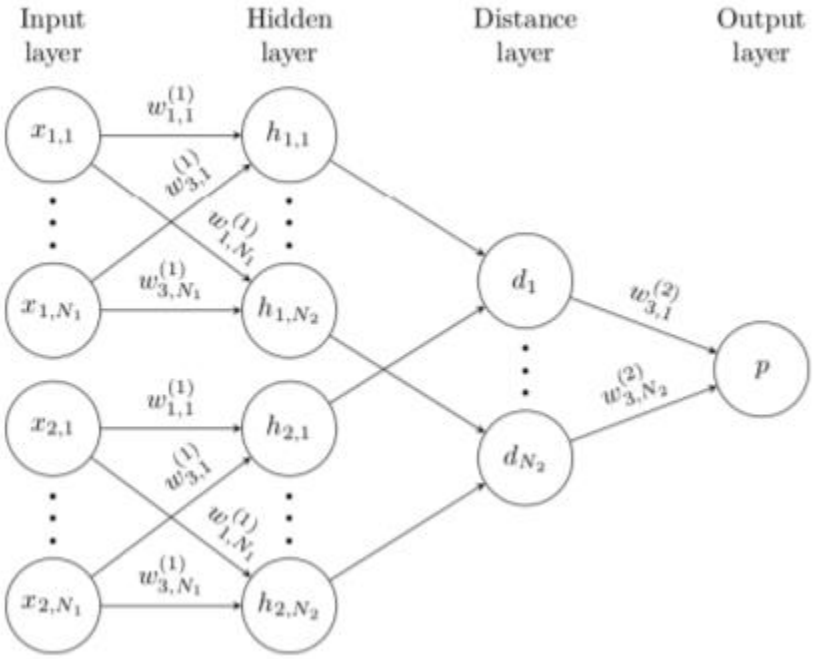
Notes

The identity vector from the top and bottom identical NNs are compared such that the same dimensions are compared with each other:



Weight tying can be achieved by adding the gradients from both identical neural networks.

The learning rate was decayed with the following formula: $\eta_j^{(T)} = 0.99\eta_j^{(T-1)}$. The learning rate $\eta_j \epsilon [10^{-4}, 10^{-1}]$.

The momentum started at 0.5 in every layer, increasing linearly each epoch until reaching the value μ_j, the individual momentum term for the jth layer. $\mu_j \epsilon [0,1]$.

“We trained each network for a maximum of 200 epochs, but monitored one-shot validation error on a set of 320 one-shot learning tasks [...]. When the validation error did not decrease for 20 epoches, we stopped and used the parameters of the model at the best epoch according to the one-shot validation error. If the validation error continued to decrease for the entire learning schedule, we saved the final state of the model generated by this procedure.”

$\lambda_j \epsilon [0, 0.1]$.

References

- https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf