



Utku Evci¹, Yani Ioannou², Cem Keskin³, Yann Dauphin¹

¹Google, ²University of Calgary, ³Meta



UNIVERSITY OF
CALGARY

Meta



1. Motivation:

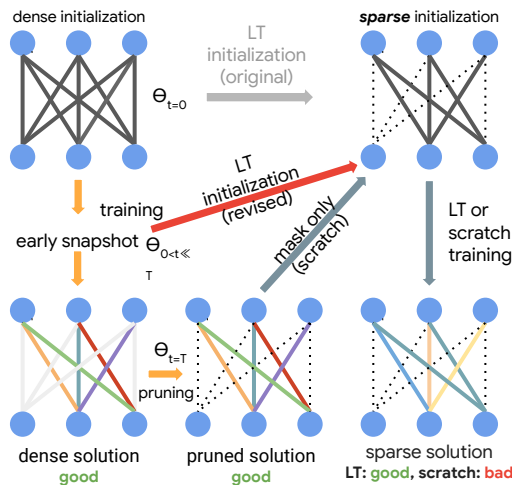
Unstructured Sparse NN Training results in poor generalization

Two exceptions:

- (a) Lottery Tickets (LTs) and,
- (b) Dynamic Sparse Training (DST).

Q: What makes DST and LTs the exceptions?

We observe **poor gradient flow at initialization and during training - even for LTs!**



2. Poor Gradient Flow at Init.:

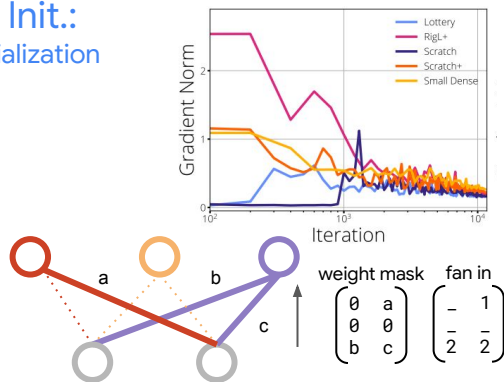
We propose a Sparsity-aware Initialization

Sparse NN training use dense initialization, but sparse NN have different fan-in/fan-out!

Our **sparsity-aware initialization**

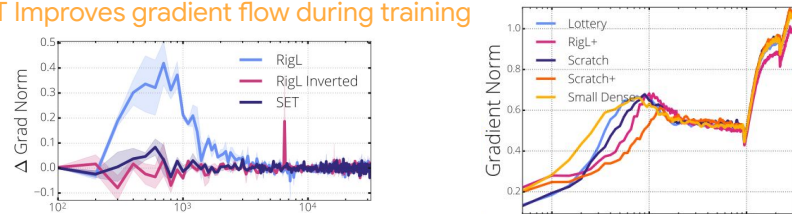
accounts for fan-in/fan-out of unstructured sparse neurons giving:

- (a) better gradient flow at init. and,
- (b) better generalization



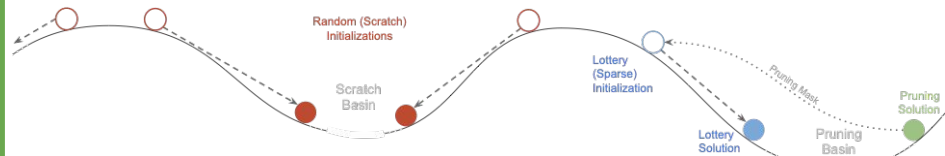
3. Poor Gradient Flow during Training:

DST Improves gradient flow during training



4. Lottery Tickets Relearn the Pruning Solution:

LTs are easier to train because they initialized within the same solution basin as the pruned solution, and effectively relearn the pruning solution



We investigate

LT solns to pruned solns:

- (a) LTs start close and move towards to the pruning solution.
- (b) LT solutions are in the same basin as the pruning solutions.
- (c) high function similarity and poor LT ensemble performance

