■ Observation (sec 3): Lottery ticket eliminates delayed generalization

(a) Dense network

Delayed generalization

(b) Lottery Ticket (Good Subnetworks)

No delay!!

■ What happen during delayed generalization? Why LT eliminate delayed generalization?

H1 [Liu+]: Reducing weight norm	H2 [Merrill+]: Higher Sparsity	H3 (Ours): Finding good structure
smaller $\|\theta\|_2$	smaller $\|\theta\|_0$	bettr structure