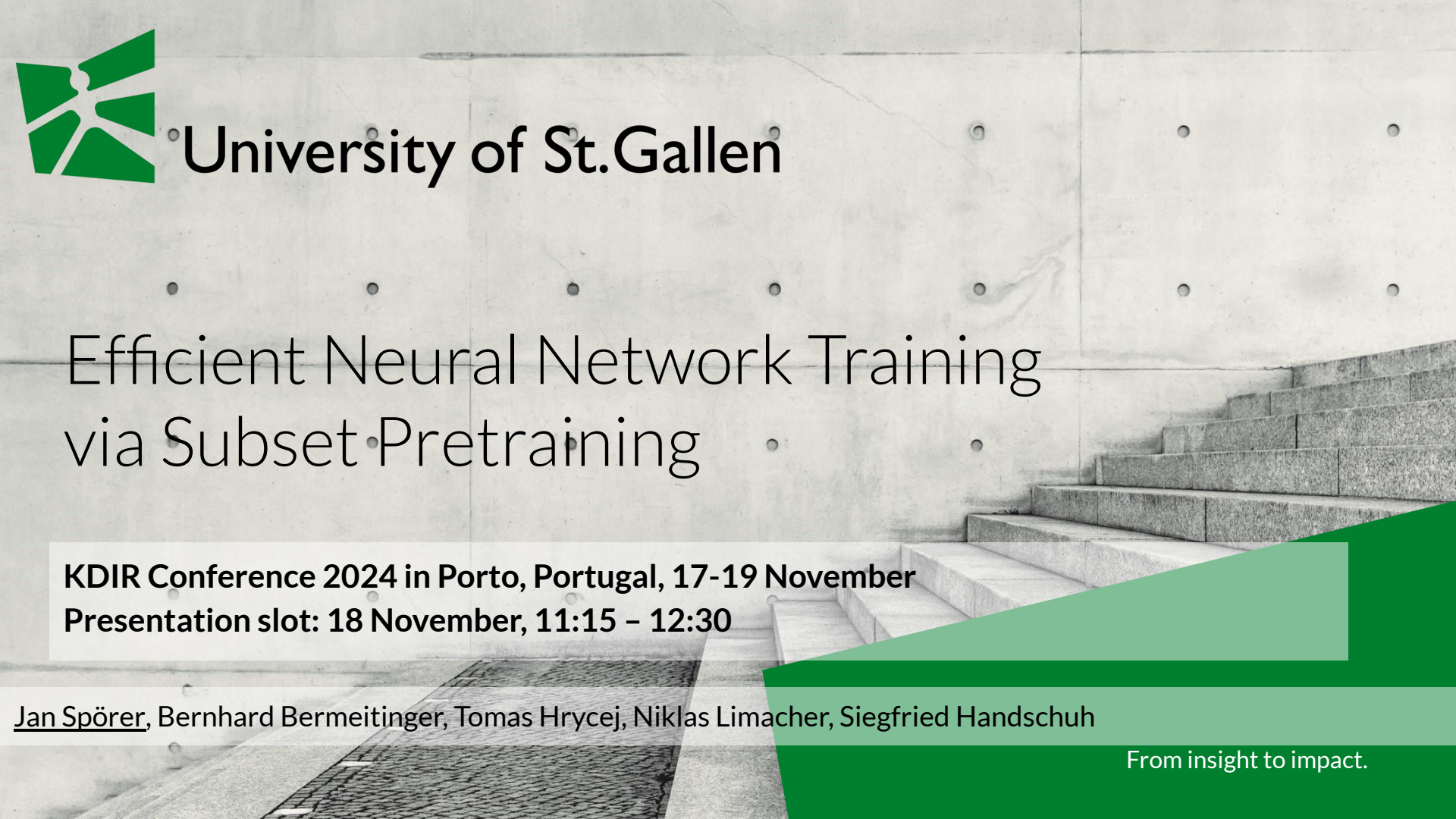


University of St.Gallen

Efficient Neural Network Training via Subset Pretraining

KDIR Conference 2024 in Porto, Portugal, 17-19 November
Presentation slot: 18 November, 11:15 – 12:30

Jan Spörer, Bernhard Bermeitinger, Tomas Hrycej, Niklas Limacher, Siegfried Handschuh

From insight to impact.

Motivation: What are *efficient subsets*?

- Do we really need to use different batches? What happens when we keep using the same batch, and only later batch from the whole training set?
- Shortcomings of batch-oriented approach:
  - Violation of stochastic approximation theory, esp. the condition of diminishing step sizes (eq. 4 of the paper)
  - Gradient approximation not sufficient
  - Convexity around minimum not exploited
- "Good" approximation of gradient not sufficient
- Hypothesis: Optimum of subset loss is close to the optimum of the training set loss and is within the convex region of the loss

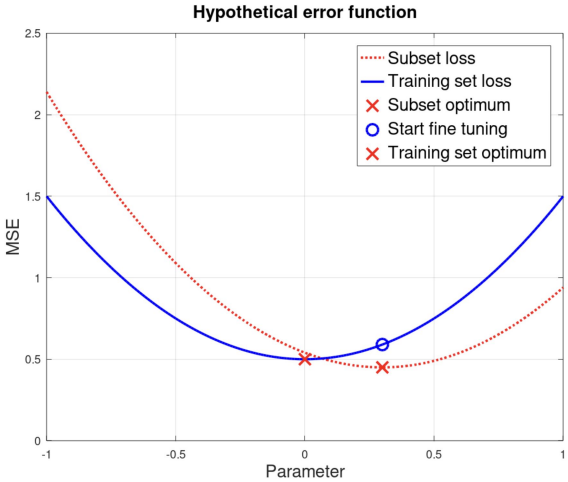


Figure 1 from the paper

Data: MNIST, CIFAR-10 and CIFAR-100

- Vision datasets MNIST and CIFAR
- CIFAR-10 contains the same images as CIFAR-100, but CIFAR-100 with more fine-grained labeling
- MNIST and CIFAR-10 augmented nine times (10x increase in data) to increase the number of training samples from 60’000 to 600’000 and from 50’000 to 500’000, respectively
- Augmentation to ensure a determination ratio of at least one (explained in detail on a later slide)

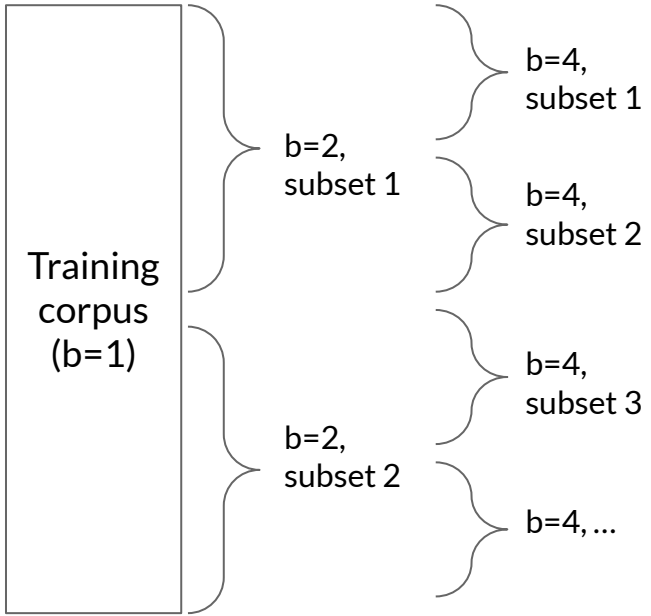
Exemplary constraint calculation:
(important for the determination)

No. of constraints = 600’000 samples x
10 labels
= 6’000’000

“Efficient Neural Network Training via Subset Pretraining”, Jan Spörer, Bernhard Bermeitinger, Tomas Hrycej, Niklas Limacher, Siegfried Handschuh, 2024, KDIR

Methods: Substitute the training set by a subset

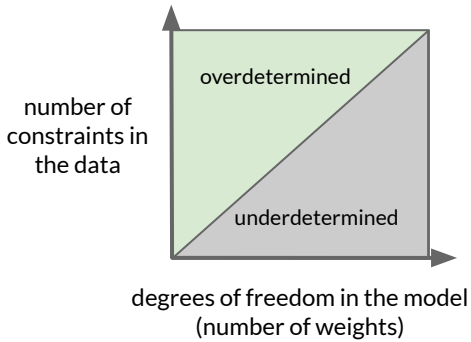
- Subset selection: Training set divided into b subsets (b ∈ {2, 4, 8, 16, 32, 64, 128})
  - We ran eight different configurations: full dataset + seven subsets: 2 … 128
  - We ran each of the seven subset configurations multiple times
- Two-phase training
  - Subset pre-training (1000 epochs)
  - Fine-tuning on full training set (100 epochs)



Methods: Overdetermination ratio

The overdetermination ratio (Q = KM/P) informs us of an appropriate subset size

- Defines a minimum data threshold for a given model size
- We used a model with ~100’000 parameters. For the overdetermination ratio to exceed 1, we should not split more than b=4 (if data not augmented)



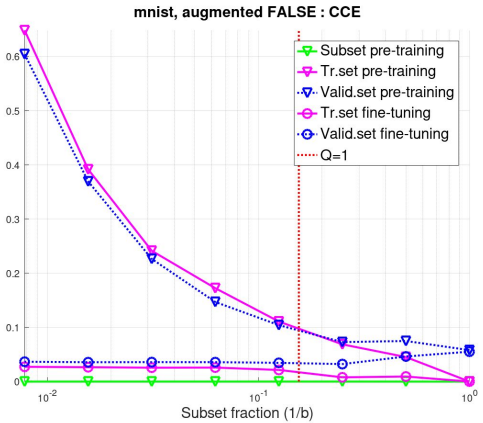
Dataset constraints (without augmentation)

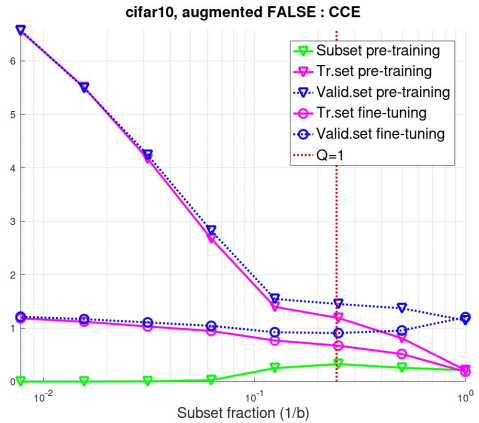
- MNIST: 60’000 training samples x 10 class labels = 600’000
- CIFAR-10: 50’000 training samples x 10 class labels = 500’000
- CIFAR-100: 50’000 training samples x 100 class labels = 5’000’000

“Efficient Neural Network Training via Subset Pretraining”, Jan Spörer, Bernhard Bermeitinger, Tomas Hrycej, Niklas Limacher, Siegfried Handschuh, 2024, KDIR

Results: Overdetermination ratio as the optimal cut-off

Result preview (figs 2+6) that illustrates
the impact of Q (red dotted vertical line):





Underdetermined

Underdetermined

The training set loss quickly approaches the subset loss as the subset size approaches the determination threshold. Generalization is only good in overdetermined settings.

Summary: Subsets are often sufficient

- Subset training leads to comparable results vs. conventional training
- Overdetermination ratio (Q) should preferably exceed unity
- Even small subsets can be representative if Q > 1
- Computational resource savings of 90% or more possible; pre-training sufficient, as shown experimentally
- Future work: Test on larger datasets, investigate second-order optimization algorithms