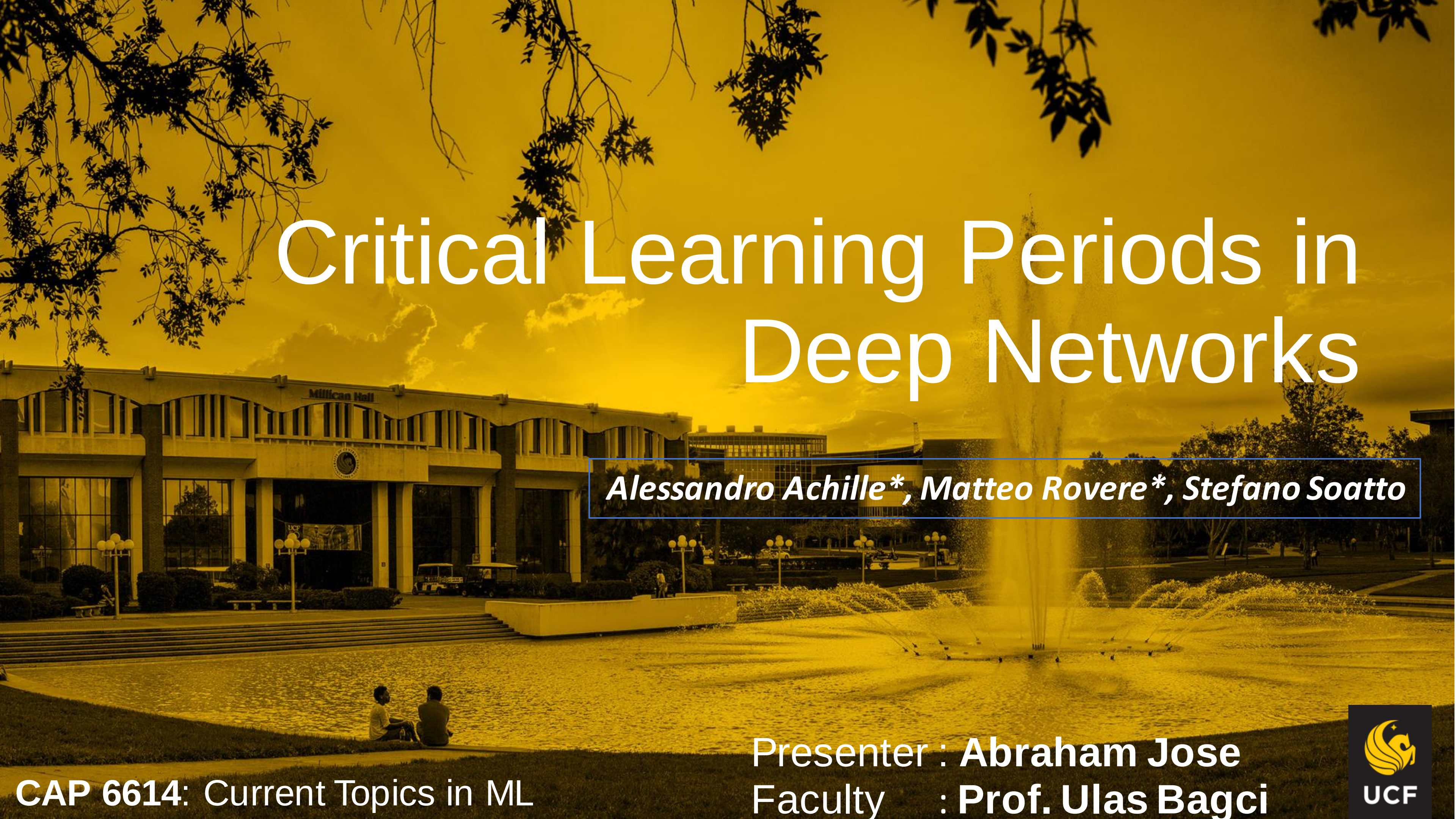


Critical Learning Periods in Deep Networks

Alessandro Achille, Matteo Rovere*, Stefano Soatto*

Presenter : **Abraham Jose**
Faculty : **Prof. Ulas Bagci**

CAP 6614: Current Topics in ML

Critical Learning Period in Nature

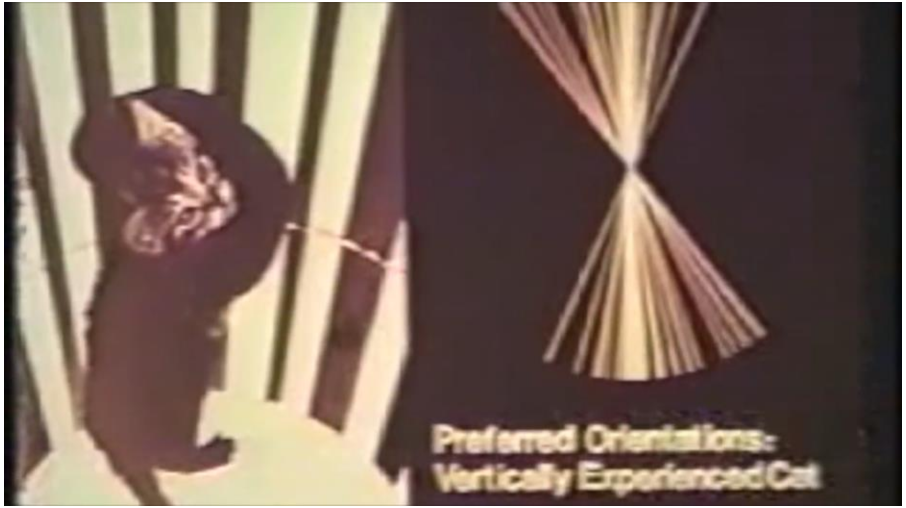
Critical Learning Periods : time windows of early post-natal development during which sensory deficits can lead to permanent skill impairment. [*biochemical phenomenon*].

Critical periods affecting a range of species and systems, from **visual acuity in kittens** (Wiesel & Hubel, 1963b; Wiesel, 1982) to **song learning in birds** (Konishi, 1985).

Kitten Experiment

Wiesel & Hubel's 1962 experiment proved that Critical Learning period exist during which we can remove the permeant visual skill impairment.

Visual experience is limited to vertical stripes. Hence responses is limited to vertical gradient!



Kitten suffer three week to three months deficit window, after which the damage is permeant.

Some important aspects of Visual perception are acquired rather than innate.

Kitten Experiment URL

Motivation

Deep Neural Network shows similar Critical Learning periods and the deficit window as we discussed.

Critical Learning Period: Maximum window of epochs in which we can remove a deficit without penalizing the performance.

Learning	
Memorization *rapid growth of Information*	**Forgetting** *achieve invariance & disentanglement in representation*
Critical Learning	Deficit causes permanent effect

Hypothesis

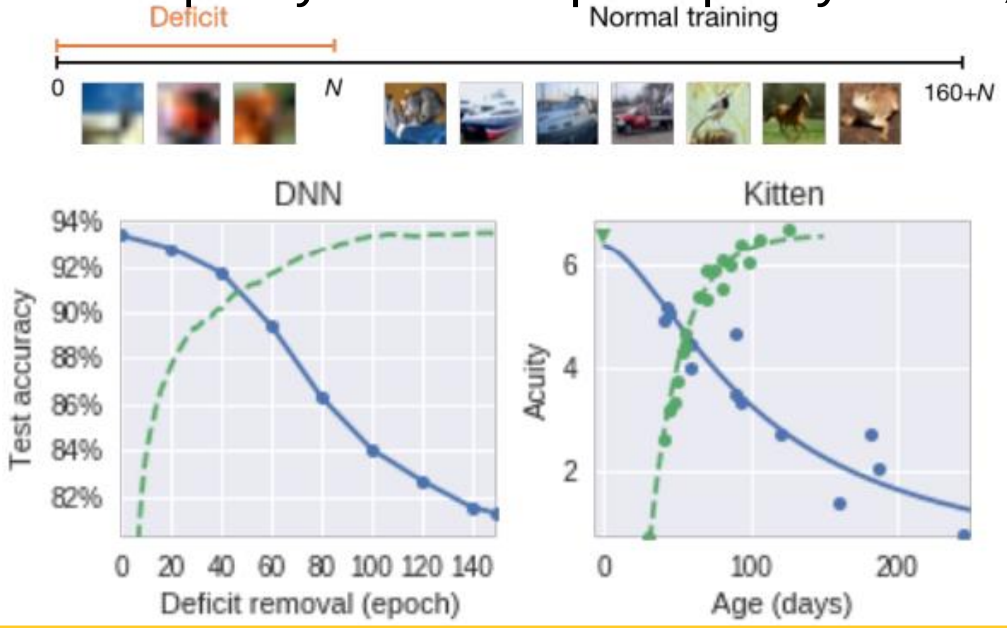
Every learning system shows Critical Learning period.

Neural Plasticity could be a consequences of ***Information Plasticity***, which is defined the distribution of information throughout a network.

Critical Periods are centered in memorization phase.

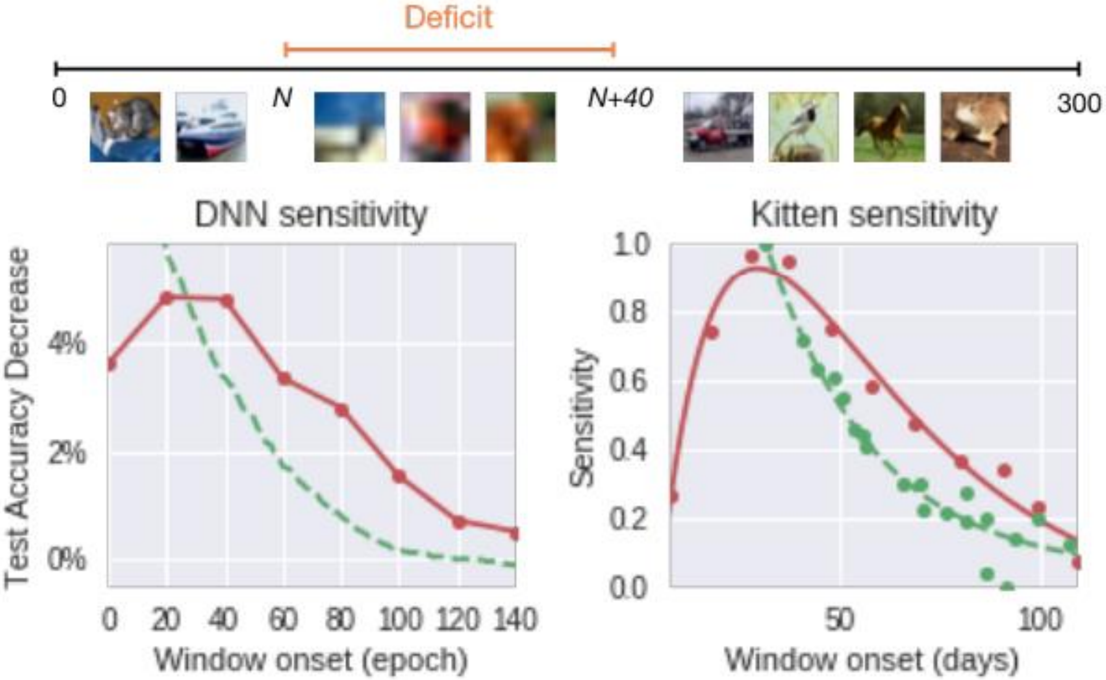
Critical Learning in DNN

Due to information processing which are mostly sensory deprivation-like deficit, which doesnot affects high level features. For example, cataract.(downsample by 8x8 and upsample by 32x32, bilinear)



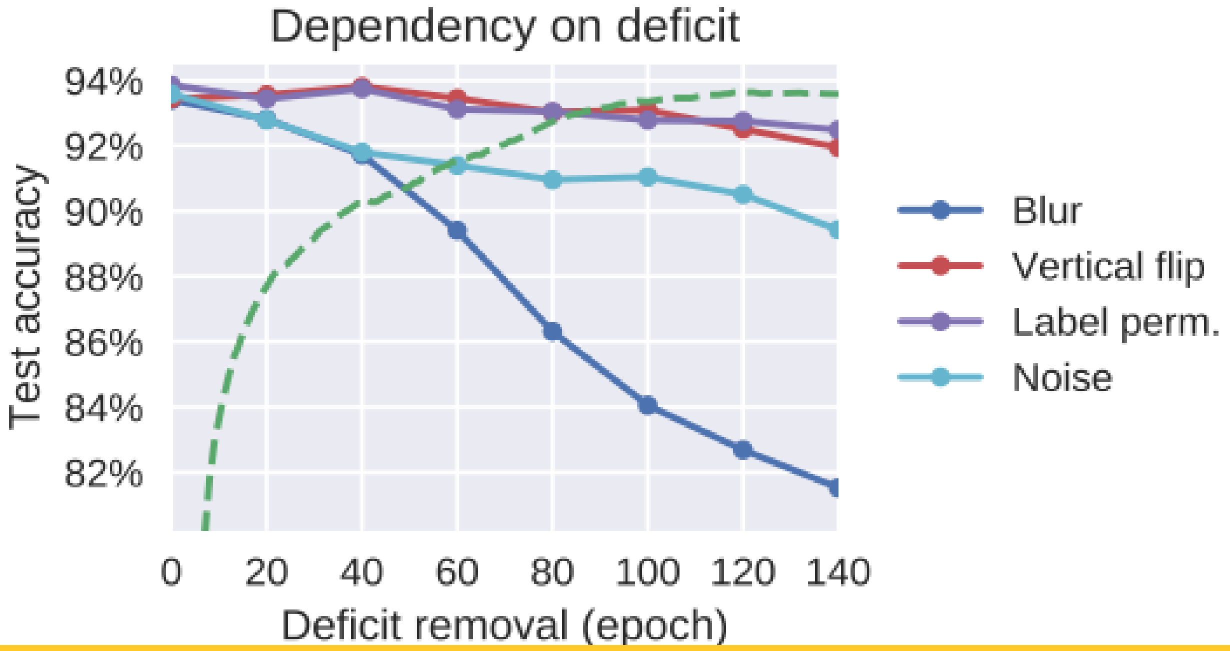
Critical Learning in DNN

However onset of a short 40-epoch deficit will cause decrease in the final performance can be used to measure the sensitivity of the deficits.



Deficits in DNN

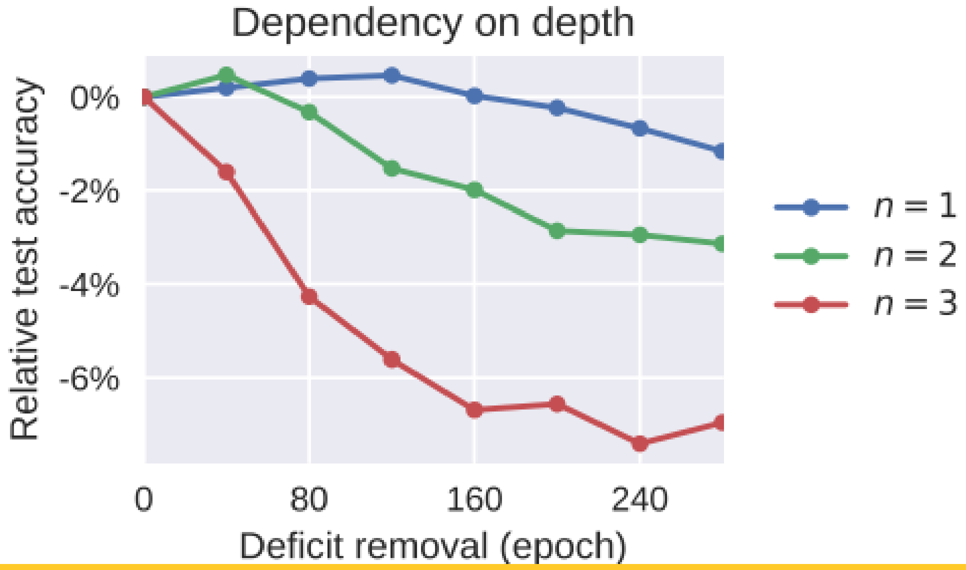
NN remain plastic enough to adapt changes in high level features(flip,label permutation) etc.. ie, by deficit correction we get to baseline performance.



Deficits in DNN

Blur, noise in image can severely affect how network gathers low level features and will be a permenent deficit if not treated within deficit window.

Critical period α Depth of n/w and hyperparams(shape & size of CP)



Fisher Information Analysis

Consider a network encoding the posterior distribution $p_w(y|x)$ parameterized by weight 'w', input image 'x' and of task variable 'y'.

The dependency of the final output w.r.t corresponding weight when you perturb the weight, Kullback-Leibler divergence or second order is

$$\mathbb{E}_x \, \mathrm{KL}(\, p_{w'}(y|x) \, \| \, p_w(y|x) \,) = \delta w \cdot F \delta w + o(\delta w^2),$$

Expectation over x is computed using the empirical data distribution,we will get the Fisher Information Matrix (FIM)

$$F := \mathbb{E}_{x \sim \hat{Q}(x)} \mathbb{E}_{y \sim p_w(y|x)} [\nabla_w \log p_w(y|x) \nabla_w \log p_w(y|x)^T]$$

Fisher Information Analysis(cont..)

The Fisher Information can be used as a measure of the effective connectivity of a DNN, or the *"synaptic strength"* of a connection.
It is local metric measuring how much the perturbation of a single weight (or a combination of weights) affects the output of the network(y).

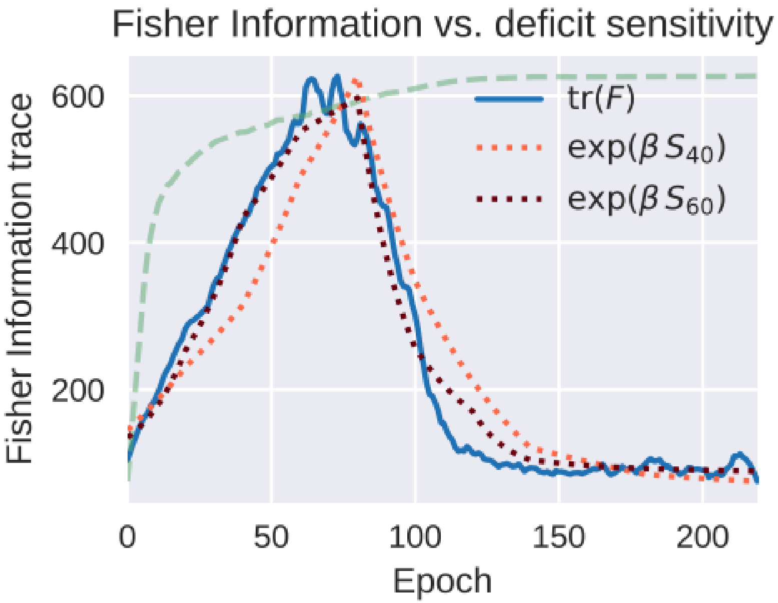
FIM is also a semidefinite approximation of the Hessian of the loss function.

FIM is expensive to compute, we will use it's trace to measure the global or layer-wise connection strength, which can be computed effectively.

$$\operatorname{tr}(F) = \mathbb{E}_{x\sim\hat{Q}(x)}\mathbb{E}_{y\sim p_w(y|x)}[||\nabla_w \log p_w(y|x)||^2].$$

Trace of FIM in Learning

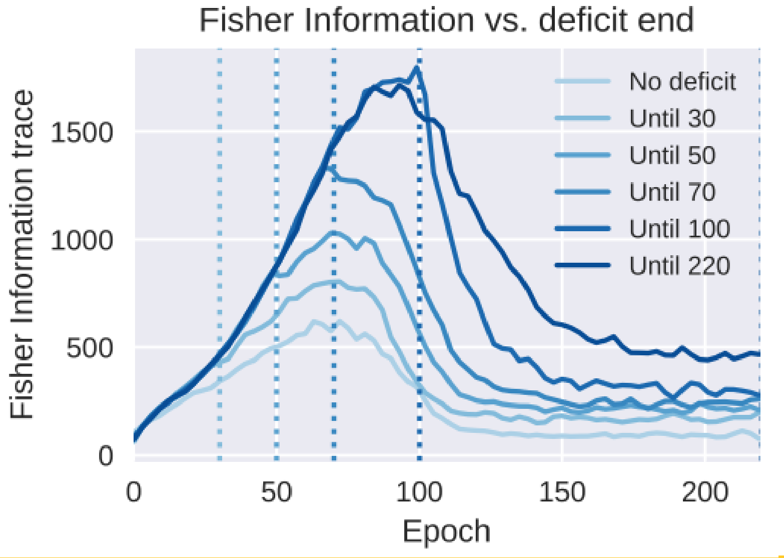
Information in network sharply increases, which means n/w holds more data and as forgetting phase starts, it will loosen synaptic strengths. FIM is local curvature of the residual landscape.



Trace of FIM in Learning

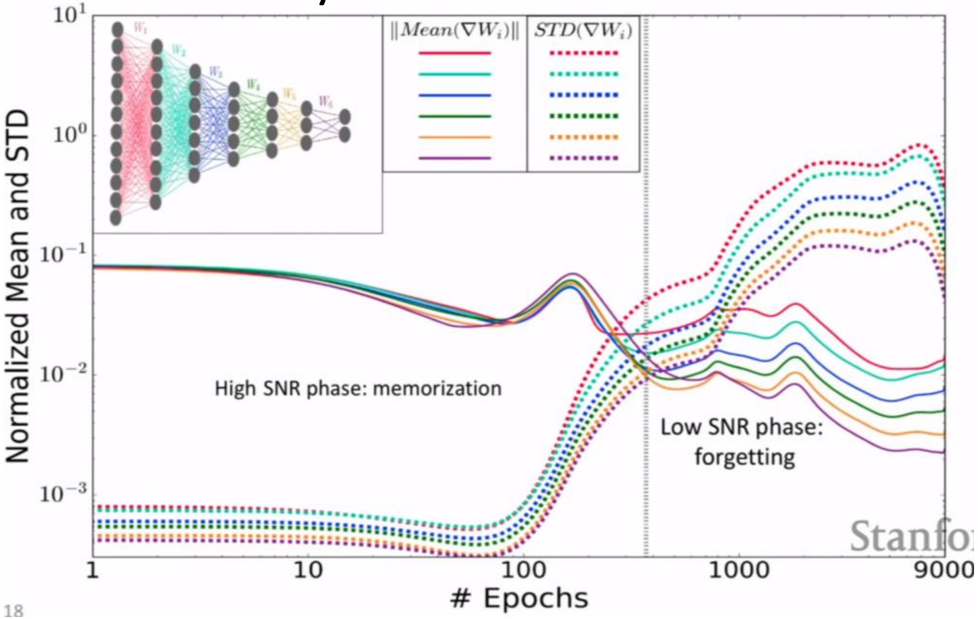
Blurring deficit : network must memorize more information to solve the task. Delay in removal of deficit results in inferior performance.

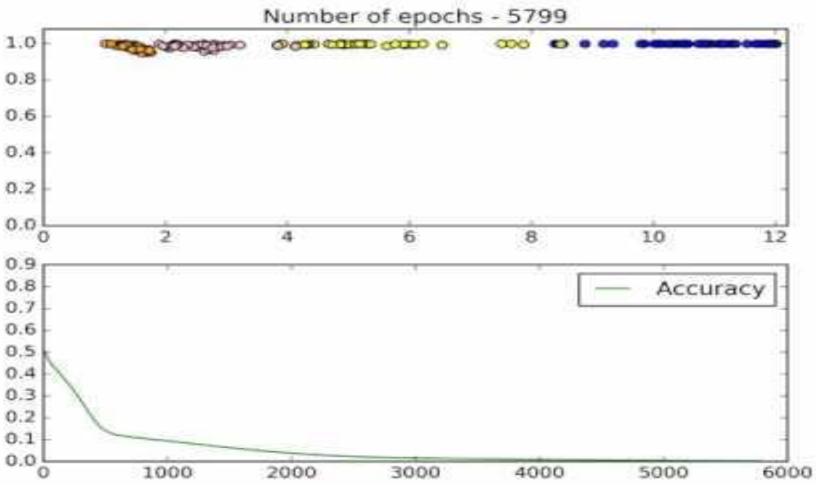
Critical Period after bottleneck crossing: It maynot be able to reach right valley in landscape, when the weightspace has low curvature.



Information Theory & DNN

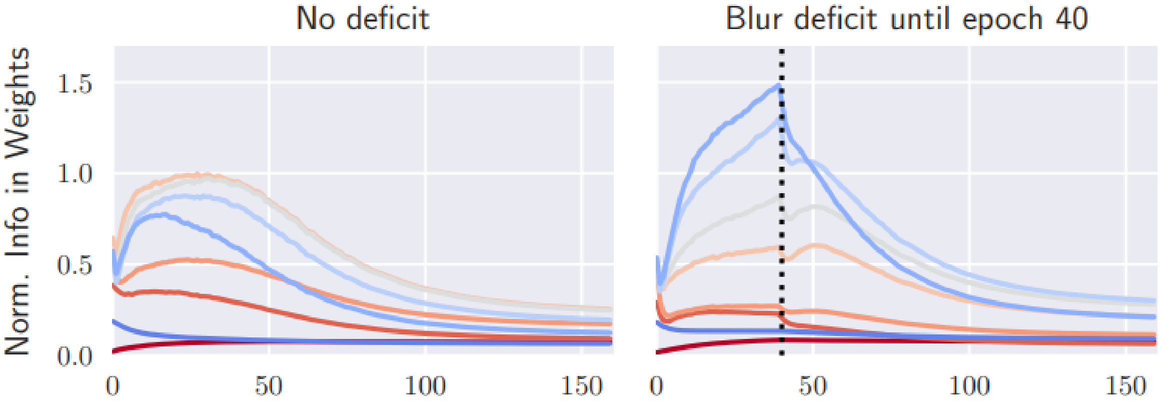
How information theory fits in DNN and why DNN works:*"Deep Learning and the Information Bottleneck Principle"* from Naftali Tishby & Noga Zaslavsky.





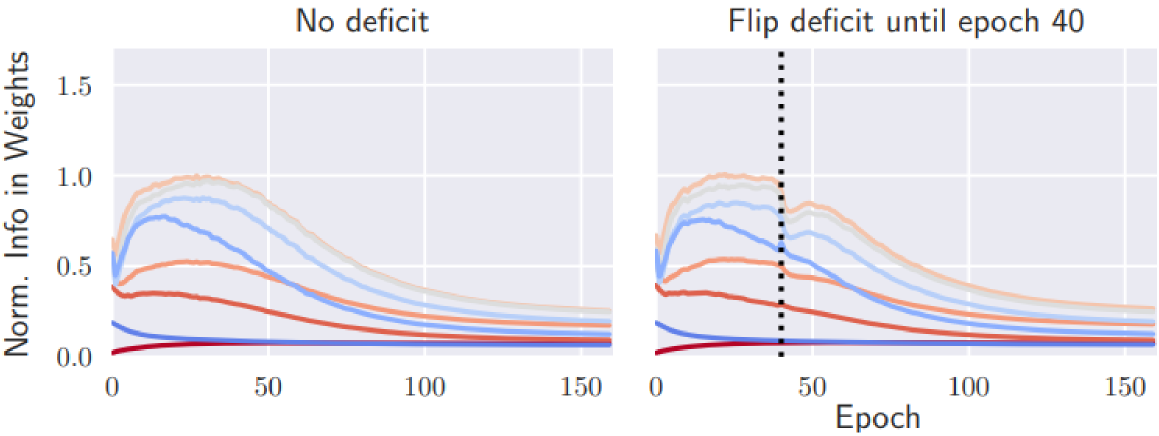
Trace of FIM in All-CNN: Layerwise

The blur deficit destroys the mid and low-level features(low & mid layers) and most information is consolidated in the high level layers(bottleneck crossing)



Trace of FIM in All-CNN: Layerwise

Prolonged defict causes severe performance loss. However, High level features will have strong trace of FIM if deficit is removed(change in info.)



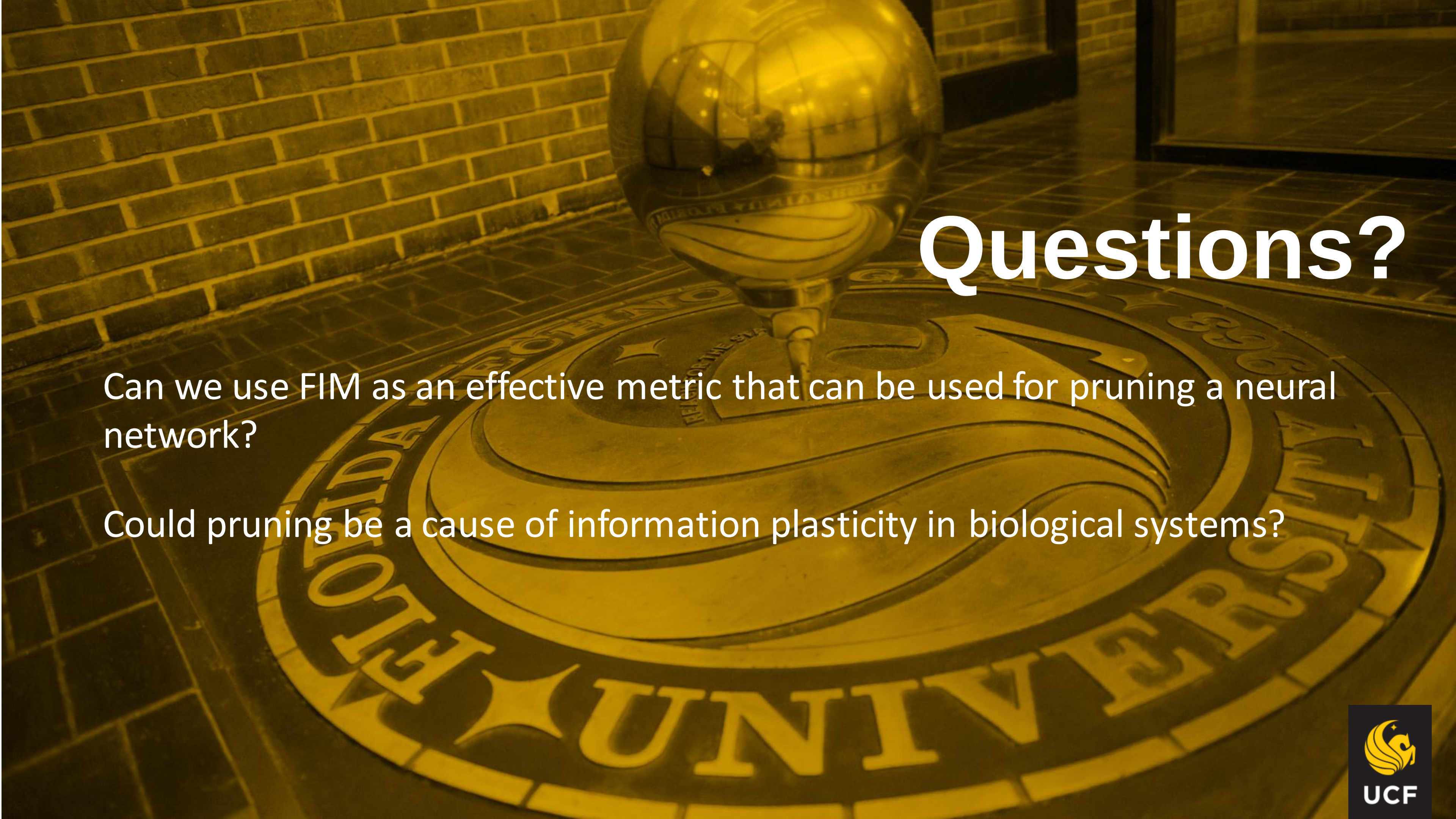
Reference

Alessandro Achille and Matteo Rovere and Stefano Soatto: *"Critical Learning Periods in Deep Networks" :* International Conference on Learning Representations 2019

Naftali Tishby & Noga Zaslavsky: "*Deep Learning and the Information Bottleneck Principle*"

Stanford Lecture by Naftali Tishby

https://blog.acolyer.org/2017/11/24/on-the-information-bottleneck-theory-of-deep-learning/- Part I and part II



Questions?

Can we use FIM as an effective metric that can be used for pruning a neural network?

Could pruning be a cause of information plasticity in biological systems?

Thank You :)