Carnegie Mellon University
Language Technologies Institute

# Accelerating Training

## Off- and on-device

Emma Strubell & Yonatan Bisk

# Training efficiency

$$Cost(R) \propto E \cdot D \cdot H$$

obtaining result

processing one example

**dataset size**

hyperparameter configurations



Equation from: Schwartz, Dodge, Smith, Etzioni. Green AI. Communications of the ACM, December 2020.

# Training efficiency

$$Cost(R) \propto E \cdot \boxed{D} \cdot H$$

obtaining result

processing one example

**dataset size**

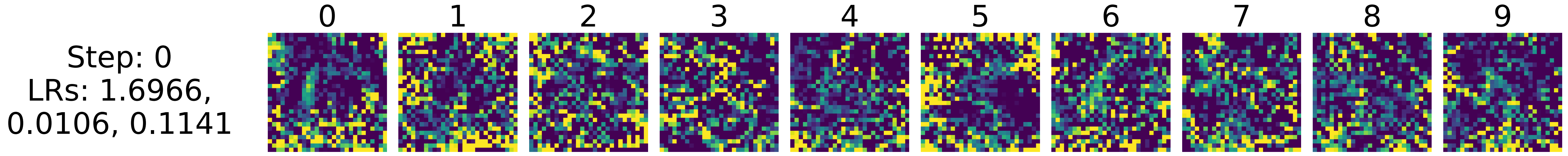
hyperparameter configurations

- Use fewer training examples
  - Change the data: Data optimization

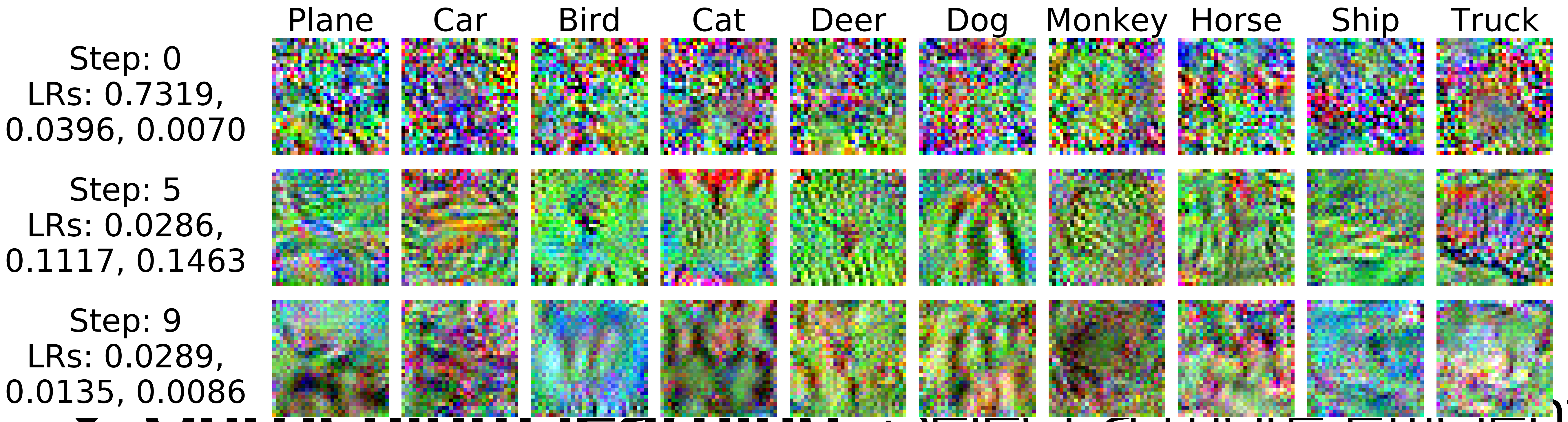
# Data Optimization

## How to reduce the number of training examples?

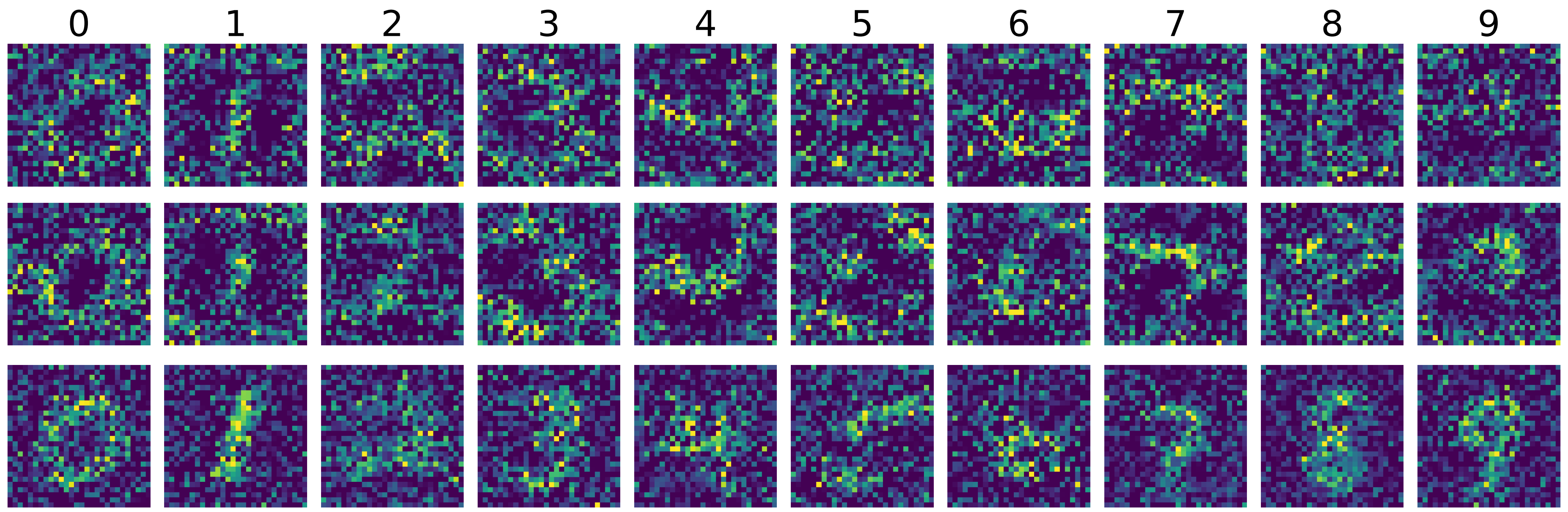
- **Data Pruning:** Select a *subset* of informative examples.
- **Data Distillation:** Generate a *synthetic* representative dataset.




(a) MNIST. Theses distilled images train a fixed initializations from 12.90% test accuracy to 93.76%.




(b) CIFAR10. These distilled images train a fixed initialization from 8.82% test accuracy to 54.03%.




These distilled images unknown random initializations to 79.50% ± 8.08% test accuracy.

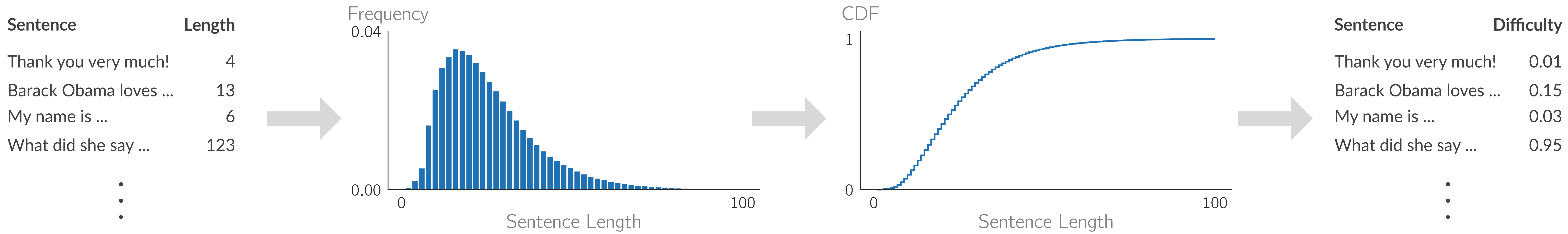
- **Curriculum learning:** Select a more efficient data *ordering.*

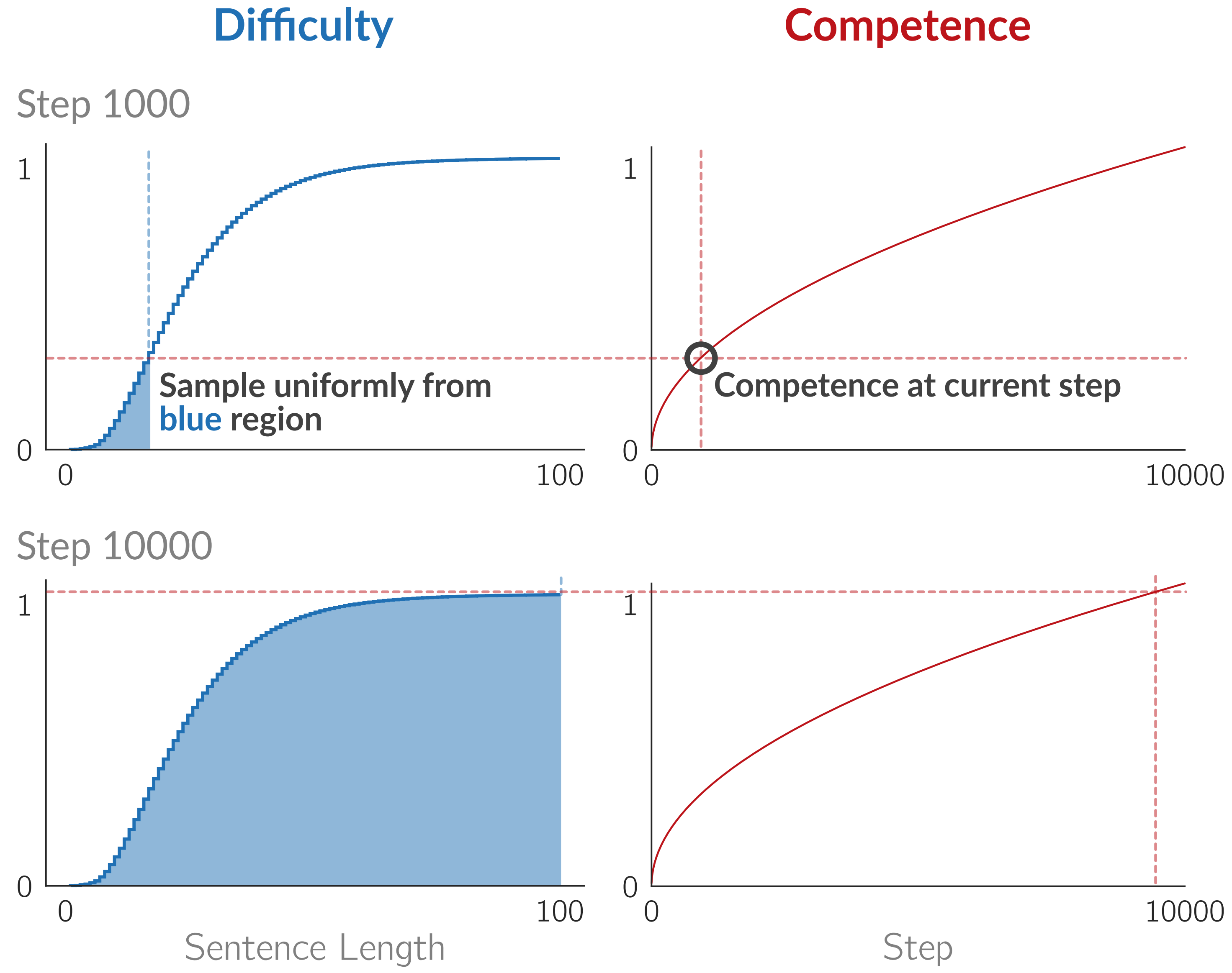


T. Wang, J-Y Zhu, A. Torralba, A. Efros. Dataset Distillation. arXiv 2019.

# Curriculum learning for machine translation

- **Key idea:** Training on examples ordered by increasing difficulty allows for faster convergence.




| | | Transformer | | | | | |
|---|---|---|---|---|---|---|---|
| | | Plain | Plain* | SL Curriculum | | SR Curriculum | |
| | | | | $c_{\text{linear}}$ | $c_{\text{sqrt}}$ | $c_{\text{linear}}$ | $c_{\text{sqrt}}$ |
| BLEU | En→Vi | 28.06 | 29.77 | 29.14 | 29.57 | 29.03 | **29.81** |
| | Fr→En | 34.05 | 34.88 | 34.98 | 35.47 | 35.30 | **35.83** |
| | En→De | – | 27.95 | 28.71 | 29.28 | 29.93 | **30.16** |
| Time | En→Vi | 1.00 | 1.00 | 0.44 | 0.33 | 0.35 | **0.31** |
| | Fr→En | 1.00 | 1.00 | 0.49 | 0.44 | 0.42 | **0.39** |
| | En→De | – | 1.00 | 0.58 | **0.55** | **0.55** | **0.55** |



E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, T. M. Mitchell. Competence-based Curriculum Learning for Neural Machine Translation. NAACL 2019.

# Training efficiency

$$Cost(R) \propto E \cdot \boxed{D} \cdot H$$

obtaining result

processing one example

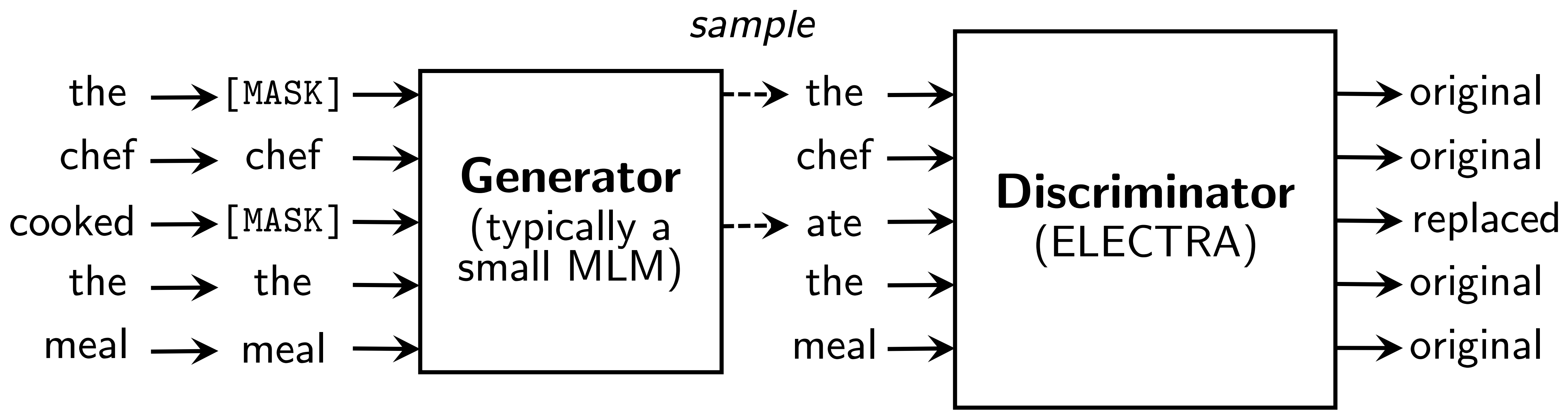
**dataset size**

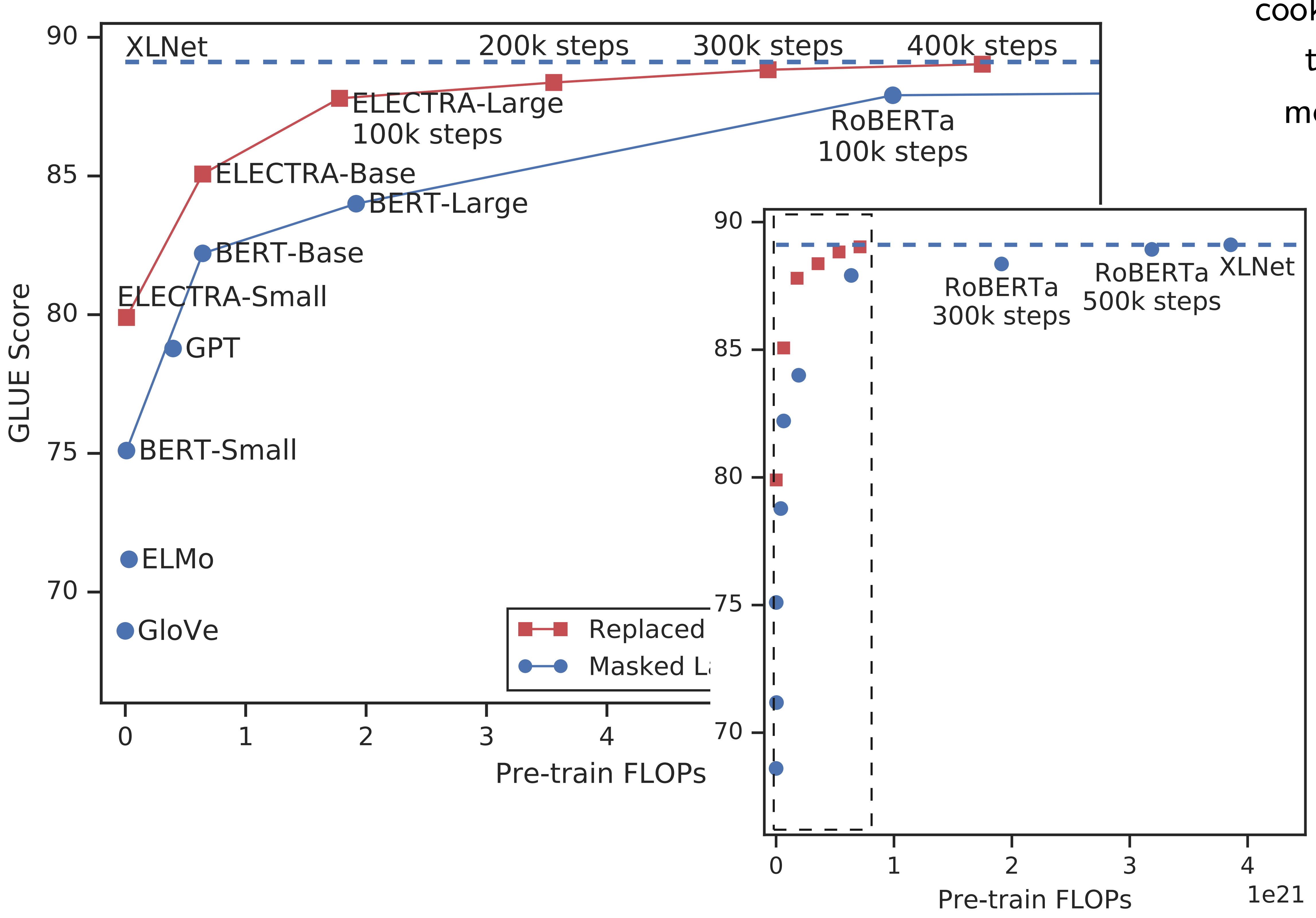
hyperparameter configurations

- Use fewer training examples
  - Change the data: Data optimization
  - Change the model: Sample efficiency

# ELECTRA: More efficient sampling for MLM

- **Key idea:** Different modeling paradigm (discrimination vs generation) enables more sample-efficient learning.




| Model | Train / Infer FLOPs | Speedup | Params | Train Time + Hardware | GLUE |
|---|---|---|---|---|---|
| ELMo | 3.3e18 / 2.6e10 | 19x / 1.2x | 96M | 14d on 3 GTX 1080 GPUs | 71.2 |
| GPT | 4.0e19 / 3.0e10 | 1.6x / 0.97x | 117M | 25d on 8 P6000 GPUs | 78.8 |
| BERT-Small | 1.4e18 / 3.7e9 | 45x / 8x | 14M | 4d on 1 V100 GPU | 75.1 |
| BERT-Base | 6.4e19 / 2.9e10 | 1x / 1x | 110M | 4d on 16 TPUv3s | 82.2 |
| ELECTRA-Small | 1.4e18 / 3.7e9 | 45x / 8x | 14M | 4d on 1 V100 GPU | 79.9 |
| 50% trained | 7.1e17 / 3.7e9 | 90x / 8x | 14M | 2d on 1 V100 GPU | 79.0 |
| 25% trained | 3.6e17 / 3.7e9 | 181x / 8x | 14M | 1d on 1 V100 GPU | 77.7 |
| 12.5% trained | 1.8e17 / 3.7e9 | 361x / 8x | 14M | 12h on 1 V100 GPU | 76.0 |
| 6.25% trained | 8.9e16 / 3.7e9 | 722x / 8x | 14M | 6h on 1 V100 GPU | 74.1 |
| ELECTRA-Base | 6.4e19 / 2.9e10 | 1x / 1x | 110M | 4d on 16 TPUv3s | 85.1 |

K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning.
ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators . ICLR 2020.

# Training efficiency

$$Cost(R) \propto E \cdot \boxed{D} \cdot H$$

obtaining result

processing one example

**dataset size**

hyperparameter configurations

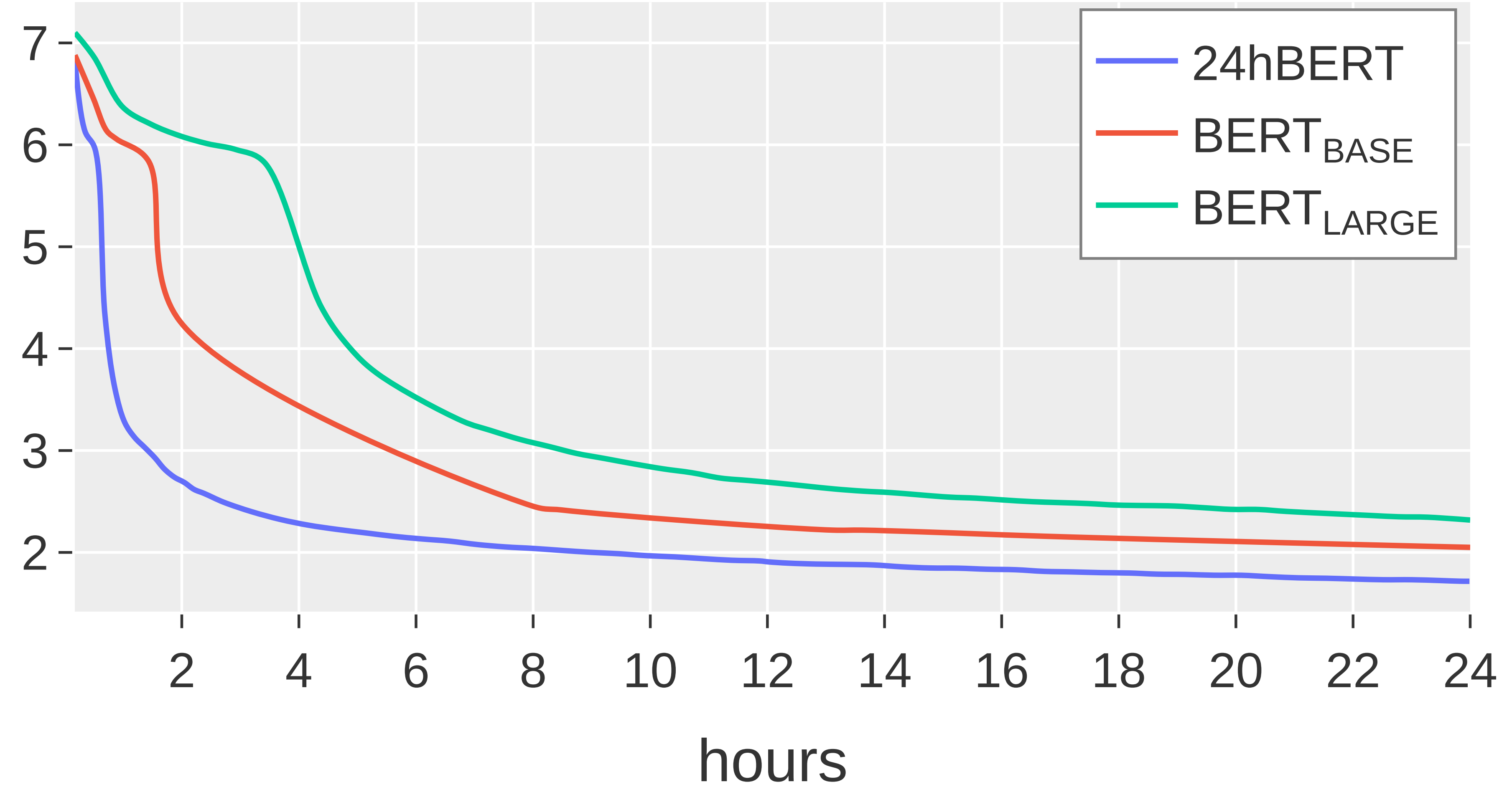
- Use fewer training examples
  - Change the data: Data optimization
  - Change the model: Sample efficiency
  - Change the optimization strategy

# 24h BERT

- **Key idea:** Efficient implementation & careful hyperparameter selection enables LM training on an "academic budget" of 8 V100 GPUs.

## Recipe:

- **Input resolution**
  - Short sequences
  - Single-sequence training
- **Implementation / framework:**
  - DeepSpeed
  - Sparse token prediction
  - Fused implementations
  - Avoiding disk I/O




- **Optimization:**
  - Training larger models
  - Large batch sizes
  - Large learning rates
  - Short warmup
  - Synchronizing schedule with time budget



P. Izsak, M. Berchansky, O. Levy. How to Train BERT with an Academic Budget. EMNLP 2021.

# Cramming

- **Key idea:** Optimize even further to enable training on a single GPU.

Data modifications:
Remove hard-to-compress examples;
Curriculum: likely sequences first (unigram).

Training modifications:
High learning rate; Big batches;
Drop dropout.

| | MNLI | SST-2 | STSB | RTE | QNLI | QQP | MRPC | CoLA | GLUE |
|---|---|---|---|---|---|---|---|---|---|
| crammed BERT | **83.9 / 84.1** | **92.2** | 84.6 | 53.8 | **89.5** | **87.3** | 87.5 | **44.5** | **78.6** |
| + original data | 82.2 / 82.7 | 92.0 | 83.6 | 49.8 | **89.5** | 87.0 | 85.9 | 42.5 | 77.3 |
| + original train | 50.0 / 50.4 | 80.7 | 13.7 | 52.0 | 59.8 | 65.1 | 73.2 | 7.2 | 50.2 |
| + original arch. | 35.4 / 35.2 | 49.1 | - | 52.7 | 49.5 | 0.0 | 0.0 | 0.0 | 27.7 |
| + minimal train mod. | 81.9 / 82.6 | 91.4 | **85.5** | **54.9** | 88.2 | 87.0 | **88.4** | 43.6 | 78.1 |
| + minimal arch. mod. | 83.2 / 83.5 | 91.7 | 82.0 | 52.0 | 88.9 | 86.8 | 83.6 | 38.3 | 76.7 |

Architecture modifications:
Layer norm inside residual (pre-norm; Xiong et al. 2020).



J. Geiping, T. Goldstein. Cramming: Training a Language Model on a Single GPU in One Day. arXiv 2022.

# Training efficiency

$$Cost(R) \propto E \cdot D \cdot H$$

obtaining result

processing one example

dataset size

hyperparameter configurations

- Make forward faster
- Make backward faster

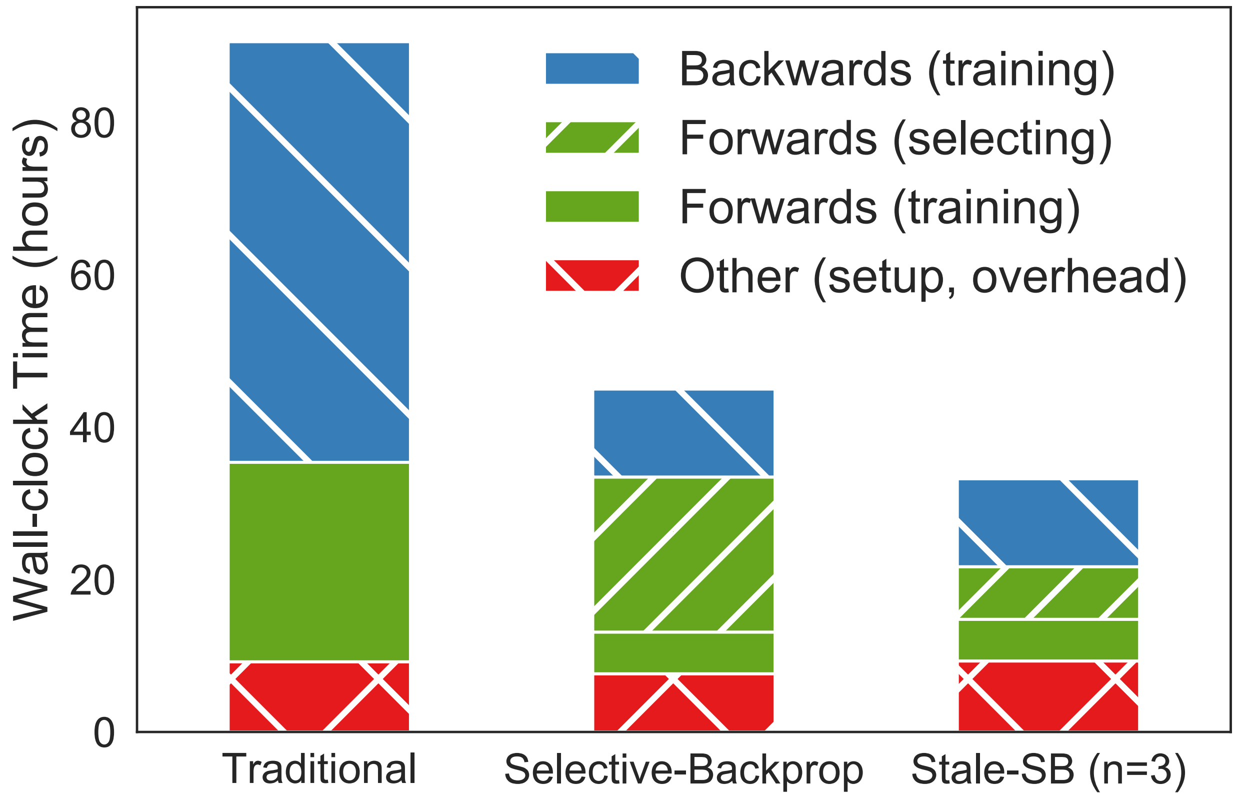



Figure from: Jiang, et al. Accelerating Deep Learning by Focusing on the Biggest Losers. 2019.

- Use fewer training examples
  - Change the data: Data optimization
  - Change the model: Sample efficiency
  - Change the optimization strategy