Andrew Saxe

Theory of Rep Learning in complex Architectures

What do DL nets actually learn?

Semantic meaning #topic

Deep linear Net

- Gains nothing from depth

SVD change

Show to V

- Cross assoc matrix of items and their properties
- Decomposes the matrix into

shit. I didn't get the nomenclature.

D: size of net (layers)

$l = 1 \cdots D$

W = weights?

Exact solutions

- about lost here

like lying on an ocean feeling the waves push through your body also made of matter

Hidden 1D Layers

- Exposing hidden linear mechanisms in a DLN

Error surface

- Getting stuck on initial plateau
- DLNs have plateaus

SVD change of Variables

Mostly been talking about speed & scaling of training speed

Semantics & classification

Semantic development

I need to read Chomsky

D linear net model
- good graph for V

Diverse structures
- Nice illustration

Prog Diff
- Theory children learn broad categories before finer distinctions

Idiosyncratic Individuals
- Is he arguing for an "ego"?
- or hidden layer as "ego"?

Transformer Heads

Neural Race Reduction

Mesoscale Architecture

- "We don't have good ideas of how architecture represents" !

Mupkes et al. 2020 #read !

Theoretical Analysis

- I think this maybe a job for me
- Inference of network activity?

Gated D Linear Net

- adding scalar gating variables "g"
  - multiply by gating var @ weight & @ node

Gating intp

- adding gating variables with inputs

Grad. Deg.

- Compressed path notation
  - to describe "path": the fall

- Pathway counting logic
- More shared pathways = more learning

Neural Race Reduction

- I think we are reducing a ReLU network to it's effective parts via studying effective paths

---

Lazy v. Feature learning

- "Large networks w/ large weights learn better"?

- lazy is not as useful, more static

Routing net

-

I don't get how these are not oversaturated,

like, yes we want as much multiplexing/ generalizability overlap of calculations

but...

How can they do this reduction easily?

I guess they mentioned it with the node centric feature stuff

- initializations must be small

- gating variables tell you the importance of edges

Saxe et al. tbp #read