5 trials for both training and inference latency measurement

The first trial has no significant difference than others. Difficult to measure iterations as time.time\_ns() is not accurate.

variance low, not outliers.

# Dataset

The train and dev set from SST-2. Bag-of-words is used as the feature.

# Network Architecture

The network contains 1 input layer which projects the input bag-of-words vectors into hidden layer size. Then the vectors are passed to certain number of identical hidden layers are and finally be converted to a logit via the output projection layer. ReLU is used as the activation function. When performing inference, sigmoid is applied to the result and if the output is considered positive if it is larger than 0.5.

# Configurations

The following options are used for all experiments:

* learning rate 0.001
* batch size 64
* epochs 2

## Baseline

The baseline model has 2 hidden layers, and its hidden size is 256.

The bag-of-word feature vectors are extracted from the top 10,000 most common words in the training data. We ignore out-of-vocabulary words when processing the dev set.

The accuracy evaluated on the dev set is 0.81.

Varying hidden layer size

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | # hidden layers | hidden size | input size (vocab size) | output size | training batch size | # params | FLOPs |
| Base | 2 | 256 | 14704 | 1 | 64 | 3896321 | 7791873 |
| Deep | 4 | 256 | 14704 | 1 | 64 | 4027905 | 8054529 |
| Shallow | 1 | 256 | 14704 | 1 | 64 | 3830529 | 7660545 |
| Wide | 2 | 512 | 14704 | 1 | 64 | 8054785 | 16108033 |
| Smaller input | 2 | 512 | 5000 | 1 | 64 | 1412097 | 2823425 |
|  |  |  |  | 1 | 64 |  |  |