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

April 6, 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

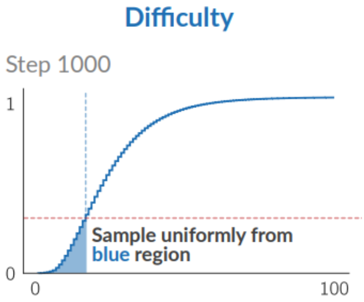
- fine-tuning
  - required time: 1-2 days

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) 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) 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

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

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

$$\downarrow$$

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

- 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

Classification task

dataset	HND (92.9%)				s140 (85.5%)			
sampler	CB	DB	Hyp	SS	CB	DB	Hyp	SS
length	55k	23k	22.5k	-	112.5k	20k	19k	-
TF-IDF	∞	19.5k	24k	23.5k	115.5k	21.5k	19.5k	16.5k
TSE	56.5k	21k	23k	22k	95.5k	16.5k	20.5k	21.5k
EE	71.5k	25.5k	22.5k	19.5k	59k	16.5k	23k	20k
max wf rk	∞	22k	20.5k	?	70k	18.5k	19.5k	?
likelihood	∞	20k	24k	?	112k	17.5k	21.5k	?
baseline	22k				18k			

- the final curriculum quality is highly dependent on the sampler
- length is not the best, but expressive enough
- TSE and EE are the most stable
- no strong acceleration

Plans

- finish the study on the pre-training problem
    - suspicious behavior of the curriculum learning is under investigation
- explore more metrics
    - model-based metric
    - avg/min word frequency rank
    - entropy

References

- Ay, N., Olbrich, E., Bertschinger, N., & Jost, J. (2006, August). A unifying framework for complexity measures of finite systems. In Proceedings of ECCS (Vol. 6).
- Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In Proceedings of the 26th annual international conference on machine learning (pp. 41-48).
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

References

- Hacohen, G., & Weinshall, D. (2019, May). On the power of curriculum learning in training deep networks. In International Conference on Machine Learning (pp. 2535-2544). PMLR.
- Kocmi, T., & Bojar, O. (2017). Curriculum learning and minibatch bucketing in neural machine translation. arXiv preprint arXiv:1707.09533.
- Kurdi, M. Z. (2020). Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL. arXiv preprint arXiv:2001.01863.
- Mermer, M. N., & Amasyali, M. F. (2017). Scalable Curriculum Learning for Artificial Neural Networks. IPSI BGD TRANSACTIONS ON INTERNET RESEARCH, 13(2).

References

- Narasimhan, S., Narasimhan, V. A. P. B. S., Karch, G., Rao, R., Huang, J., Zhang, Y., Ginsburg, B., Chitale, P., Sreenivas, S., Mandava, S., Ginsburg, B., Forster, C., Mani, R., & Kersten, K. (2020, October 13). NVIDIA Clocks World's Fastest BERT Training Time and Largest Transformer Based Model, Paving Path For Advanced Conversational AI. NVIDIA Developer Blog. https://developer.nvidia.com/blog/training-bert-with-gpus/
- Platanios, E. A., Stretcu, O., Neubig, G., Poczos, B., & Mitchell, T. M. (2019). Competence-based curriculum learning for neural machine translation. arXiv preprint arXiv:1903.09848.
- Sajjad, H., Dalvi, F., Durrani, N., & Nakov, P. (2020). Poor Man's BERT: Smaller and Faster Transformer Models. arXiv preprint arXiv:2004.03844.

References

- Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M. W., & Keutzer, K. (2020). Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8815–8821. https://doi.org/10.1609/aaai.v34i05.6409
- van der Sluis, F., & van den Broek, E. L. (2010, August). Using complexity measures in information retrieval. In Proceedings of the third symposium on information interaction in context (pp. 383-388).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

References

- Xu, B., Zhang, L., Mao, Z., Wang, Q., Xie, H., & Zhang, Y. (2020). Curriculum Learning for Natural Language Understanding. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6095–6104. https://doi.org/10.18653/v1/2020.acl-main.542
- Zhang, X., Kumar, G., Khayrallah, H., Murray, K., Gwinnup, J., Martindale, M. J., ... & Carpuat, M. (2018). An empirical exploration of curriculum learning for neural machine translation. arXiv preprint arXiv:1811.00739.