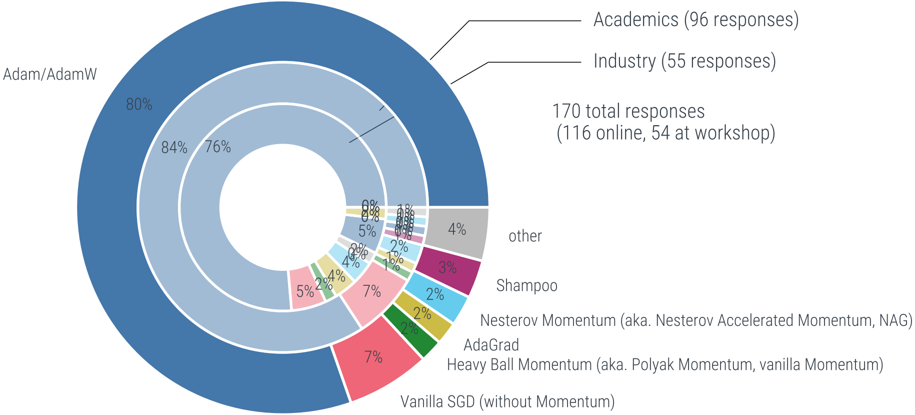
Results of the Poll

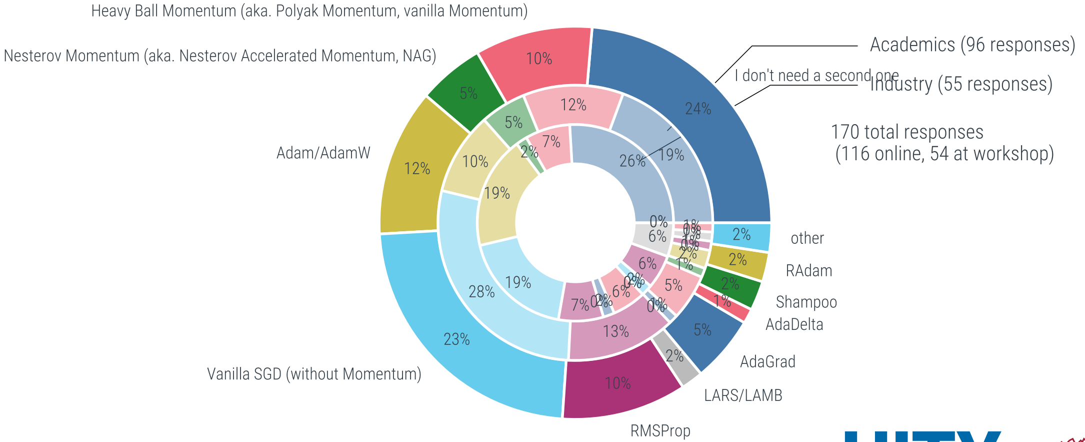
Note: Answers labelled “academic” are mostly PhD students

Which optimizer (i.e. update rule) do you try out first to train a neural network?



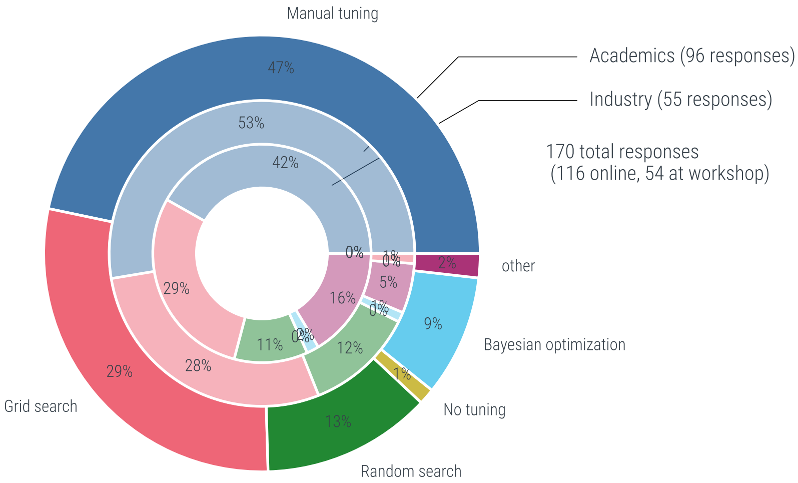
170 total responses
(116 online, 54 at workshop)

Which optimizer (i.e. update rule) do you try out second to train a neural network? This is not a duplicate question! This question is about your second choice.



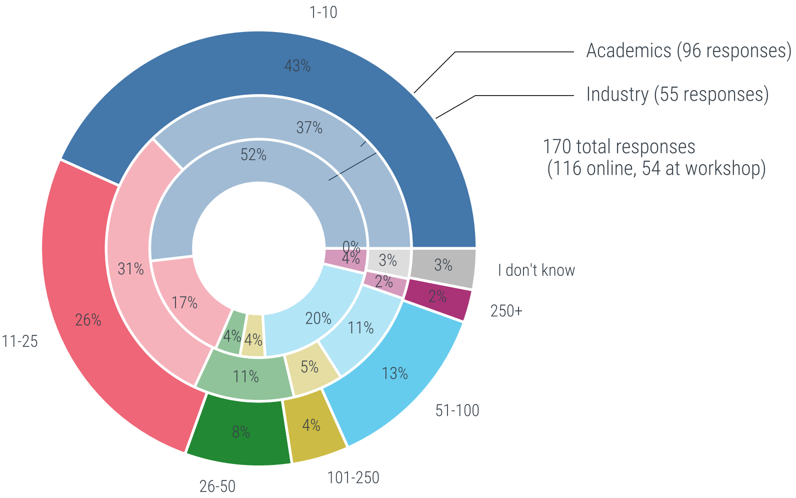
170 total responses
(116 online, 54 at workshop)

How do you tune your hyperparameters?



170 total responses
(116 online, 54 at workshop)

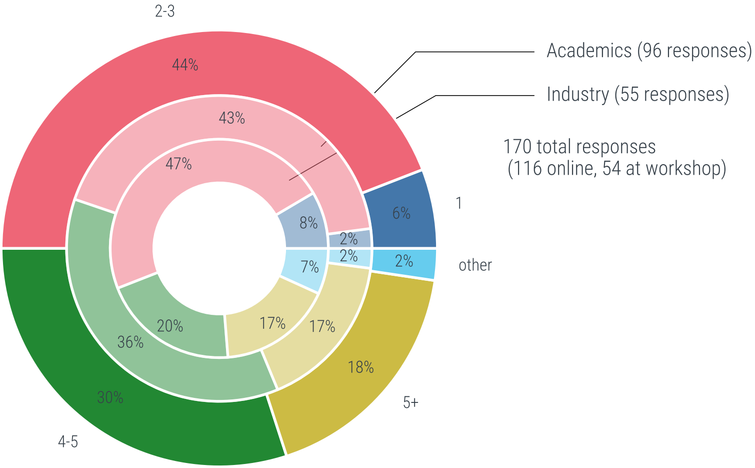
How many tuning trials do you use? Just use your best estimate.



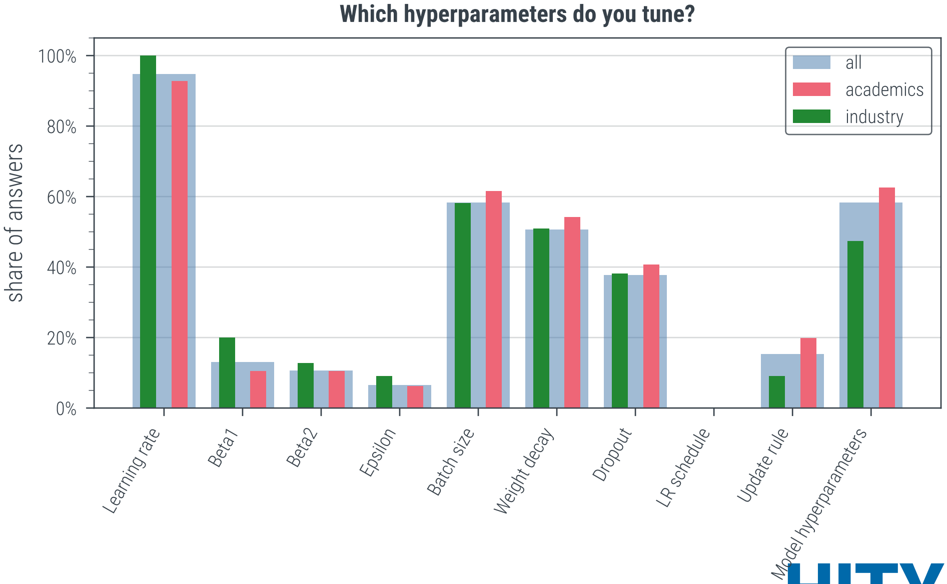
170 total responses
(116 online, 54 at workshop)

HITY
NeurIPS Workshop 2022

How many hyperparameters do you tune?

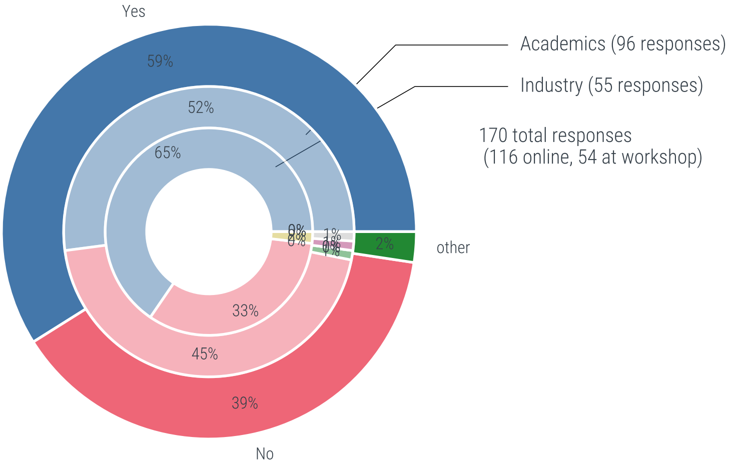


170 total responses
(116 online, 54 at workshop)



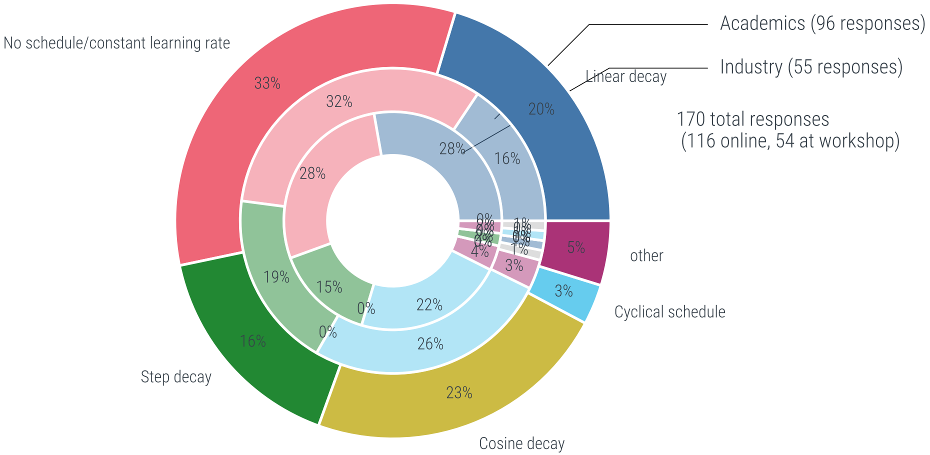
HITY
NeurIPS Workshop 2022

Do you determine your batch size by memory, i.e. such that one batch fits exactly in memory?



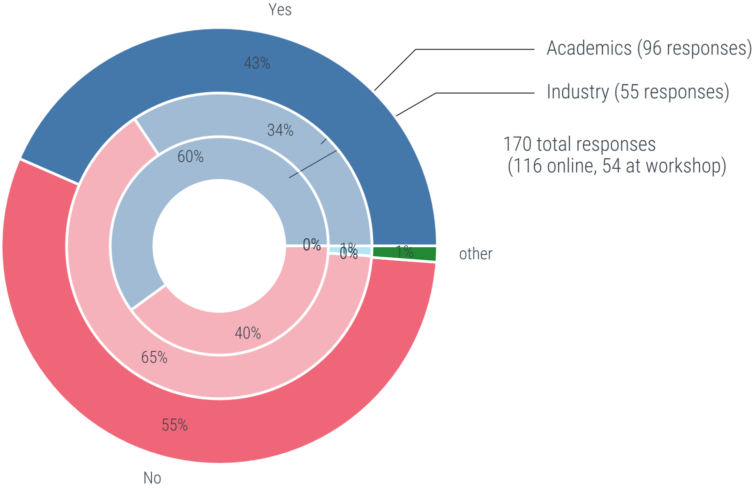
170 total responses
(116 online, 54 at workshop)

Which learning rate schedule do you try out first?

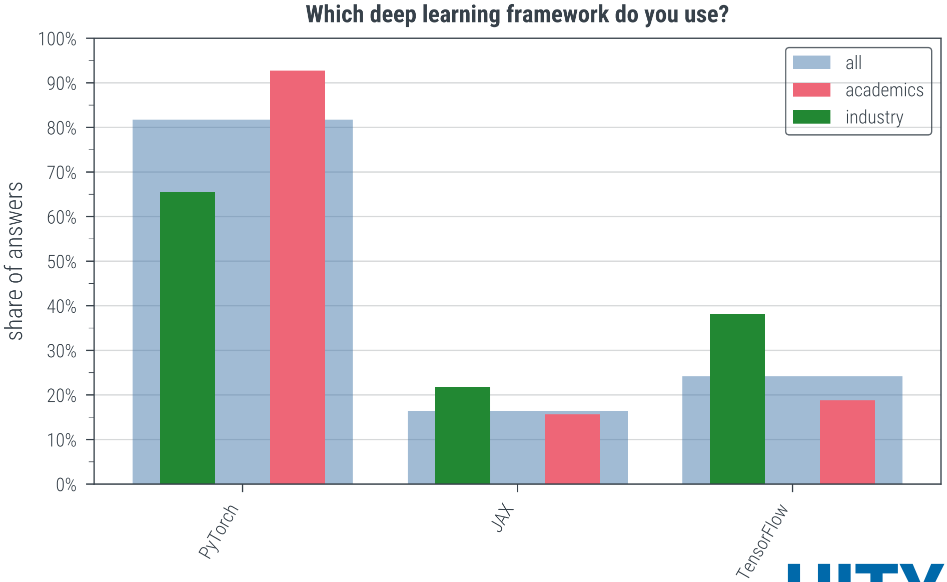


170 total responses
(116 online, 54 at workshop)

Do you use a learning rate warmup?



170 total responses
(116 online, 54 at workshop)



HITY
NeurIPS Workshop 2022

What are your favorite training tricks for neural networks that more people should know? Slide 1 of 2

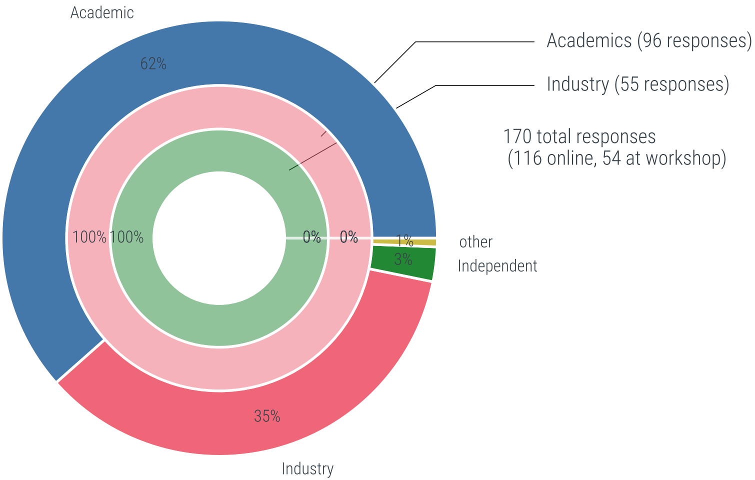
- Marginal likelihood
- OneCycle scheduler, gradient checkpointing
- Genetic Algorithm for Hyperparameters
- W&B
- avoid batches that lead to nan/inf losses
- One cycle, low fidelity training, sgd with restarts
- Use proximal optimization for regularizers
- The Composer library for PyTorch (MosaicML)
- using line search to choose the maximum learning rate possible at each step
- Normalized updates
- We recently published a survey of our tricks: `https://arxiv.org/abs/2209.05310`
- Use Distributed Shampoo, Normformer, GLU
- Weight averaging

What are your favorite training tricks for neural networks that more people should know? Slide 2 of 2

- FreezeOut
- L2 regularised features
- try a different epsilon value!
- Use sample prioritization techniques, and perform calibration during training.
- free nats (ignoring losses that are smaller than some threshold)
- Check consistency of hyperparameter performance over multiple seeds
- Lowering the learning rate!
- In some contexts, normalizing data as pre-processing step works a lot better than batch or layer norm. Better than batch or layer norm.
- Mixed precision training
- label smoothing
- Train with a small subset
- Ablation spreadsheets and oncalls. :)
- Cyclic and one cycle LR
- Label smoothing (easy & works well), looking at outputs for failure modes

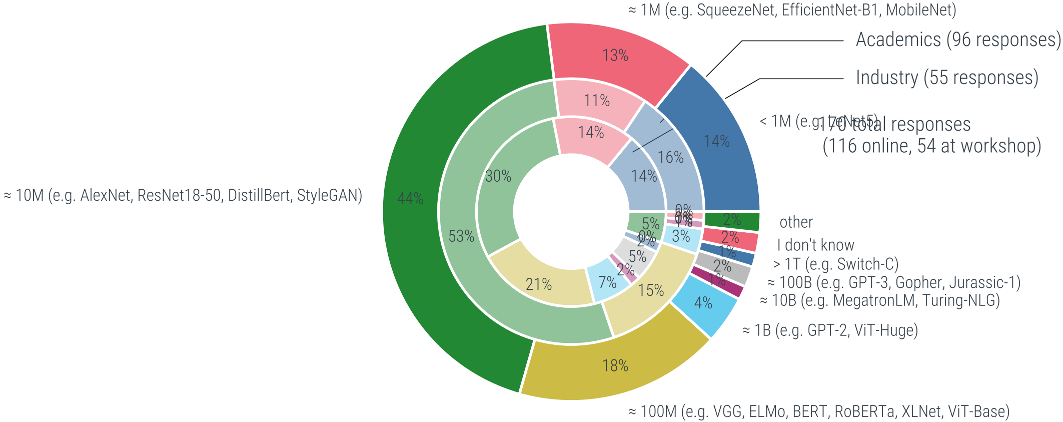
– Backup –

What is your background?

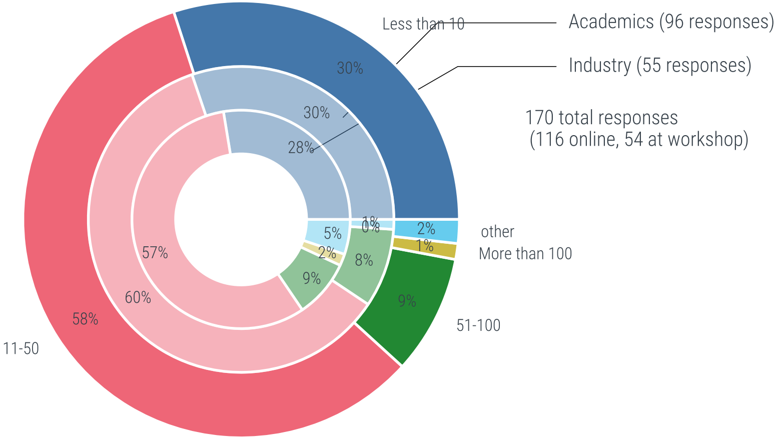


170 total responses
(116 online, 54 at workshop)

How many weights do you usually have in your neural network architectures? Please estimate the order of magnitude.

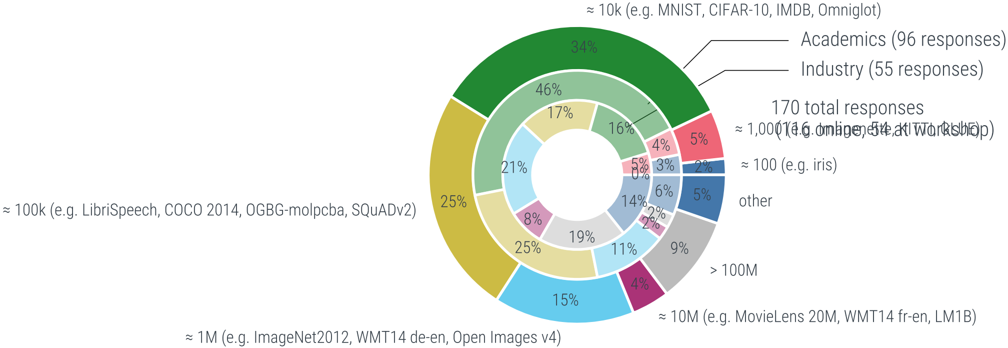


How many layers do you usually have in your neural network architectures?

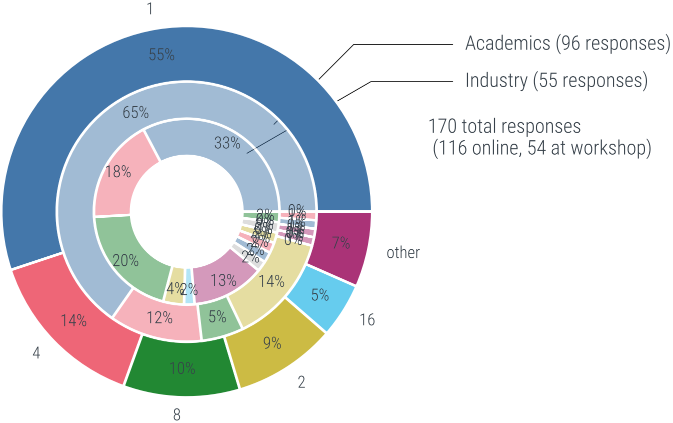


170 total responses
(116 online, 54 at workshop)

How many training samples are typically in your data sets? Please estimate the order of magnitude.



How many GPUs (or similar accelerators) do you usually use to run a single training run? Please select the closest match. Note, this question does not ask how many GPUs you have available to perform runs in parallel, but how many GPUs you use to train a single model.



170 total responses
(116 online, 54 at workshop)