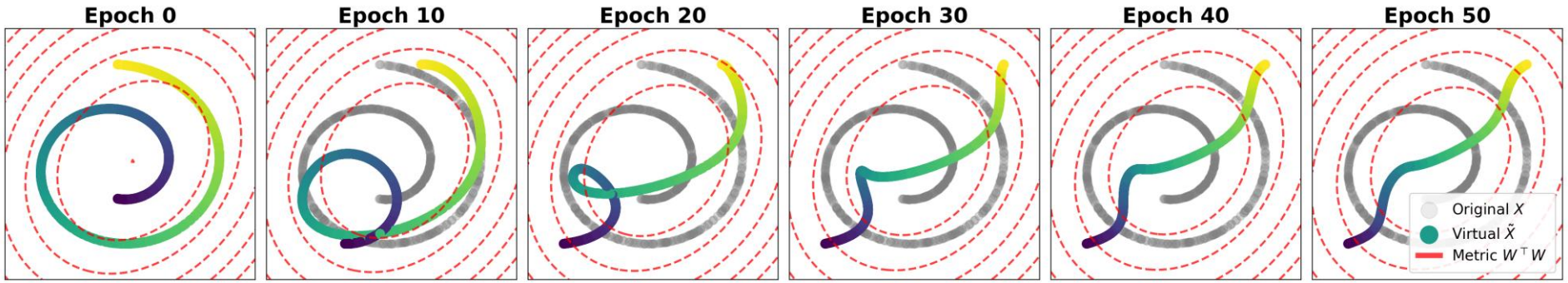
The Weight Gram Matrix Captures Sequential Feature Linearization in Deep Networks



Taehun Cha (Korea Univ.),

Daniel Beaglehole (UCSD),

Adityanarayanan Radhakrishnan (MIT),

and Donghun Lee (Korea Univ.)

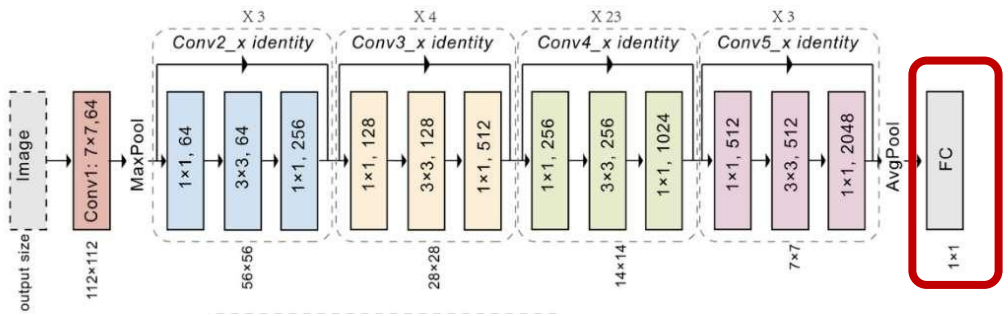
AIML@K KOREA UNIVERSITY

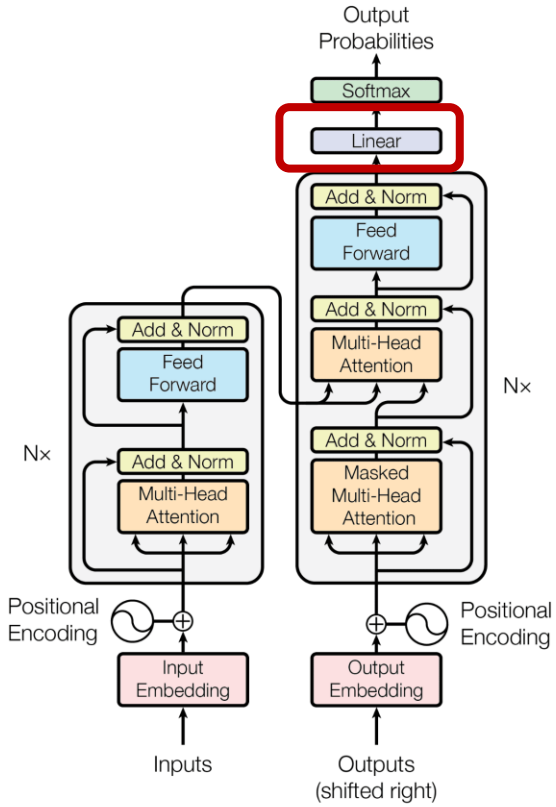
What is Feature Learning?

- Deep neural networks outperform kernel methods
  - Feature Learning (FL) is a key to understand this success
  - Yet there us no widely agreed-upon definition of FL
- Let's begin with a basic principle:

All DNNs end with linear last layer

- Can this guide us to define and analyze FL?





AIML@K KOREA UNIVERSITY

Revisiting the Chain rule

- For any function f that takes $z = Wh$ as its input,

$$\nabla_W f = \nabla_z f \cdot h^\top \text{ and } \nabla_h f = W^\top \nabla_z f, \text{ so}$$

$$\boldsymbol{W^\top \nabla_W f = \nabla_h f \cdot h^\top}$$

  which shows the exact relationship between weight grads and input grads.

- How the weights are updated is directly connected to how the inputs would be updated if they were optimization variables.

- We refer to this identity as "**feature learning equation.**" Let's see what it gives.

AIML@K KOREA UNIVERSITY

GD is Virtual Covariance Learning

- Let's define:
  - L: loss function which takes $z = Wh$ as its input
  - Actual update: $W^+ = W - \gamma \nabla_W L$
  - Virtual update: $h^+ = h - \gamma \nabla_h L$
- **Theorem 1** The following holds up to a residual term of order γ^2:

$$(W^+)^\top (W^+) - W^\top W \approx (h^+)(h^+)^\top - hh^\top.$$

  - ΔCov(Actually updated weight) ≈ ΔCov(Virtually updated hidden state)

AIML@K KOREA UNIVERSITY

Relationship between Actual/Virtual Updates

- **Proposition 1** Under single-layer/single-sample training, **the actually updated feature does not deviate far from the virtually updated feature**,

$$|\sigma(W_{l-1}^{+}h_{l-1}) - \sigma(W_{l-1}h_{l-1})| \leq c^2|h_l^{+} - h_l| \text{ } (c\text{: Lipschitz constant) and}$$

$$\text{sgn}(\sigma(W_{l-1}^{+}h_{l-1}) - \sigma(W_{l-1}h_{l-1})) = \text{sgn}(h_l^{+} - h_l), \text{ elementwisely.}$$

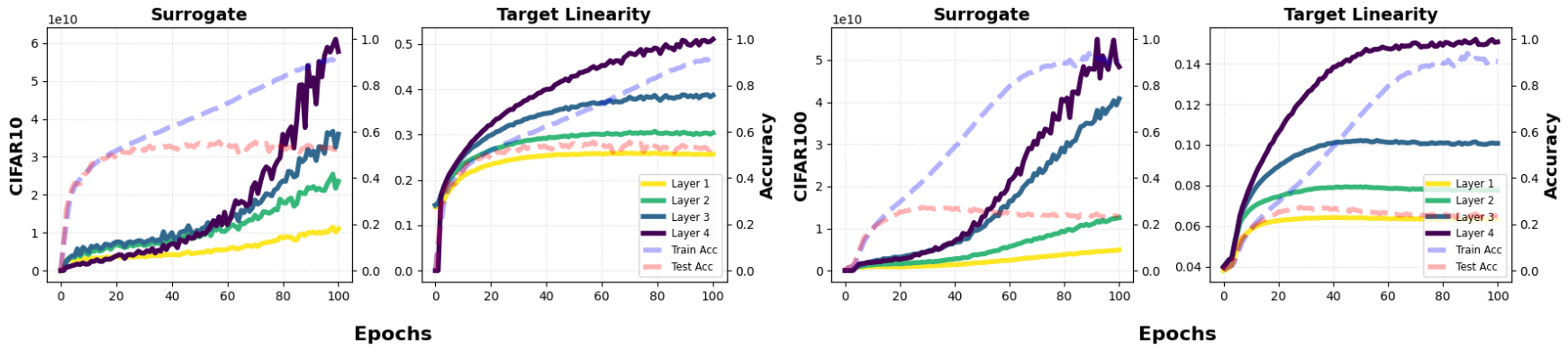
- **Proposition 2** Under full training in the gradient-flow regime, **the actual update is a pre-conditioned virtual update**, i.e. there exists a psd matrix M s.t.

$$\frac{dh_l}{dt} = -M\nabla_{h_l}L$$

AIML@K KOREA UNIVERSITY

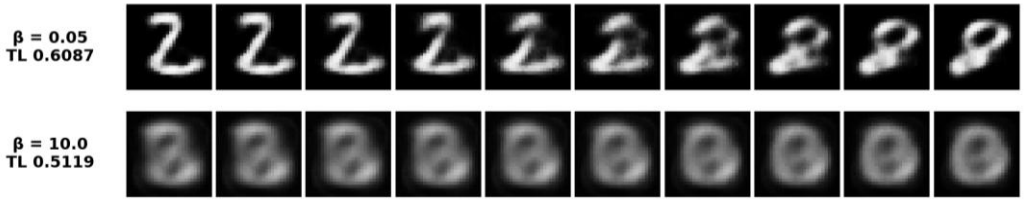
DNN is a Sequential Linearization Machine

- We define **Target Linearity (TL)** as R^2 **of a linear prober** on hidden states
  - It measures the strength of the linear relationship between hidden states and labels
- Theoretically, we show that TL increases
  - during standard gradient-descent training (**Theorem 3**)
  - as data pass through deeper layers (**Theorem 4**)



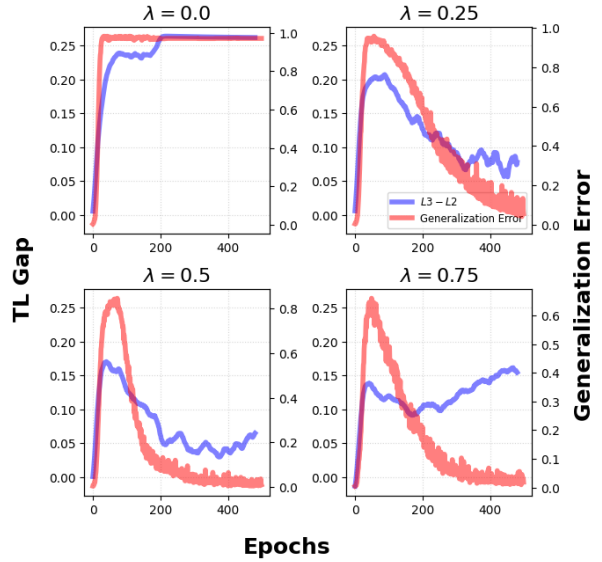
TL as a Unifying Framework

- TL provides theoretical explanations for other FL phenomena
  - **Neural Collapse** is equivalent to TL maximization (**Proposition 3**)
  - High TL is a sufficient condition for **latent-space interpolation** in VAEs (**Proposition 4**)



- We also present empirical observations on
  - Generalization (random-label memorization and grokking)
  - and pre-trained language models

See our paper for details!



TL and Grokking

AIML@K KOREA UNIVERSITY