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

Base model:

Training time (seconds): [58.94187879562378, 59.73922681808472, 58.75348138809204, 60.40507936477661, 60.226938247680664]

Accuracy: 0.8222477064220184

Inference time (ms): [0.30280367 0.29592259 0.29816307 0.29357718 0.30024839]

Deep:

Training time (seconds): [64.73041129112244, 64.3940851688385, 63.025757789611816, 63.243123292922974, 62.9337215423584]

Accuracy: 0.8279816513761468

Inference time (ms): [0.34394025 0.3621539 0.33378154 0.33759679 0.35382546]

Shallow:

Training time (seconds): [55.409019231796265, 54.435054063797, 54.541584491729736, 54.59976005554199, 54.524529457092285]

Accuracy: 0.8268348623853211

Inference time (ms): [0.25778463 0.26768555 0.26778899 0.25114633 0.25300241]

Wide

Training time (seconds): [116.7300660610199, 116.35143113136292, 115.42436933517456, 116.74917149543762, 116.07944297790527]

Accuracy: 0.8256880733944955

Inference time (ms): [0.68692936 0.70068452 0.66972408 0.69724862 0.69621147]

Smaller input (top-5000 common vocab)

Training time (seconds): [21.529683589935303, 22.65859055519104, 22.14132809638977, 22.82339072227478, 23.089998483657837]

Accuracy: 0.8153669724770642

Inference time (ms): [0.20641181 0.18807466 0.20986284 0.21330206 0.19609989]

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | # 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 |  |  |