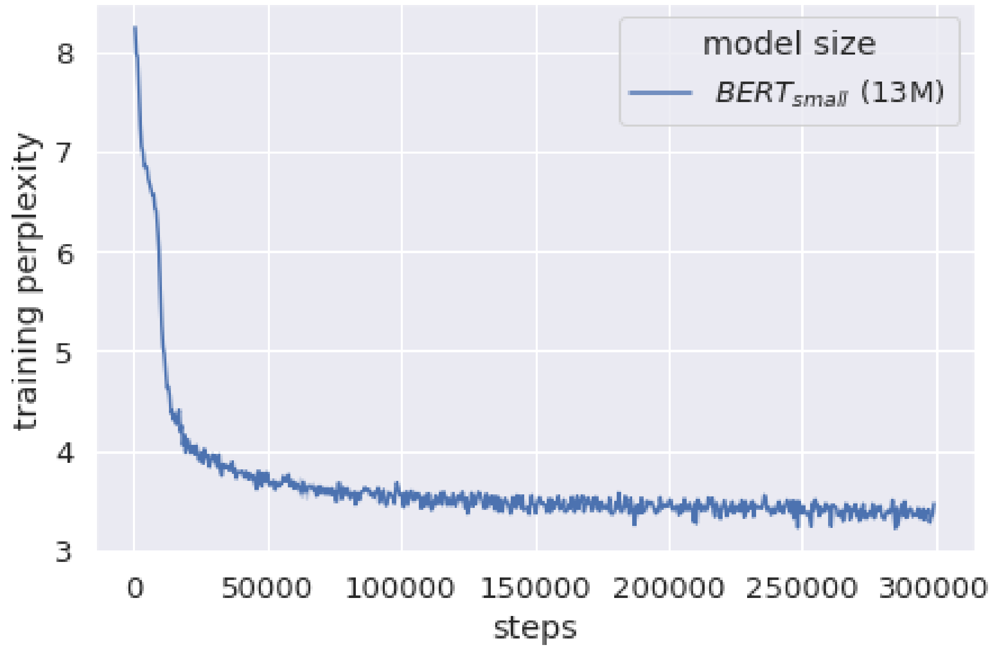
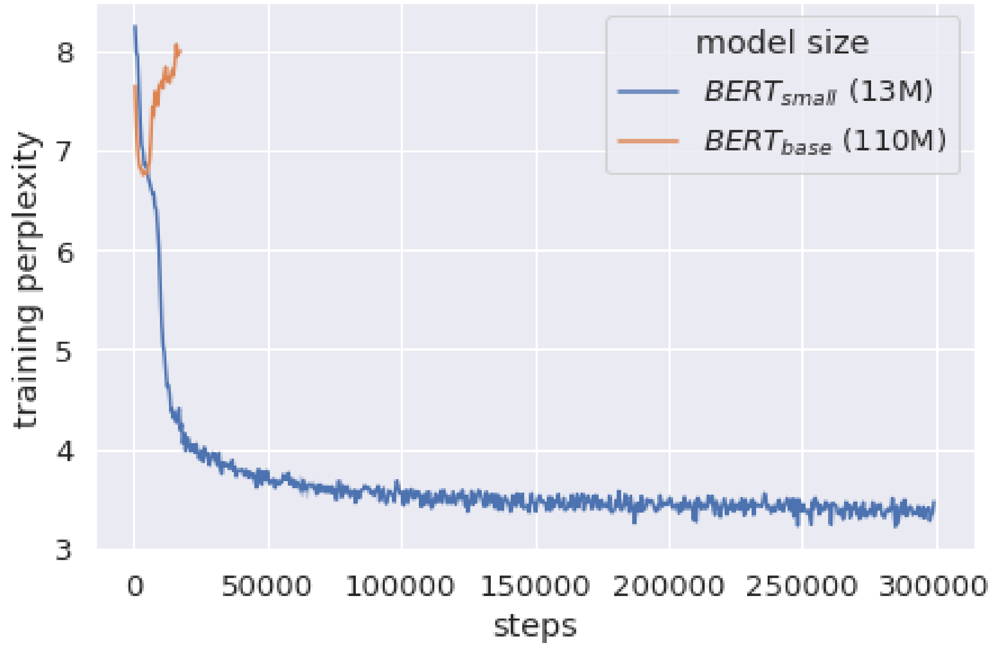
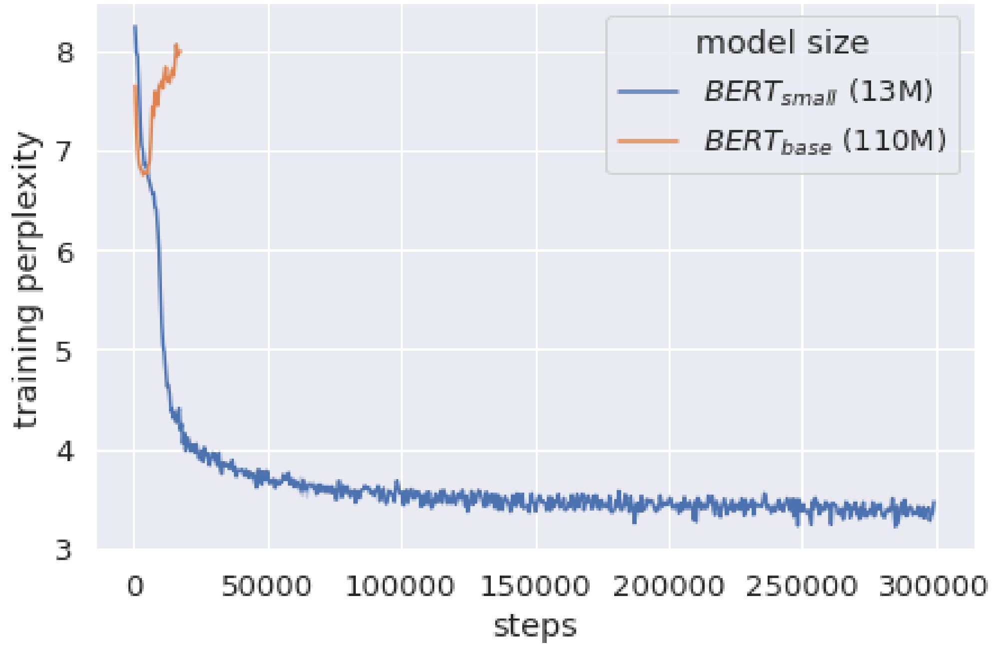
Explore with a small model



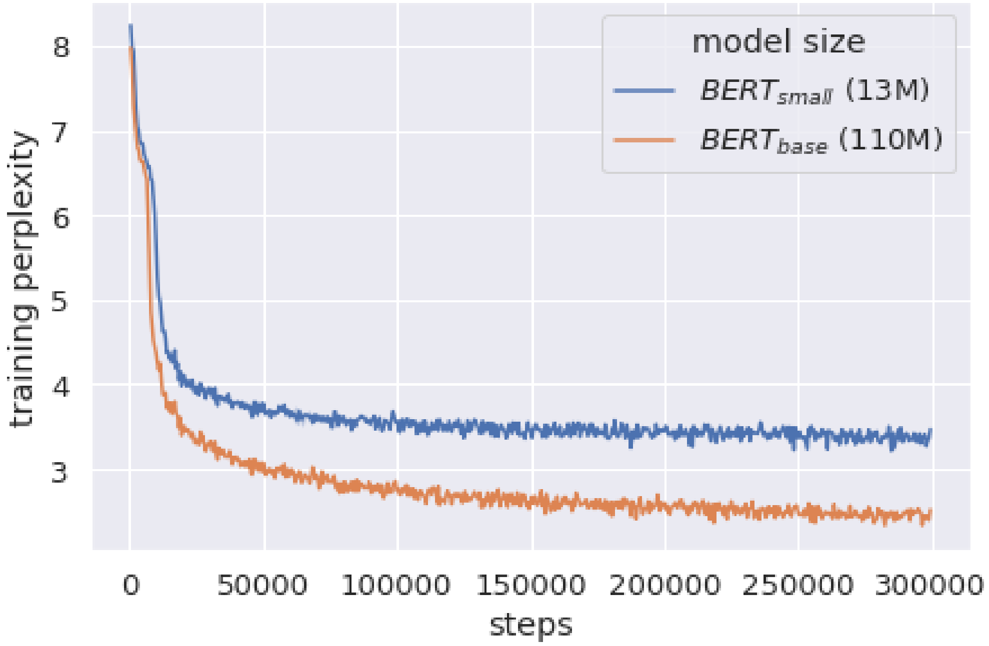
Model fails to train when scaled up with the same hyperparameters



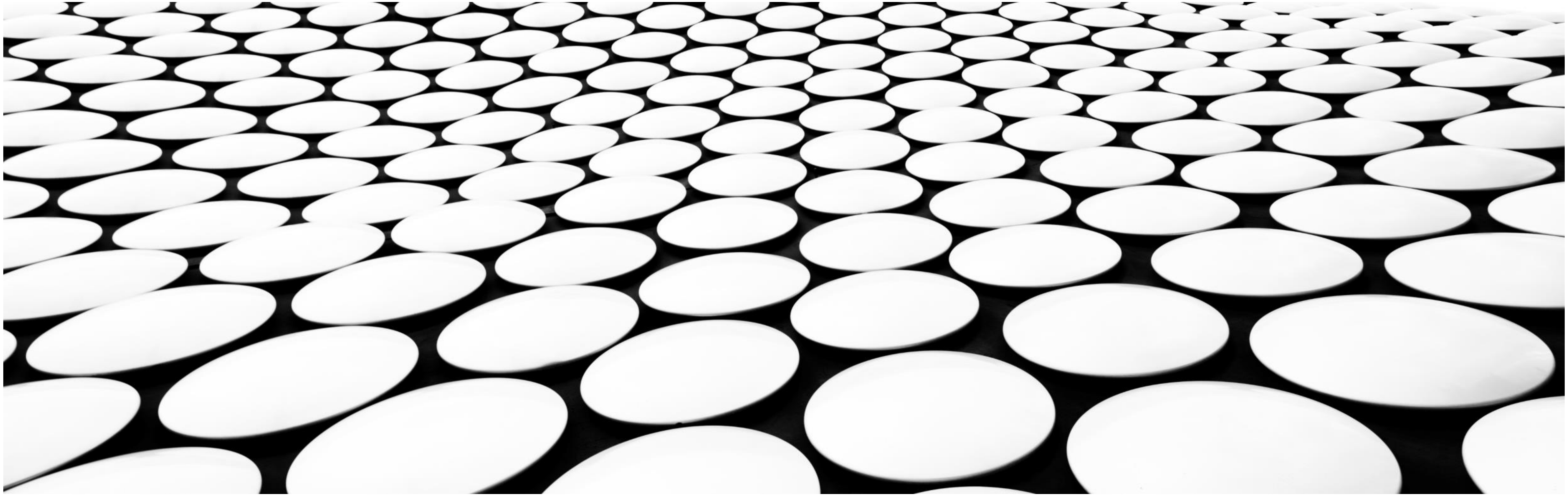
What *This Work* Allows You To Do



Before



After



Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer

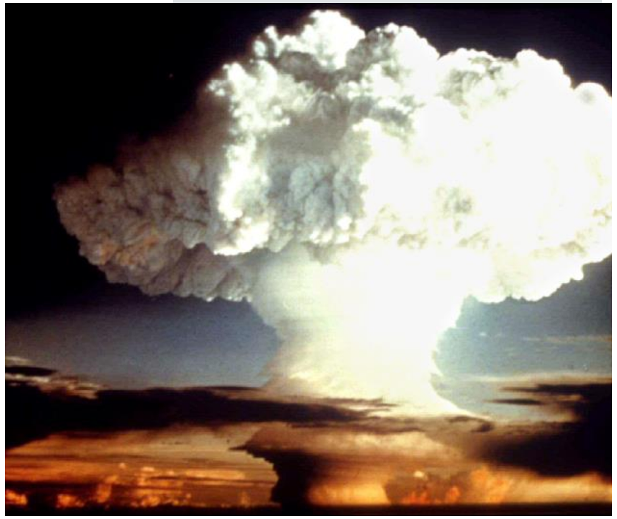
Greg Yang

in Collaboration with Edward Hu, Igor Babuschkin, Szymon Sidor, David Farhi, Nick Ryder, Jakub Pachocki, Xiaodong Liu, Weizhu Chen, Jianfeng Gao

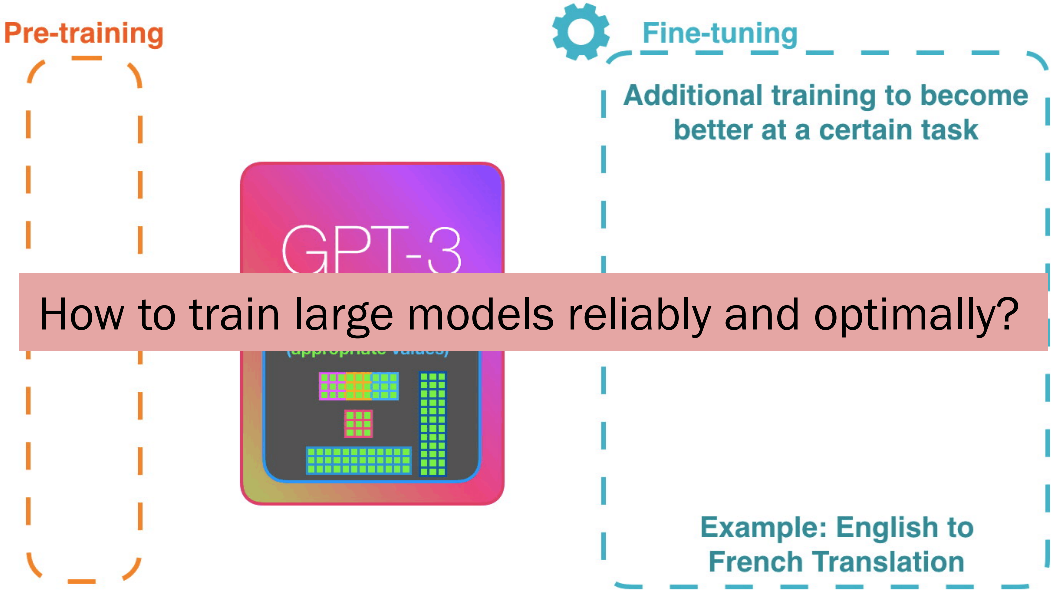
Microsoft OpenAI

WHAT DO THESE HAVE IN COMMON?

Manhattan Project

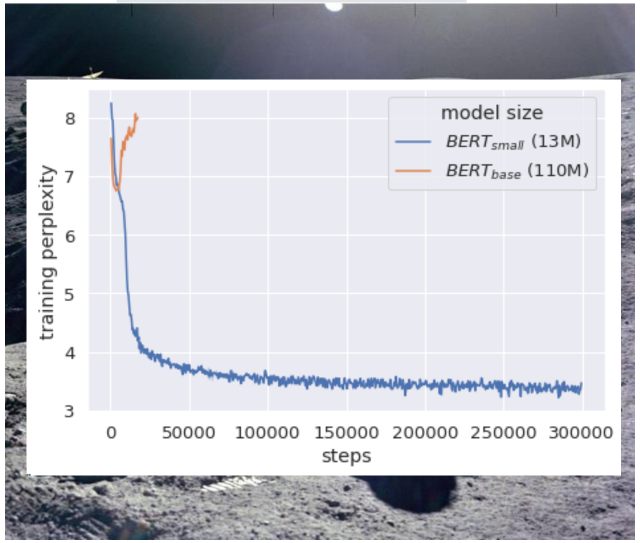


Large pretrained language/vision models



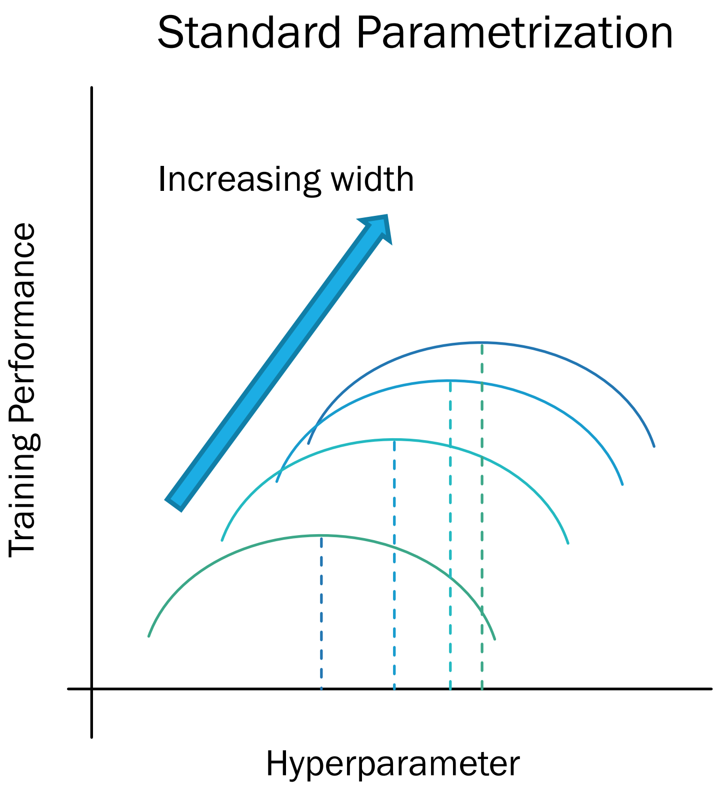
How to train large models reliably and optimally?

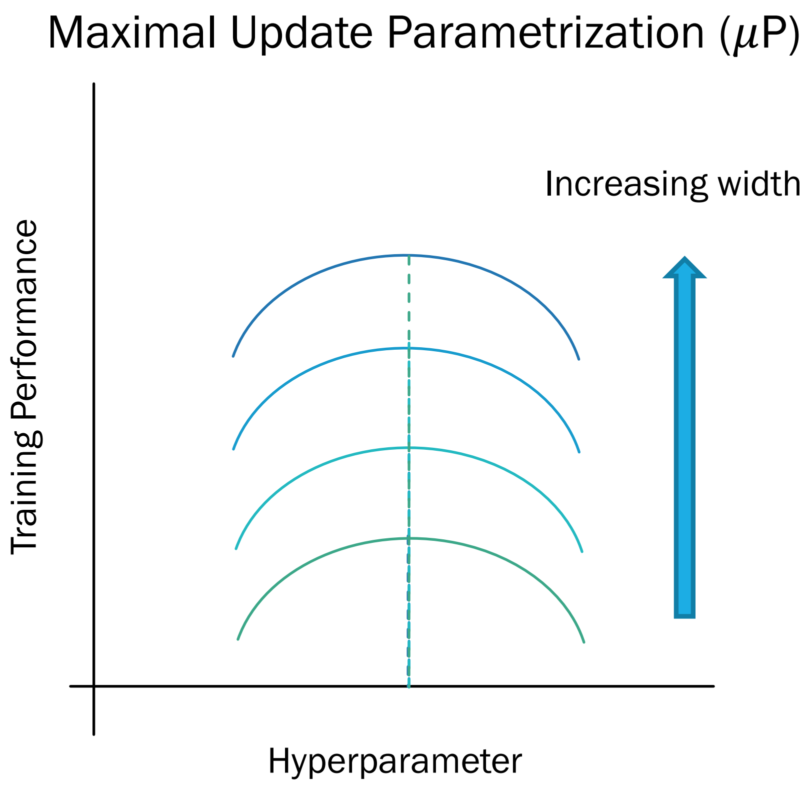
Space Program



- Revolutionary achievements, paradigm shifts of their times
- Started races between nation-states
- Each empirical test is very expensive
- Require extensive theoretical calculation first before launching any empirical test

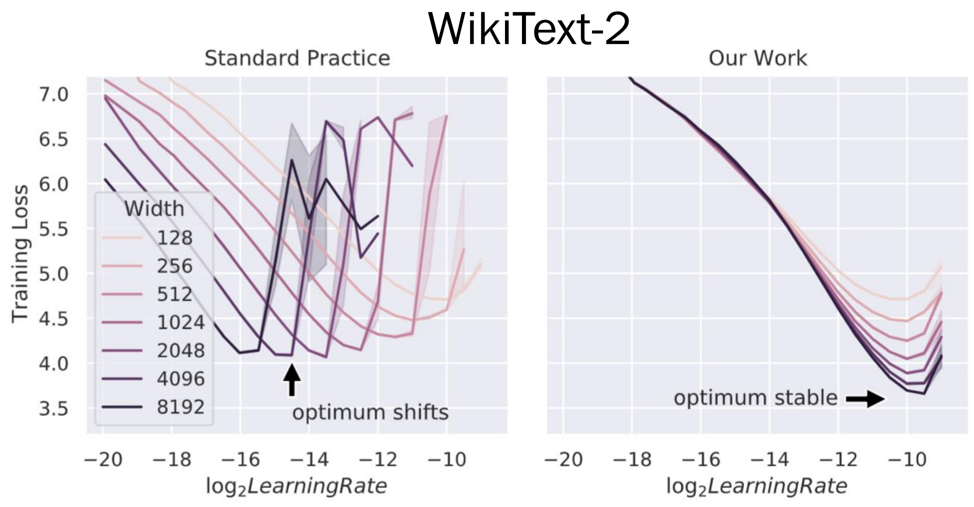
μTransfer in a Gist

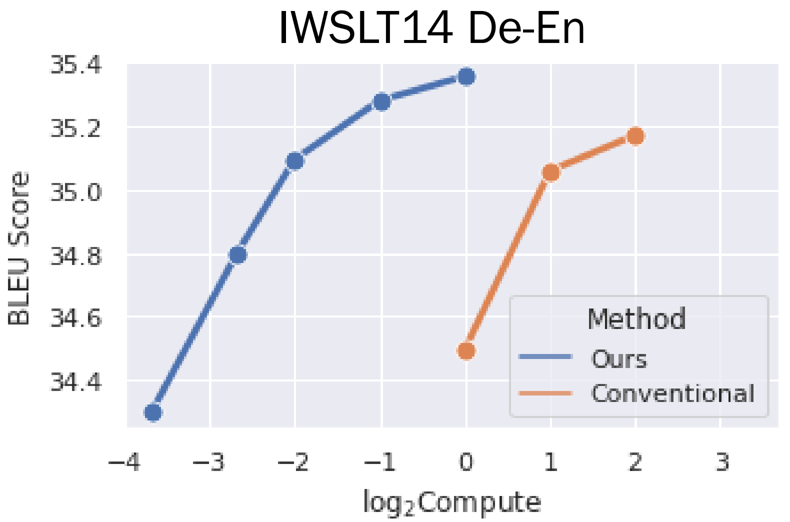


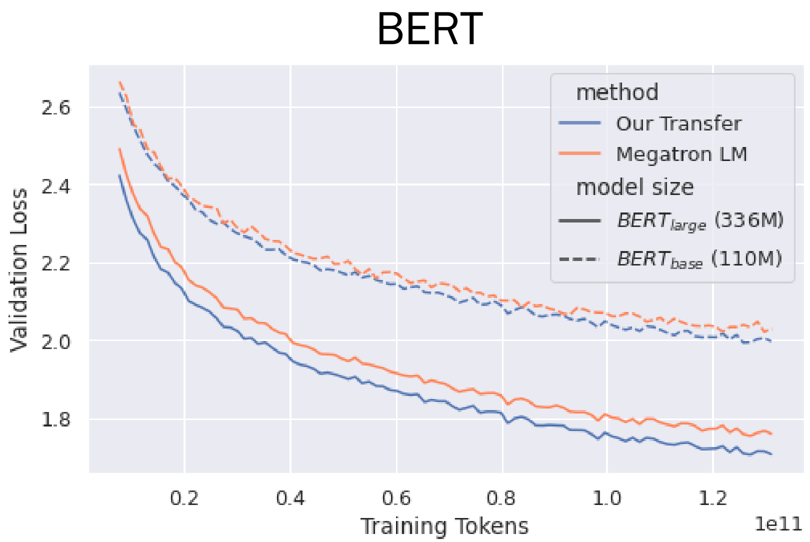


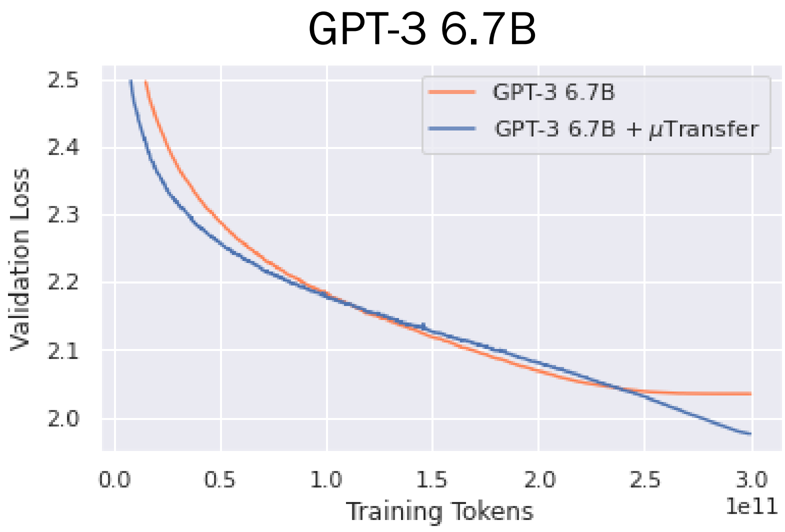
"Transfer" = optimal hyperparameter remains stable with model size

Key Empirical Results



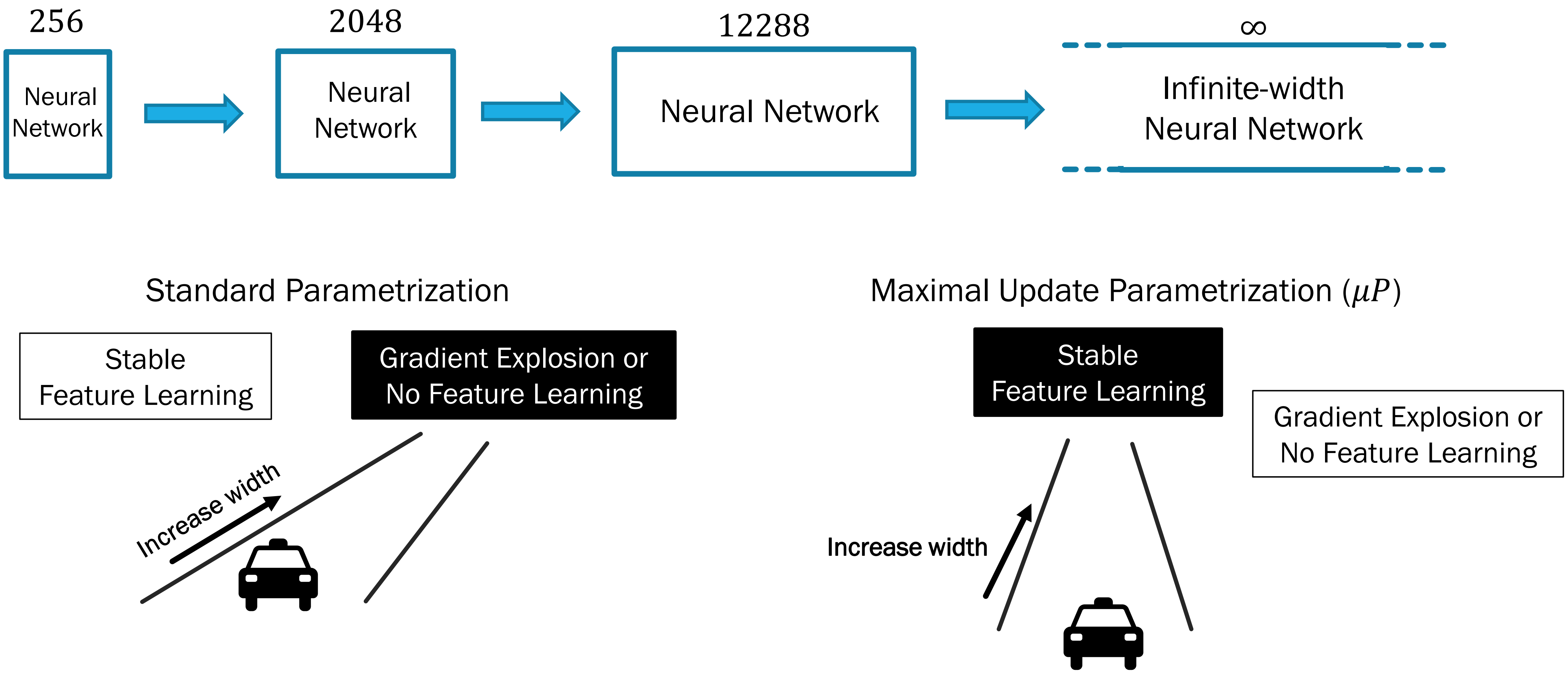




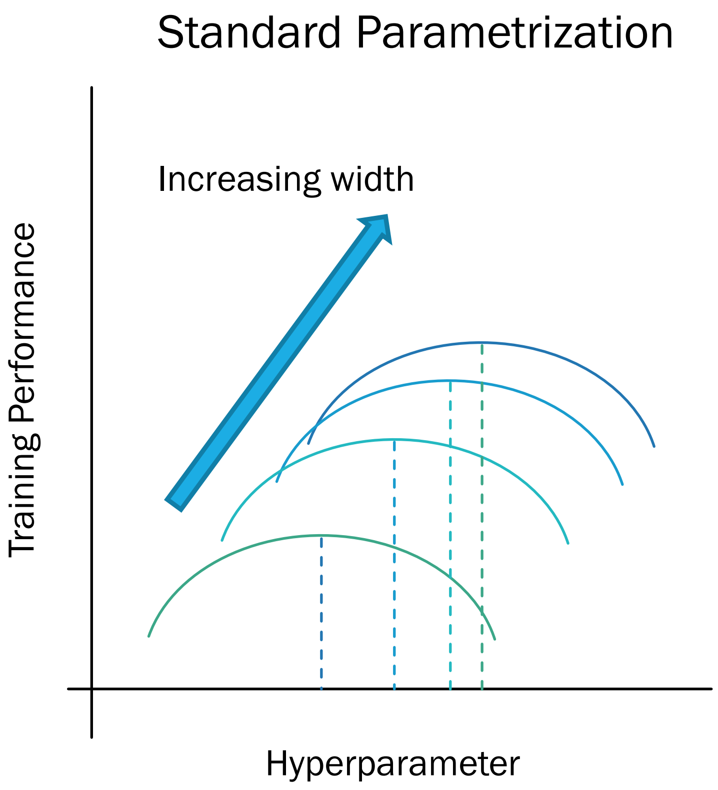


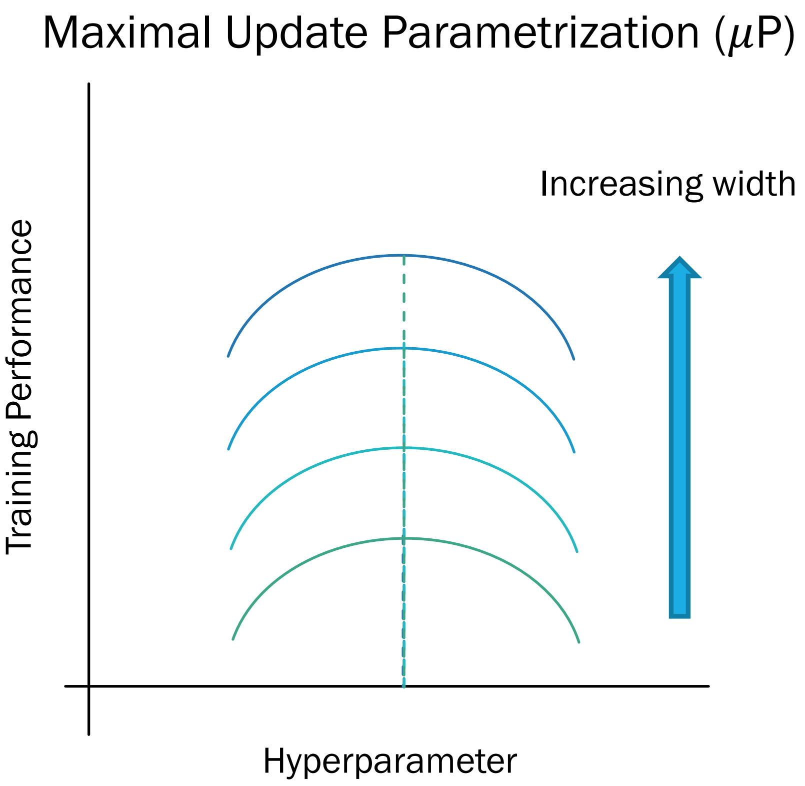
Theoretical Foundation

Neural Network Infinite-width Limits



μTransfer in a Gist





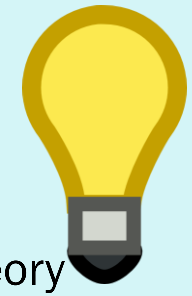
"Transfer" = optimal hyperparameter remains stable with model size

Big O or Θ suppress constants not dependent on width n, including input and output dim

Desiderata for a Good Parametrization

Any time during initialization or training:

1. Every (pre)activation vector should have $\Theta(1)$-sized coordinates
2. Neural network output should be $O(1)$
3. All parameters should be updated as much as possible (in terms of scaling in width) without leading to divergence.



- Given these desiderata, deriving μP ~= deriving the renormalizability of an effective field theory
- i.e. dimension analysis in width (compared to dimensional analysis in cutoff)

Maximal Update Parametrization (μP)

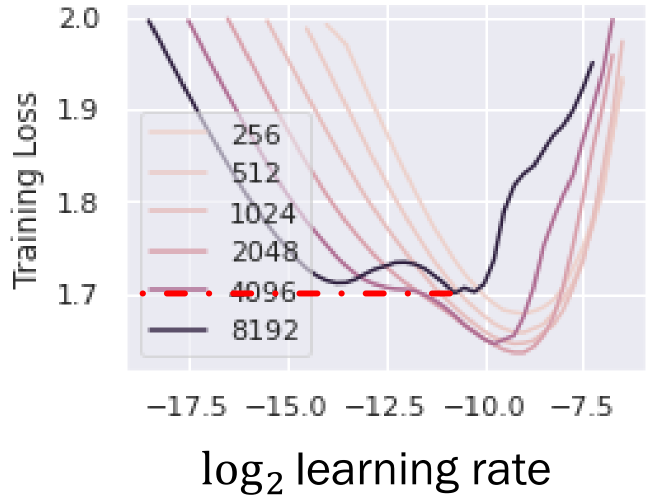
	Input weights & all biases	Output weights	Hidden weights
Init. Var.	$1/\text{fan\_in}$	$1/\text{fan\_in}^2$ $(1/\text{fan\_in})$	$1/\text{fan\_in}$
SGD LR	$\eta \cdot \text{fan\_out}$ (η)	$\eta/\text{fan\_in}$ (η)	η
Adam LR	η	$\eta/\text{fan\_in}$ (η)	$\eta/\text{fan\_in}$ (η)

Note: focus on scaling with fan_in or fan_out; everything else is a tunable constant

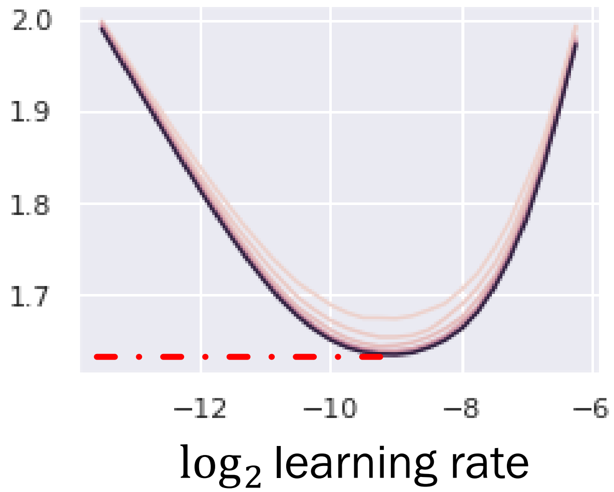
Empirical Evidence

2-hidden Layer MLP on CIFAR-10

Standard Parametrization

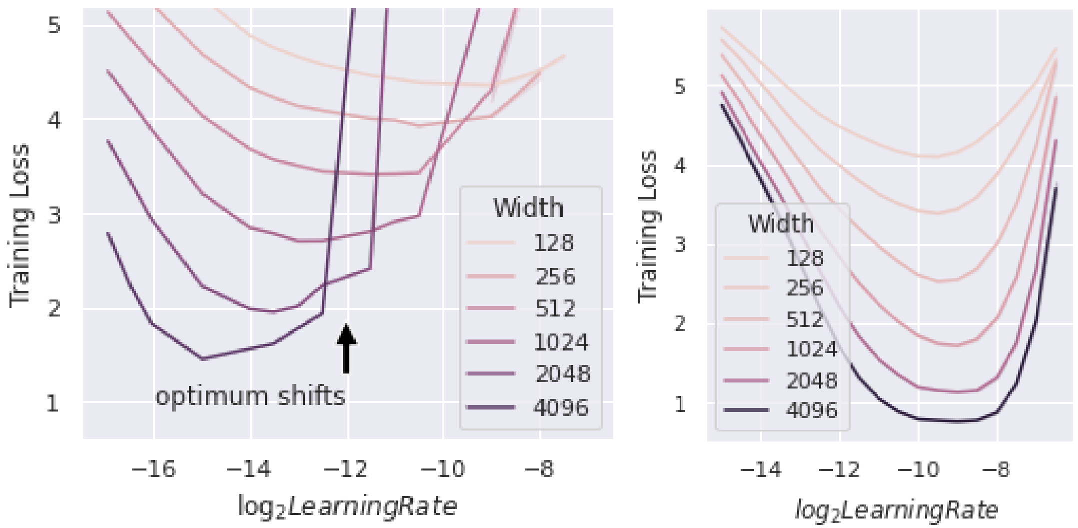


Max Update Parametrization

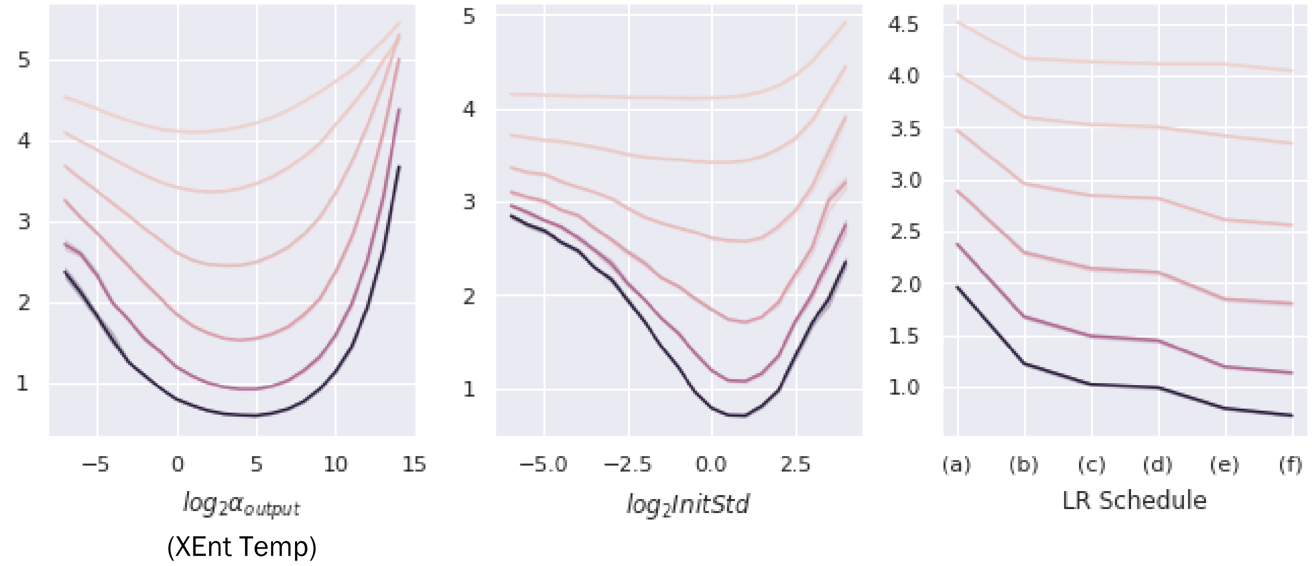


Transformer on Wikitext-2

Standard Parametrization



Max Update Parametrization



Tuning BERT with μTransfer

Model	# of params	Tuning cost (V100 yr)	Our Speedup
BERT SMALL	13M	1.8	1x
BERT BASE	110M	7.2	4x
BERT LARGE	336M	40	22x

Step 1:

Parameterize BERT in μP

Step 2:

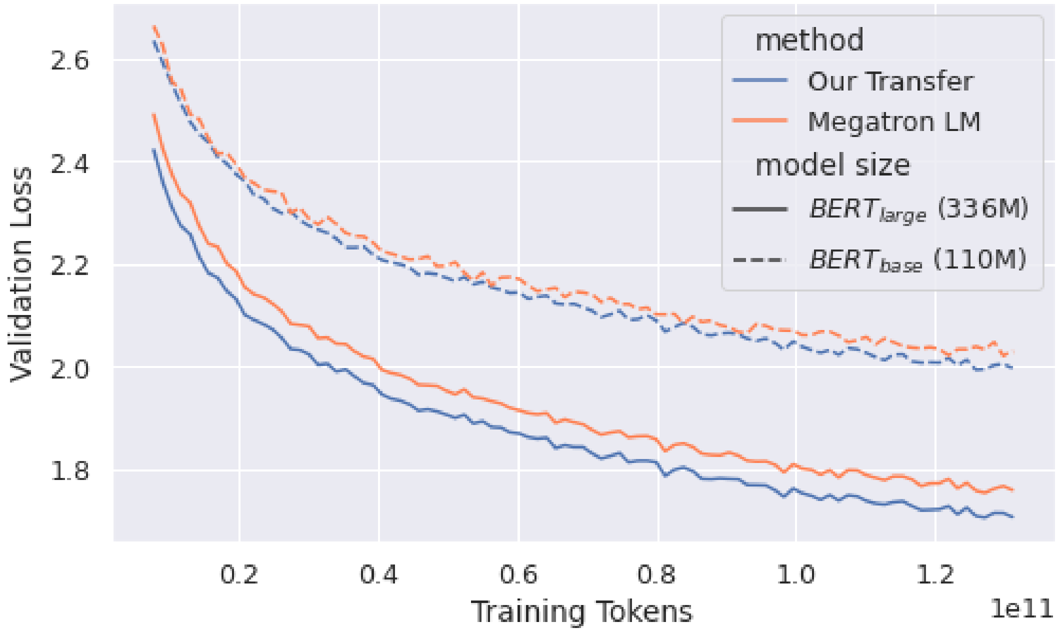
Tune hyperparameters on BERT SMALL via random search (256 combinations)

Step 3:

Copy the best hyperparameter combination to BERT BASE and BERT LARGE

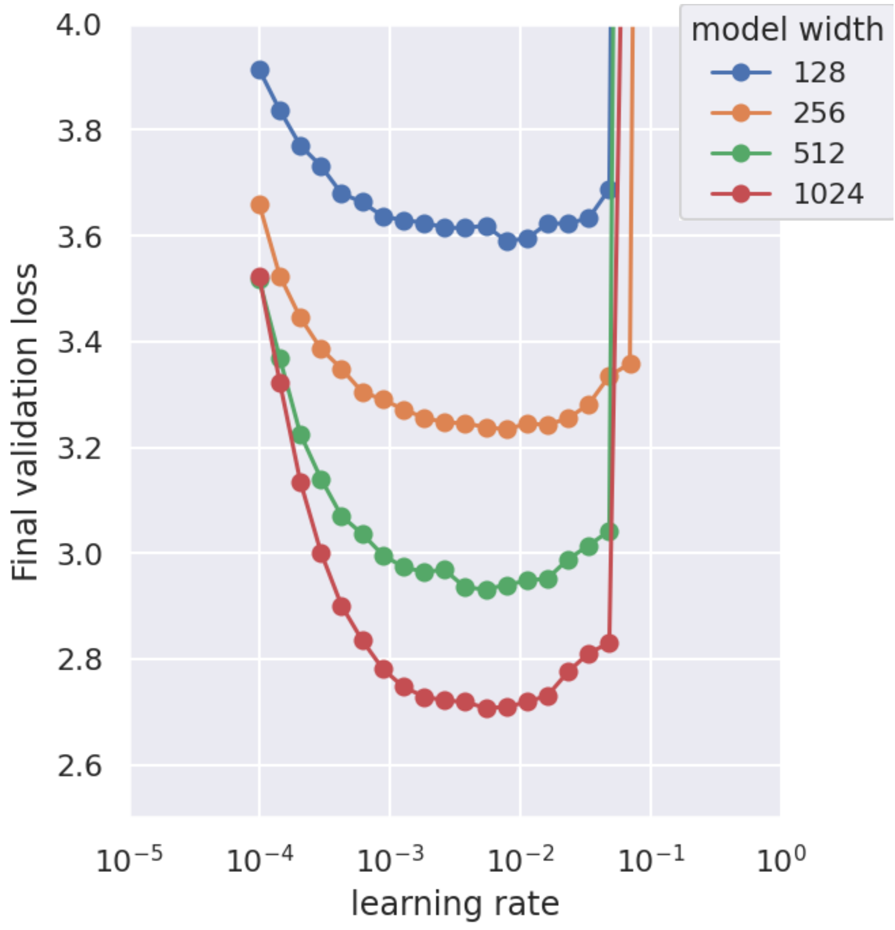
✓ Tune once, use for a family of models

✓ Only run the large models once

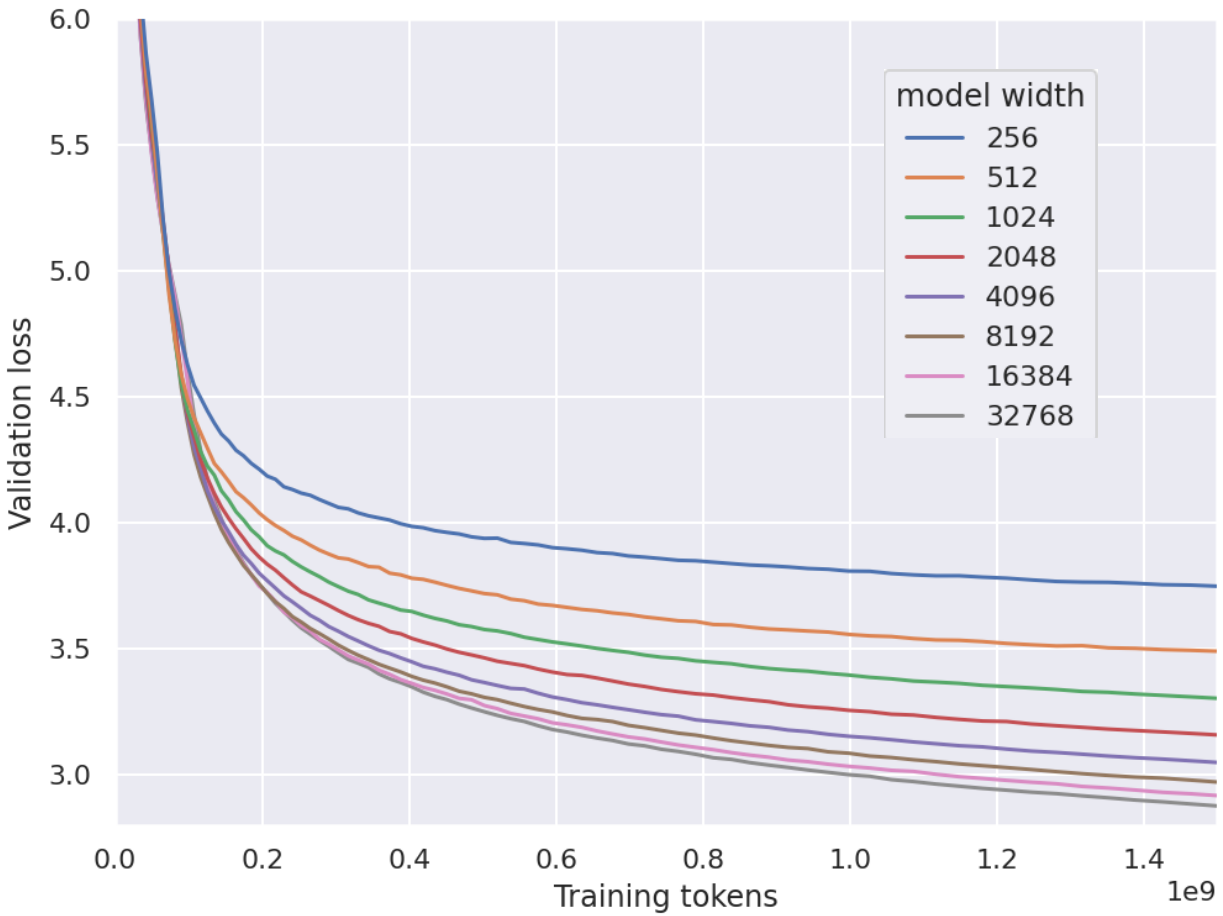


OpenAI GPT-3 Family + μP

Hyperparameter Optimum is Stable



Wider is Always Better Given the Same HPs



OpenAI GPT-3 6.7B + μTransfer

μTransfer Outperforms the Heuristics Used in Brown et al. 2020

Total tuning compute budget is only 7% of training budget!!!

Connection with Physics

	Input weights & all biases	Output weights	Hidden weights
Init. Var.	$1/\text{fan\_in}$	$1/\text{fan\_in}^2$ $(1/\text{fan\_in})$	$1/\text{fan\_in}$
SGD LR	$\eta \cdot \text{fan\_out}$ (η)	$\eta/\text{fan\_in}$ (η)	η
Adam LR	η	$\eta/\text{fan\_in}$ (η)	$\eta/\text{fan\_in}$ (η)

ANALOGY: LARGE MODEL TRAINING VS EFFECTIVE FIELD THEORY

Abbrev: HP = hyperparameter

Large Model Training

- Model size, or other compute HP like training time
- Non-compute HP, like learning rate
- Parametrization
- Trained model is a function of
  - Compute HP: model size, training time, batch size, etc
  - Non-compute HP: learning rate, weight decay, etc
- Model predicts next word of sentence/image label/etc
- Objective: find best HP for a given size to train model to reproduce human language and vision as closely as possible
- “Optimal” hyperparameters

Effective Field Theory

- Momentum/energy cutoff
- Coupling constants
- Theory skeleton (with unspecified coupling constants)
- A concrete effective field theory is a function of
  - Momentum/energy cutoff
  - Instantiations of “bare” coupling constants
- Theory predicts fundamental physics of our universe
- Objective: find coupling constants that reproduce experimental results as closely as possible
- “correct” coupling constants

NOW CONSIDER WIDTH AS THE MEASURE OF MODEL SIZE

Abbrev: HP = hyperparameter

Large Model Training

- Model width
- Infinite-width limit
- Hyperparameter transfer

$$HP' = F(HP, width, width')$$

- Parametrization admitting hyperparameter transfer
  - Optimal HPs have infinite-width limits

Effective Field Theory

- Momentum/energy cutoff
- Ultraviolet limit
- Renormalization

$$coupling' = F(coupling, cutoff, cutoff')$$

- Renormalizable theory
  - "physical" coupling constants have ultraviolet limits

NOW CONSIDER WIDTH AS THE MEASURE OF MODEL SIZE

Large Model Training

- Goal:
  - Train large models reliably and optimally using parametrizations admitting hyperparameter transfer
- Example
  - Parametrization: Maximal Update Parametrization (μP)
  - Infinite-width limit: the feature learning limit (aka μ-limit)
- Counterexample
  - Parametrization: Neural Tangent (NT) parametrization
  - Infinite-width limit: Neural Tangent Kernel (NTK) limit
  - Failure: does not transfer optimal hyperparameters

Effective Field Theory

- Goal (?):
  - Come up with theory that describes nature at any energy cutoff
- Example
  - Theory: QCD
  - Ultraviolet limit: asymptotic freedom
- Counterexample
  - Theory: classical electromagnetism
  - Ultraviolet limit: itself (?)
  - Failure: ultraviolet catastrophe

OPEN QUESTIONS

- μP solves the transfer problem for width in a principled way. Can we do it for all other compute hyperparameters?
  - Naïve transfer seems to work OK empirically, but as we go to larger scales, likely they will break down
  - Analogy in physics: we have renormalizable QCD but are looking for a renormalizable theory unifying all fundamental forces
- How can techniques from physics, like effective field theory, help?

WHY DOES ONE CARE ABOUT HYPERPARAMETER TRANSFER?

- High impact
  - Large model training is a modern space race
  - Highly heated race between large corporations and nation-states
  - These large neural networks are the closest we have to human intelligence
  - They can significantly reshape everyone's lives in the upcoming years
- High leverage (for theorists)
  - Each model training run can cost $10+ million dollars
  - so theorists are absolutely crucial here to provide guidance, as empirical approaches are absurdly expensive
- Distillation of theory
  - The current field of theoretical deep learning has a lot of "spurious explanations" with no predictive power
  - The high stakes mean that these fluff theories will be weeded out quickly
    - Akin to testing physical predictions using data from LHC
  - In particular, the *correct* limits of neural networks should necessarily admit HP transfer
    - So anything based on NTK should not be correct

PAPER