with real text complexity

# Problem Statement

**Goal:** speed up the BERT model's training process at the expense of applying effective text complexity estimation metrics within the framework of curriculum learning on pre-training and classification tasks

**Problems:**

1. Suggest alternative text complexity metrics
2. Implement a practical algorithm for metrics calculation on large datasets
3. Carry out comparative analysis between the proposed metrics and the existing ones
4. Study the impact of the found metrics on the BERT training time

# Literature Review

| | |
|---|---|
| Bengio et al. (2009) | it was first shown that curriculum has a great potential for improving ML models |
| Hacohen and Weinshall (2019)<br>Mermer et al. (2017) | application of curriculum learning in computer vision |
| Platanios et al. (2019) | the first application of curriculum learning in machine translation |
| Xu et al. (2020) | model-based metric investigation |
| Tom Kocmi et al. (2017)<br>Xuan Zhang et al. (2018) | good results on the machine translation task were shown using curriculum learning with **length** and **word frequency rank** metrics |
| Nihat Ay et al. (2006) | Excess Entropy and TSE metrics description |

# Methodology: metrics

- filtered metrics
  - length
  - word frequency rank
  - log-likelihood
  - language metrics are **not** used
- Information Retrieval
  - TF-IDF
- Information Theory
  - *Excess Entropy*
  - *TSE*

# Methodology: metrics calculation

- Information Theory metrics adaption to texts

$$T = (t_1, t_2, \dots, t_{i-1}, t_i, \dots, t_n)$$

$$t_i \to \xi^i_{t_i} =: \mu_i - \text{binary random value}$$

$$\downarrow$$

$$\xi = (\xi^1_{t_1}, \xi^2_{t_2}, \dots, \xi^{i-1}_{t_{i-1}}, \xi^i_{t_i}, \dots, \xi^n_{t_n})$$

- Statistics collection
  1. divide the dataset into parts
  2. collect statistics on multiple processors in parallel
  3. join
- Excess Entropy and TSE metrics calculation
  1. $\mathcal{O}^*(2^n)$
  2. $\mathcal{O}(n^2)$ - dynamic programming
  3. $\mathcal{O}(n)$ - math equations and text's far-placed parts' independence assumption

# Methodology: comparison method

1. fix the dataset, model, and sampling algorithm
2. train BERT model
3. fix a sufficiently large accuracy value
    - train BERT model without curriculum learning until convergence
    - take the best accuracy value $\pm\varepsilon$
4. compare average number of steps required to reach this threshold

# Results

| dataset | HND (92.9%) | | | s140 (85.5%) | | |
|---|---|---|---|---|---|---|
| sampler | CB | DB | Hyp | CB | DB | Hyp |
| length | 55k | 23k | 22.5k | 112.5k | 20k | 19k |
| TF-IDF | $\infty$ | 19.5k | 24k | 115.5k | 21.5k | 19.5k |
| TSE | 56.5k | 21k | 23k | 95.5k | 16.5k | 20.5k |
| EE | 71.5k | 25.5k | 22.5k | 59k | 16.5k | 23k |
| max wf rank | $\infty$ | 22k | 20.5k | 70k | 18.5k | 19.5k |
| log-likelihood | $\infty$ | 20k | 24k | 112k | 17.5k | 21.5k |
| | | 22k | | | 18k | |

# References