Tuning Deep Learning Models

- Optimizer Hyperparameters
  - Learning rate, batch size, epochs
- Model Hyperparameters
  - Number of layers, hidden units, model-specific

**Learning Rate**

Seemingly most important hyperparameter.
Good initial value = 0.01

**Good Cases**
Validation error decreases

**Bad Cases**
Validation error increases or does not decrease fast enough
Bad Case
Validation Error stops decreasing and somehow oscillates.

Solutions → learning rate decay
- ↓LR linearly (divide by half every k epochs)
- ↓LR exponentially (mult by 0.5 every k epochs)
- Adaptive Learning → ↓/↑ as needed

Batch Size

Affects both resource requirements and training.
A good initial value ≈ 32, 64

↑ Batch Size requires more memory.

↓ Batch size has more noise which helps the GD not to get stuck in local minima.

Accuracy

Training too slow (1, 2, ..., 16)

~ 32

Batch size

Computationally taxing poorer performance
epochs

- We must monitor validation loss when it stops decreasing, we should stop.
- Early stopping technique: stop training if the validation loss has not decreased in k epochs.

Model Size

- Number of hidden units. Intuitively, it controls the model's capacity to learn a function.
- Having too much capacity → overfit

But the general rule is "more is better"

- Number of layers. As a rule of thumb, 3 layers is better than 2. But more than that does not usually help. (exception: CNN)
Cell Type

- In an RNN, we can choose LSTM, GRU, or just the vanilla RNN. The first 2 are usually better.
- Choosing LSTM or GRU is task-dependent.
Selecting Reasonable Hyperparameters

- Training loss \(\ll\) Validation loss?

  - Network is overfitting
    - Option 1: Increase dropout
    - Option 2: Decrease model size
training loss \approx validation loss?

network is underfitting

increase model size

- \# of layers
- \# of neurons per layer

- the number of model parameters should be about the same magnitude as the size of dataset.

100 MB dataset (~ 100 million chars)
number of model params $\ll$ size of dataset?

the model underfits