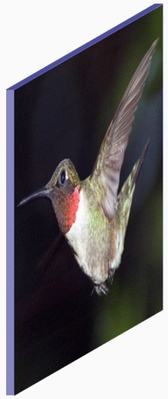
Information, Noise and Emergent Properties of Learning Deep Representations

Alessandro Achille
University of California, Los Angeles

Test Image



Minimal Invariant Representation

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

Weights

Training Set

{  , (car, horse, deer, …) }

Some notation

Cross-entropy: The standard loss function in machine learning

$$H_{q,p}(x) = \mathbb{E}_{x\sim q(x)}[-\log p(x)]$$

Kullback-Leibler divergence: “Distance” between two distribution (used in variational inference)

$$\begin{aligned}\mathrm{KL}(q(z)\|p(z)) &= \mathbb{E}_{z\sim q(z)}\left[\log \frac{q(z)}{p(z)}\right] \\ &= H_{q,p}(x) - H_q(x)\end{aligned}$$

Mutual Information: Expected divergence between the posterior *p(z|x)* and the prior p(z).

$$\begin{aligned}I(x;z) &= \mathbb{E}_{x\sim p(x)}[\mathrm{KL}(p(z|x)\|p(z))] \\ &= H_p(z) - H_p(z|x)\end{aligned}$$

Compression without loss of *useful* information

Task *Y* = Is this the picture of a dog?

Original X

X ~ 350KB

Compressed Z

Z ~ 5KB

Z is as useful as *X* to answer the question *Y*, but it is much smaller.

Image source https://en.wikipedia.org/wiki/File:Terrier_mixed-breed_dog.jpg

The “classic” Information Bottleneck

The Information Bottleneck Lagrangian

Tishby et al., 1999

Given data x and a task y, find a representation z that is **useful** and **compressed**.

$$\begin{aligned} \text{minimize}_{p(z|x)} \quad & I(x;z) \\ \text{s.t.} \quad & H(y|z) = H(y|x) \end{aligned}$$

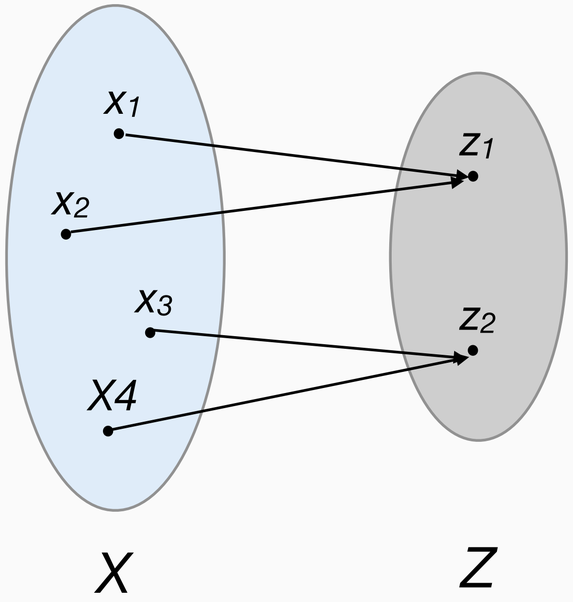
Consider the corresponding Lagrangian (the Information Bottleneck Lagrangian)

$$\mathcal{L} = H_{p,q}(y|z) + \beta I(z;x)$$

Trade-off between accuracy and compression governed by parameter β.

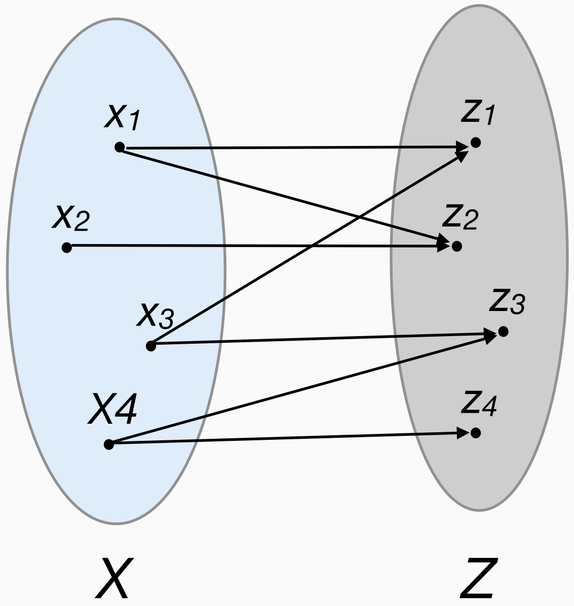
Compression in practice

Reduce the dimension



Examples: max-pooling, dimensionality reduction

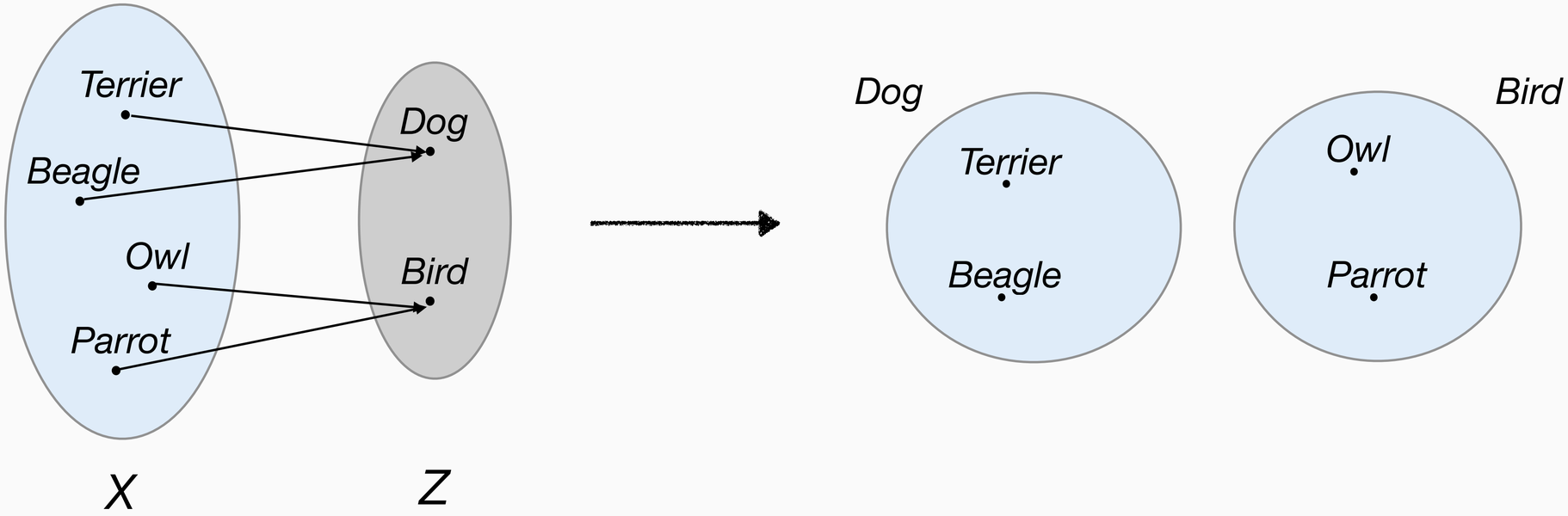
Increase dimension +
Inject noise in the map



Examples: Dropout, batch-normalization

Application to Clustering

An important application is task-based clustering, or summaries extraction.



See also Deterministic Information Bottleneck for hard-clustering vs soft-clustering.

Strouse and Schwab, *The Deterministic Information Bottleneck, 2016*

Information Bottleneck and Rate-Distortion

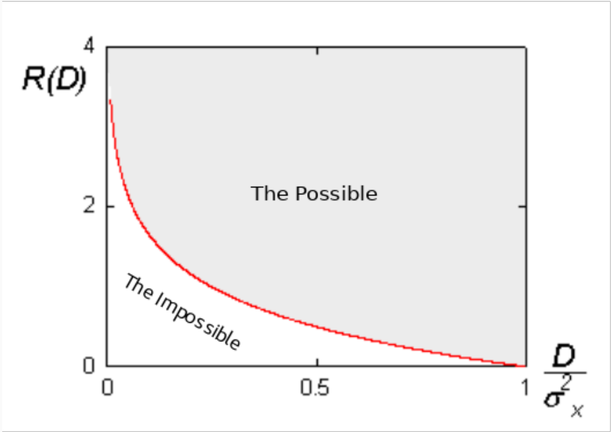
Rate-Distortion theory: What is the least distortion D obtainable with a given capacity R?

$$\min_{p(z|x)} \quad \mathbb{E}_{x,z}[d(x,z)]$$
$$\text{s.t} \quad I(z;x) \leq R$$

Equivalent to IB when *d(x, z)* is the information that *z* retains about *y*:

$$d(x,z) = KL(p(y|x)\|p(y|z))$$

Rate-distortion/IB curve:



Blahut-Arimoto algorithm

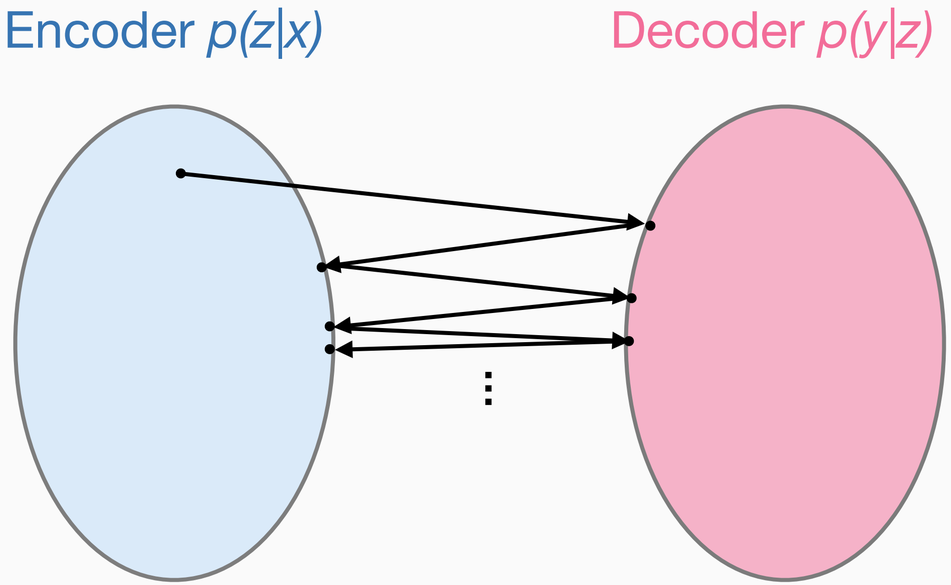
Blahut, 1972; Arimoto, 1972; Tishby et al., 1999

In general, no closed form solution. But we have the following iterative algorithm:

$$p_t(z|x) \leftarrow \frac{p_t(z)}{Z_t(x,\beta)} \exp(-1/\beta d(x,z))$$

$$p_{t+1}(z) \leftarrow \sum_x p(x) p_t(z|x)$$

$$p_{t+1}(y|z) \leftarrow \sum_y p(y|x) p_t(x|z)$$



But what happens if *p(z|x)* is too large, or parametrized in a non-convex way?

The variational approximation: IB and Representation Learning

Desiderata for representations

An optimal **representation** z of the **data** x for the **task** y is a stochastic function $z \sim p(z|x)$ that is:

Sufficient	$I(z; y) = I(x; y)$
Minimal	$I(x; z)$ is minimal among sufficient z
Invariant to nuisances	If $n \perp\!\!\!\perp y$, then I(n; z) = 0
Maximally disentangled	$TC(z) = \mathrm{KL}(p(z) \| \prod_i p(z_i))$ is minimized

Information Bottleneck Lagrangian

Minimal sufficient representations for deep learning

A minimal sufficient representation is the solution to:

$$\begin{aligned} \text{minimize}_{p(z|x)} \quad & I(x;z) \\ \text{s.t.} \quad & H(y|z) = H(y|x) \end{aligned}$$

Information Bottleneck Lagrangian:

$$\mathcal{L} = \underbrace{H_{p,q}(y|z)}_{\text{cross-entropy}} + \underbrace{\beta I(z;x)}_{\text{regularizer}}$$

Trade-off: between sufficiency and minimality, regulated by the parameter.

Invariant if and only if minimal

We only need to enforce minimality (easy) to gain invariance (difficult)

Proposition. (Achille and Soatto, 2017) Let z be a sufficient representation and n a nuisance. Then,

$$\underbrace{I(z;n)}_{\text{invariance}} \leq \underbrace{I(z;x)}_{\text{minimality}} - \underbrace{I(x;y)}_{\text{constant}}$$

Moreover, there exists a nuisance n for which equality holds.

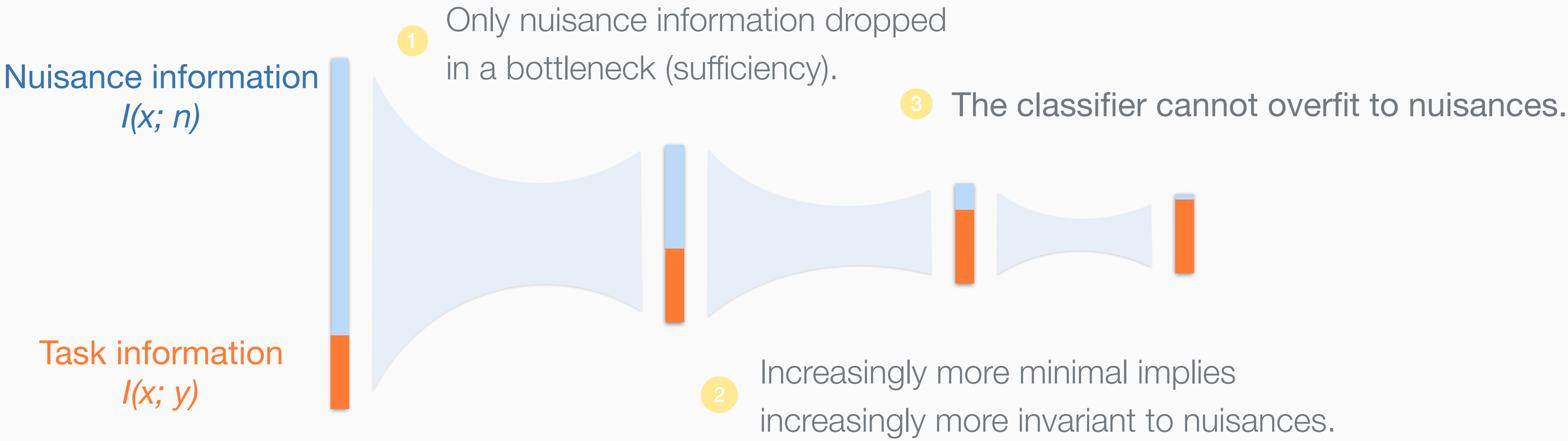
- A representation is maximally insensitive to all nuisances iff it is minimal

Corollary: Ways of enforcing invariance

The standard architecture alone already promotes invariant representations

Regularization by architecture

Reducing dimension (max-pooling) or adding noise (dropout) increases minimality and invariance.



Stacking layers

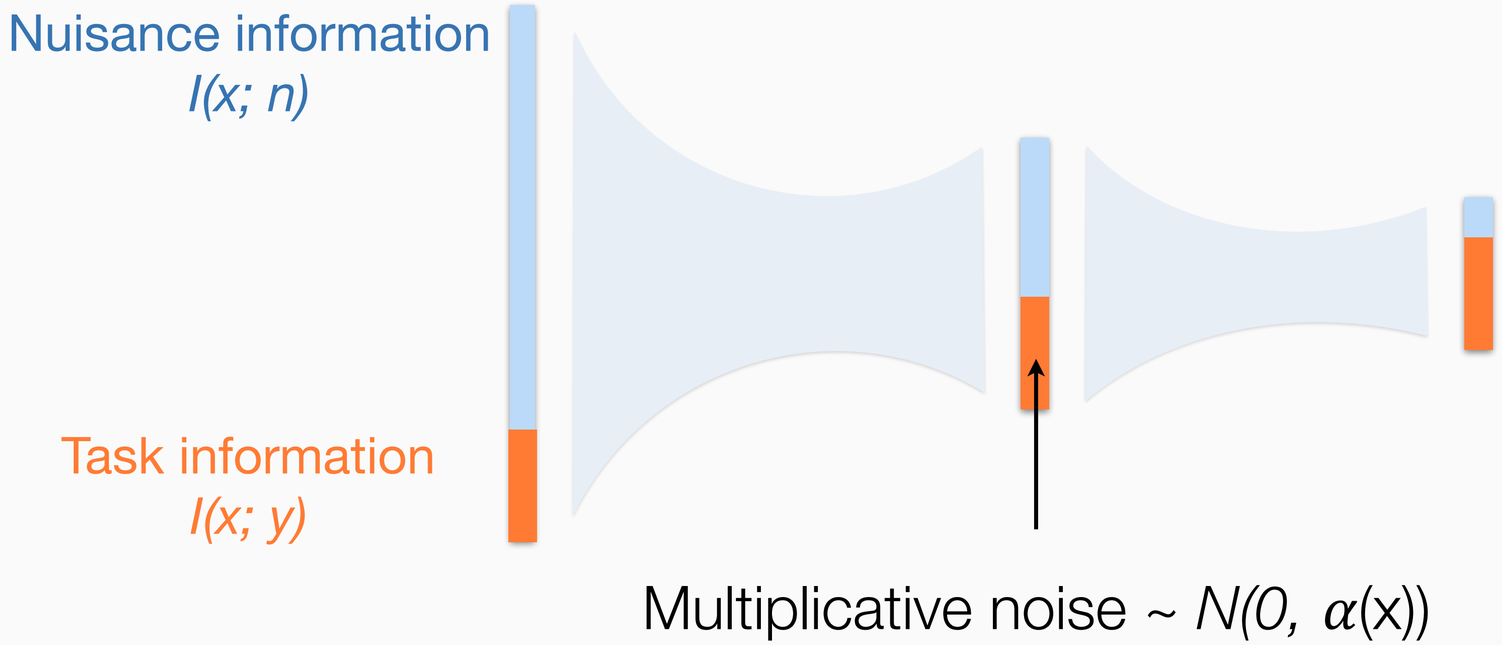
Stacking multiple layers makes the representation increasingly minimal.

Information Dropout: a Variational Bottleneck

Creating a soft bottleneck with controlled noise

$$\mathcal{L} = H_{p,q}(y|z) + \beta\, I(z;x) = H_{p,q}(y|z) - \beta \log \alpha(x)$$

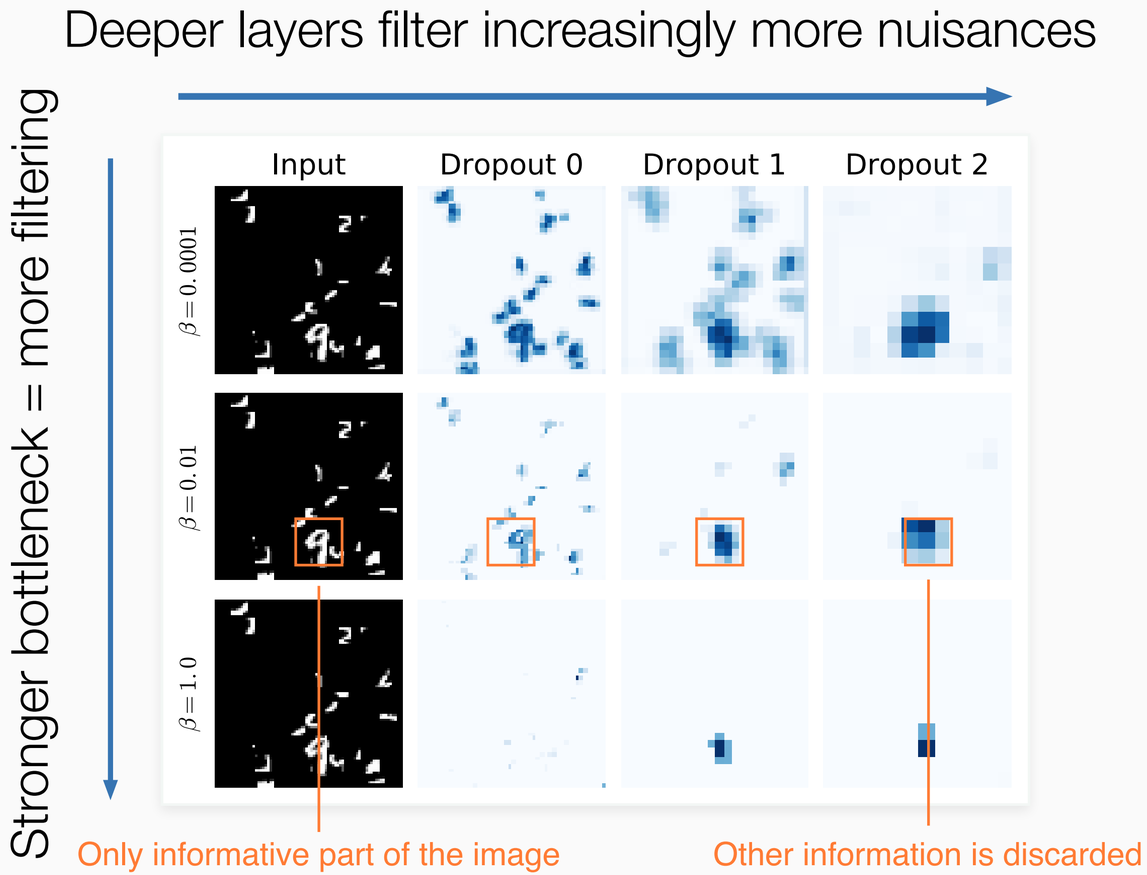
bottleneck



Achille and Soatto, *"Information Dropout: Learning Optimal Representations Through Noisy Computation"*, PAMI 2018 (arXiv 2016)

Learning invariant representations

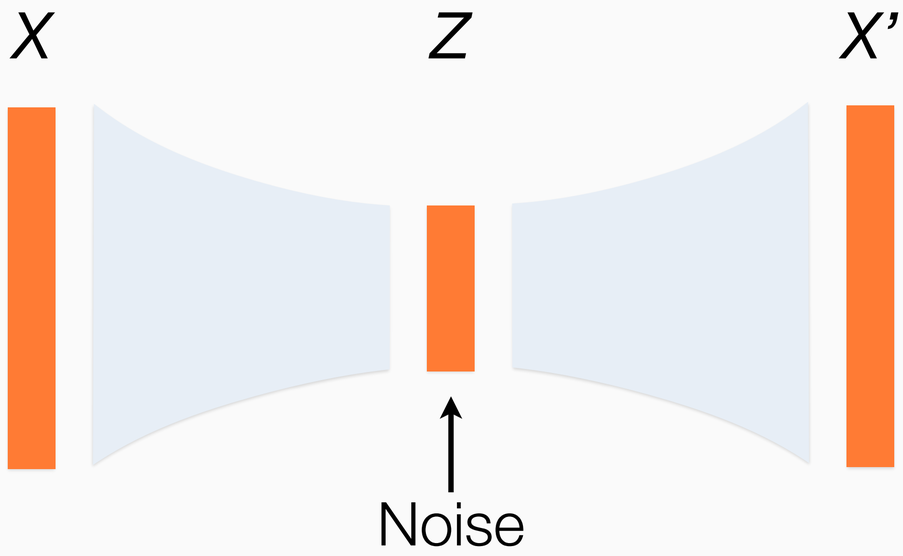
(Achille and Soatto, 2017)



Achille and Soatto, *"Information Dropout: Learning Optimal Representations Through Noisy Computation"*, PAMI 2018 (arXiv 2016)

Variational Auto-Encoders

(Kingma and Welling, 2014)



Loss function (VLBO):

$$\mathcal{L} = H_{p,q}(x|z) + \mathbb{E}_x[\mathrm{KL}(q(z|x)\|p(z))]$$

Factorized prior p(z)
rather than marginal q(z)

Or more generally with a β:

$$\mathcal{L} = H_{p,q}(x|z) + \beta\mathbb{E}_x[\mathrm{KL}(q(z|x)\|p(z))]$$

Kingma and Welling, *Auto-Encoding Variational Bayes, ICLR 2014*

Higgins et al., *β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, ICLR 2017

Achille and Soatto, *"Information Dropout: Learning Optimal Representations Through Noisy Computation"*, PAMI 2018 (arXiv 2016)

VAEs and disentanglement

A β-VAE minimizes the loss function

$$\mathcal{L} = H_{p,q}(x|z) + \beta \mathbb{E}_x[\mathrm{KL}(q(z|x)||p(z))]$$

If p(z|x) and p(z) are factorized (noise is independent), this is equivalent to

$$= H_{p,q}(x|z) + \beta \{\underbrace{I(z;x)}_{\text{Minimality}} + \underbrace{\mathrm{TC}(z)}_{\text{Disentanglement}}\}$$

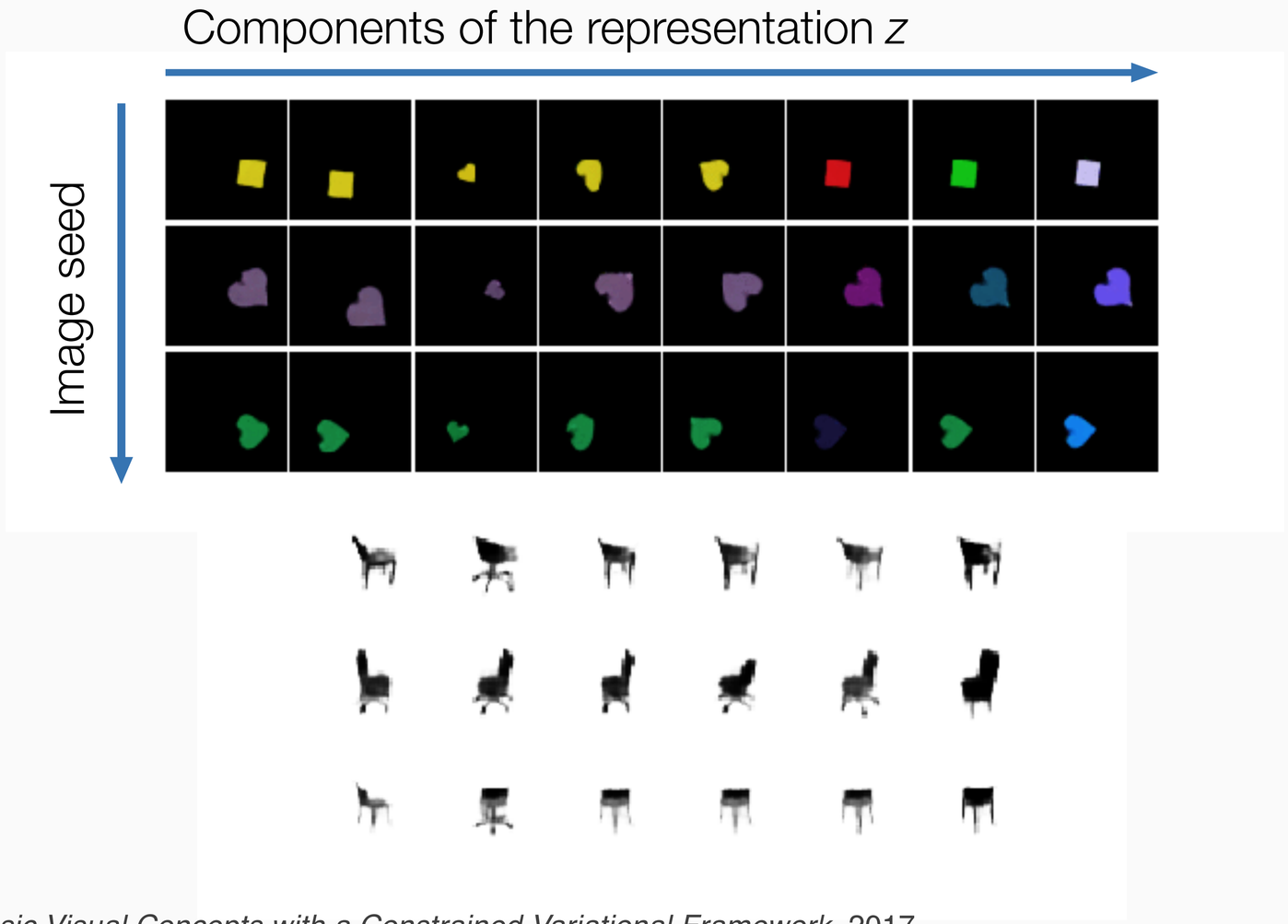
Proposition (Achille and Soatto, 2017). Assuming a factorized prior for z, a β-VAE optimizes both for the IB Lagrangian and for disentanglement.

Higgins et al., *β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, ICLR 2017
Achille and Soatto, *"Information Dropout: Learning Optimal Representations Through Noisy Computation"*, PAMI 2018 (arXiv 2016)

Learning disentangled representations

(Higgins et al., 2017, Burgess et al., 2017)

Each component of the learned representation corresponds to a different semantic factor.



Higgins et al., *β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, 2017
Burgess et al., *Understanding Disentangling in beta-VAE"* 2017

Pictures courtesy of Higgins et al., Burgess et al.

Is the representation “semantic” and domain invariant?

(Achille et al., *2018)*

Original

Reconstruction

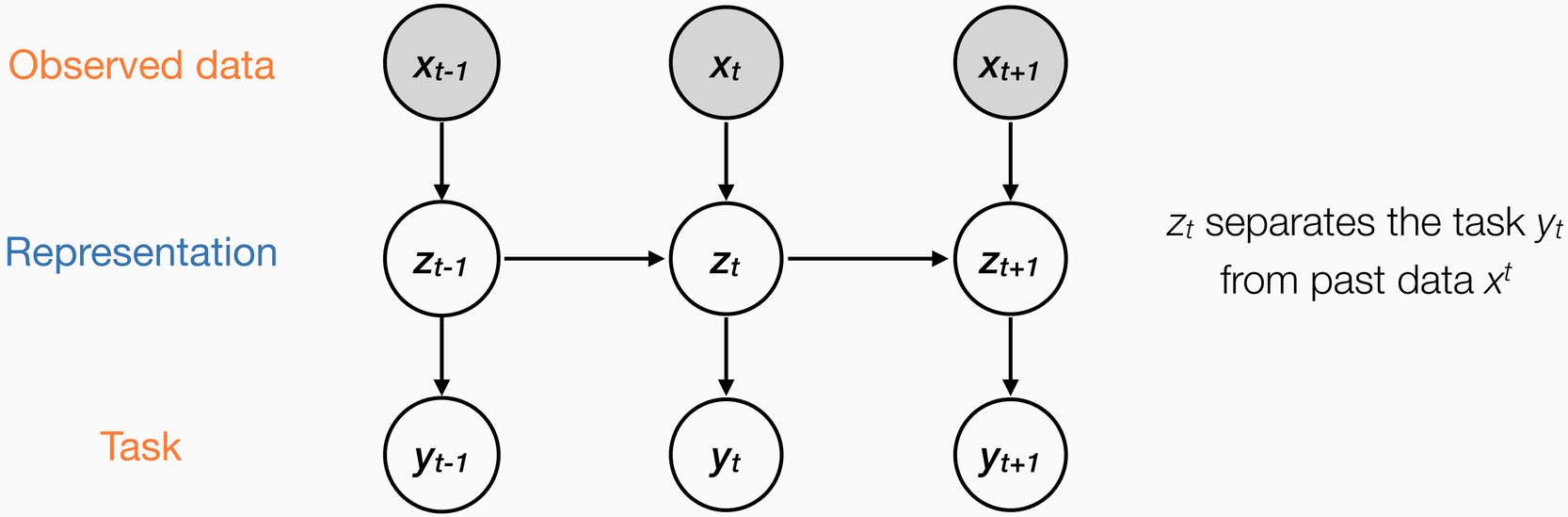
Cross-domain reconstruction

Achille et al., *Life-Long Disentangled Representation Learning with Cross-Domain Latent Homologies*, 2018

A Separation Principle for control

(Achille and Soatto; 2017)

Separating representation: a representation of all past data that is sufficient for control (*i.e*, a short term memory).



For linear control the hidden state of the Kalman filter is a separating representation.

But its complexity scales with the complexity of input data!

A Separation Principle for control

Proposition: A representation z_t that minimizes

$$L = \frac{1}{T} \sum_{t=0}^{T} \sum_{k=0}^{n} H(y_{t+k}|z_t) + \beta I(x^t; z_t)$$

is minimal and sufficient for n-step prediction of the task variable.

Note: The parameter β determines trade-off between complexity of the representation and sufficiency: Allows to ignore non-important long-term temporal dependencies.

Achille and Soatto, *"A Separation Principle for Control in the Age of Deep Learning"*, 2017

Summary so far

We want to learn minimal sufficient invariant representations of the input.

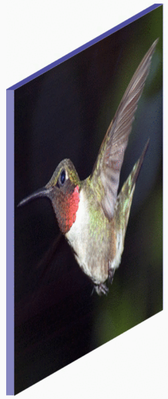
Reduce dimensionality (e.g., max-pooling)

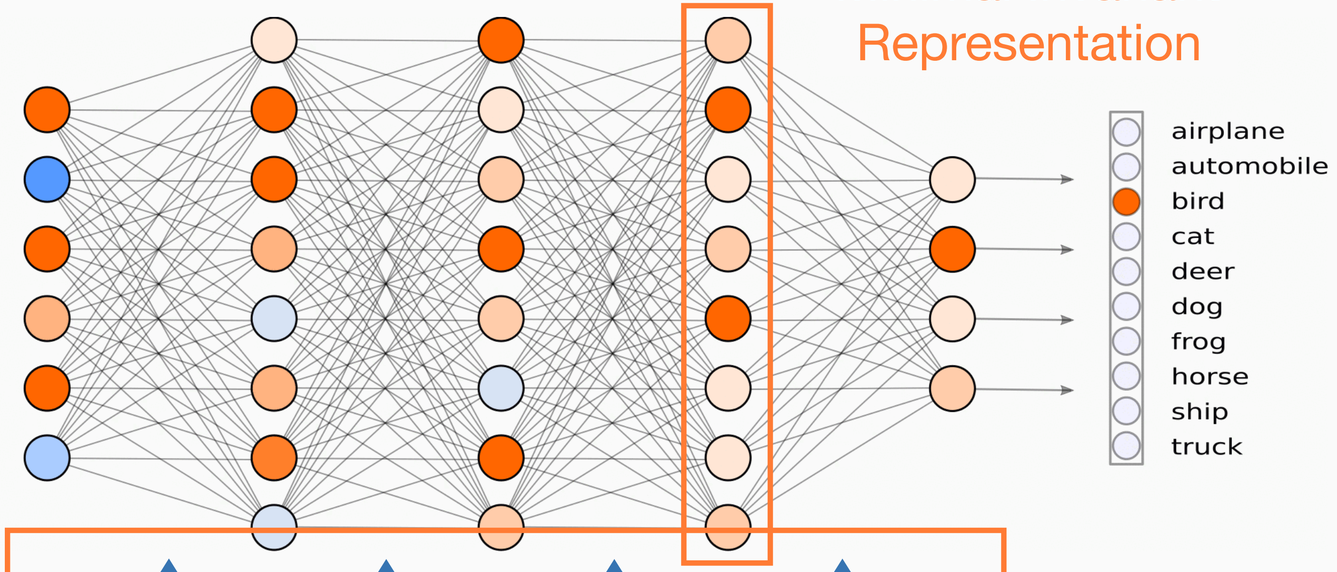
Or add noise to the activations!

We also gain some form of (semantic) disentanglement.

Minimal Invariant Representation

Test Image





Weights

Training Set

{  , (car, horse, deer, …) }

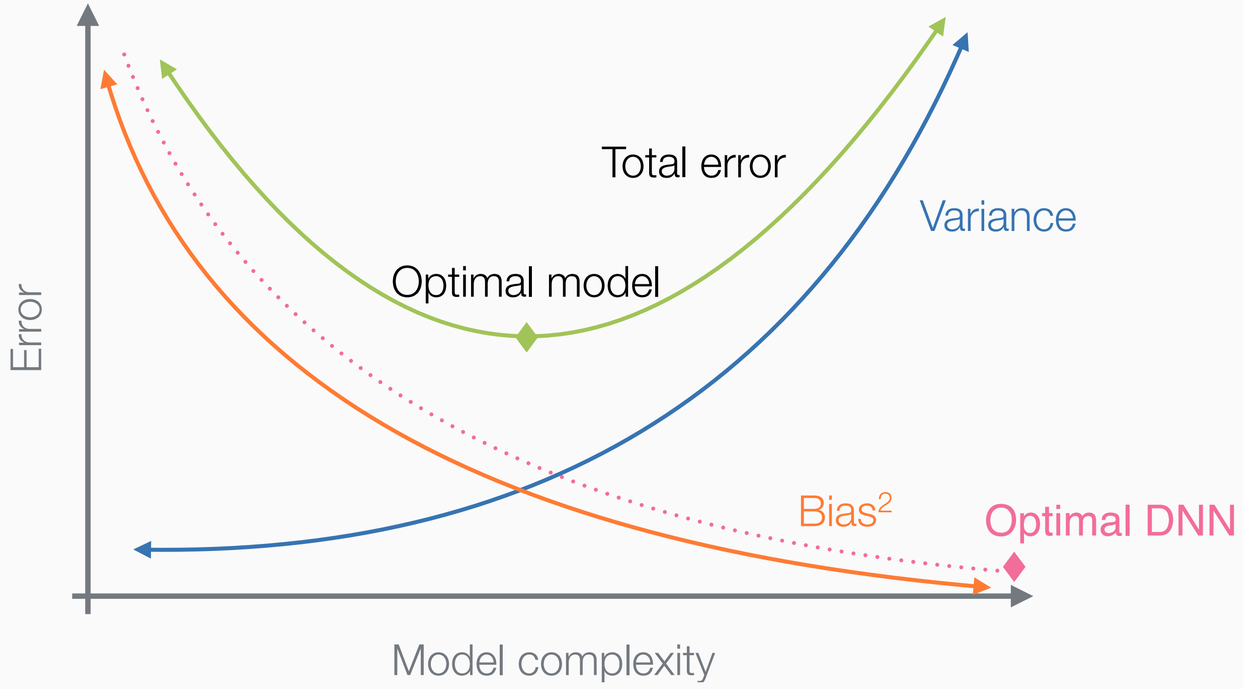
What about Deep Learning?
An IB for the weights

SGD and cross-entropy loss

Why deep networks do not overfit?

$$\mathcal{L} = H_{p,q}(y|z) + \beta I(z;x)$$

cross-entropy regularizer



Empirical loss and overfitting

Regularize the information in the weights to prevent memorization

$$\underbrace{H_{p,q}(\mathcal{D}\|w)}_{\text{cross-entropy}} = \underbrace{H(\mathcal{D}|\theta)}_{\text{intrinsic error}} + \underbrace{I(\theta;\mathcal{D}|w)}_{\text{sufficiency}} + \underbrace{\mathrm{KL}(q\,\|\,p)}_{\text{model efficiency}} - \underbrace{I(\mathcal{D};w|\theta)}_{\text{overfitting}}$$

Eliminate the negative term:

$$L(w) = H_{p,q}(\mathcal{D}|w) + I(\mathcal{D};w|\theta)$$

Intractable. Rather, add Lagrange multiplier to bound I(D; w):

$$L(w) = H_{p,q}(\mathcal{D}|w) + \beta\ I(\mathcal{D};w)$$

Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018

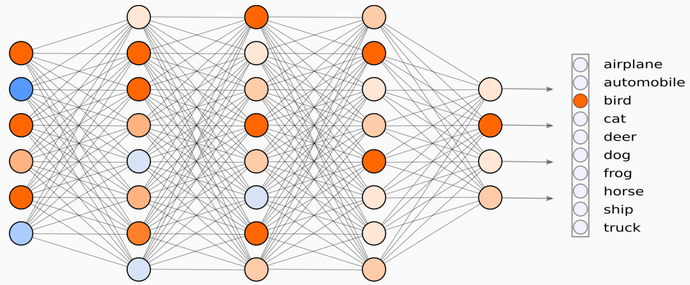
An Information-Bottleneck for the weights

Training data

{  , (car, horse, deer, …) }

Weights

Testing



The weights should be a minimal sufficient statistic of the training set for test prediction.

In practice: a generalized Bayesian loss

Recall that we have the upper-bound:

$$\begin{aligned} I(\mathcal{D}; w) &= \mathbb{E}_{\mathcal{D}}[KL(p(w|\mathcal{D})\|p(w))] \\ &\leq \mathbb{E}_{\mathcal{D}}[KL(q(w|\mathcal{D})\|p(w))] \end{aligned}$$

Then, for a fixed dataset *D*, the loss becomes:

$$L(w) = H_{p,q}(\mathcal{D}|w) + \beta\, KL(q(w|\mathcal{D})\|p(w))$$

For $\beta=1$ we recover the standard Bayesian Variational Lower-Bound!

Methods and observations to minimize the VLBO still applies.

Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018

The PAC-Bayes generalization bound

Catoni, 2007; McAllester 2013

PAC-Bayes bound (Catoni, 2007; McAllester 2013).

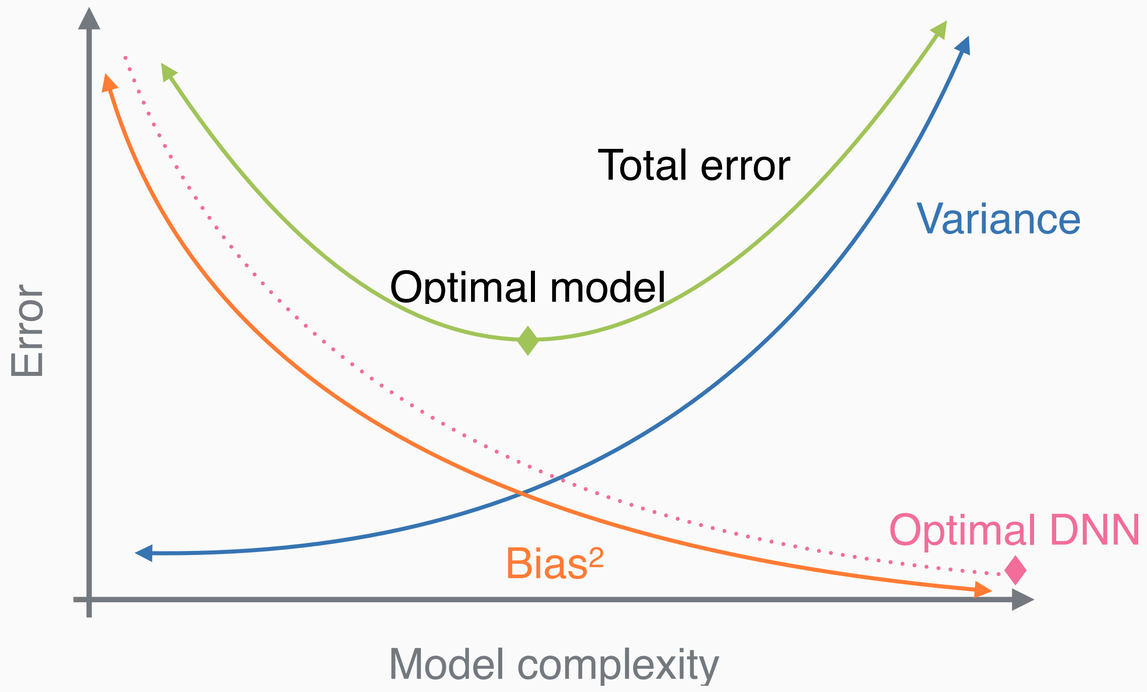
$$L_{\text{test}}(q(w|\mathcal{D})) \leq \frac{1}{N(1-1/2\beta)} \underbrace{\left[H_{p,q}(y|x,w) + \beta\ \mathrm{KL}(q(w|\mathcal{D})\|p(w))\right]}_{\text{IB Lagrangian for the weights}}$$

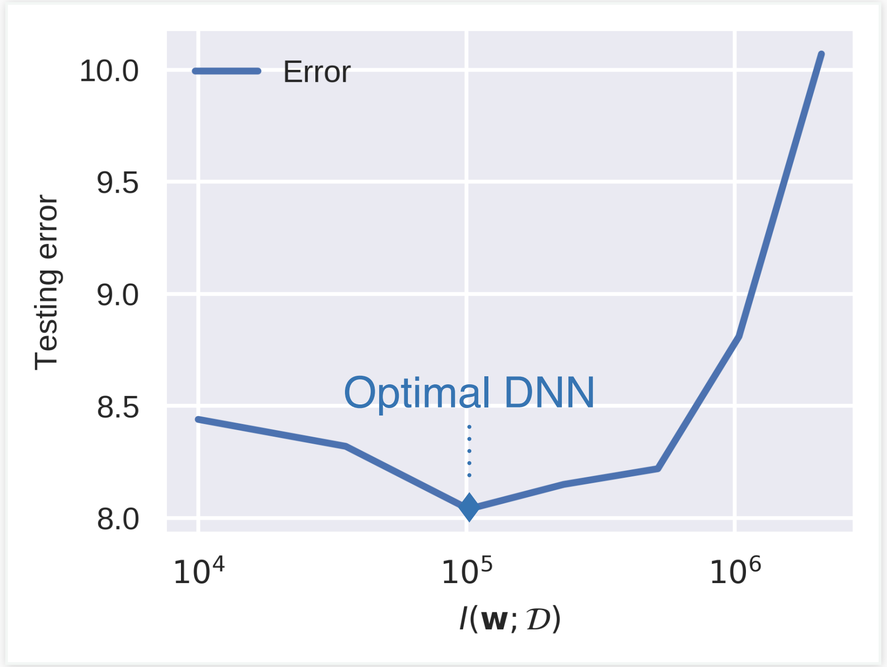
Corollary. Minimizing the IB Lagrangian for the weights minimizes an upper bound on the test error (Dziugaite and Roy, 2017; Achille and Soatto, 2017)

This gives non-vacuous generalization bounds! (Dziugaite and Roy, 2017)

Bias-variance tradeoff

Information is a better measure of complexity





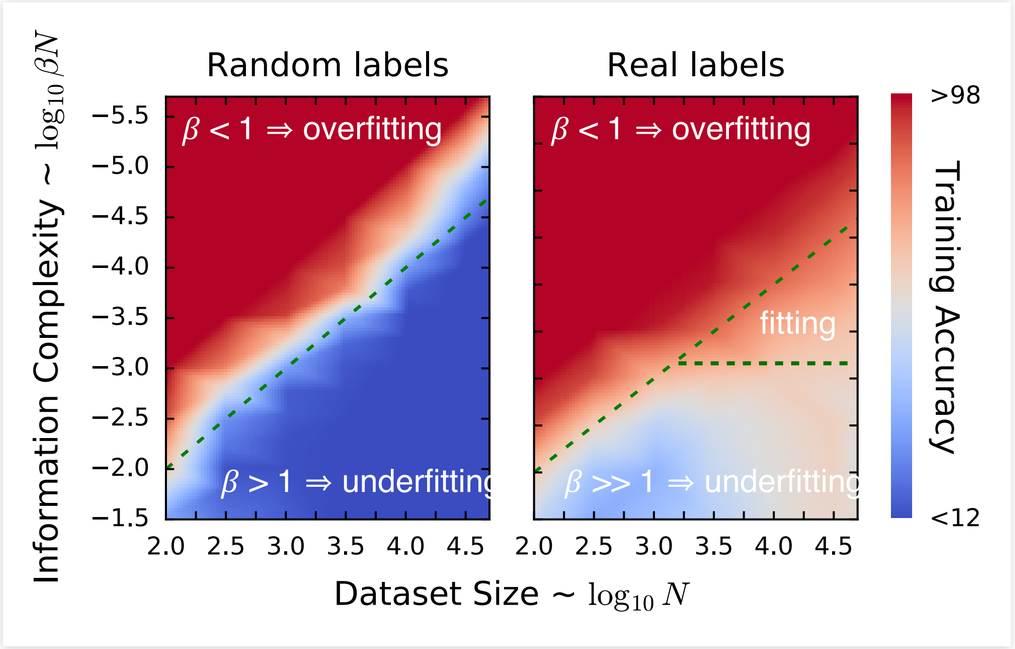
Information complexity

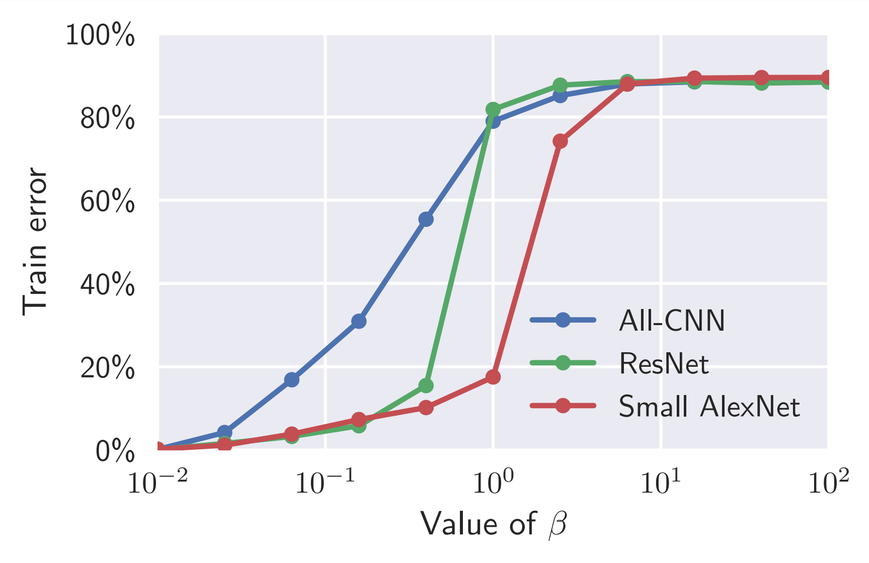
Parametrizing the complexity with information in the weights, we recover bias-variance trade-off trend.

Phase transition

For random labels sharp transition from overfitting to underfitting

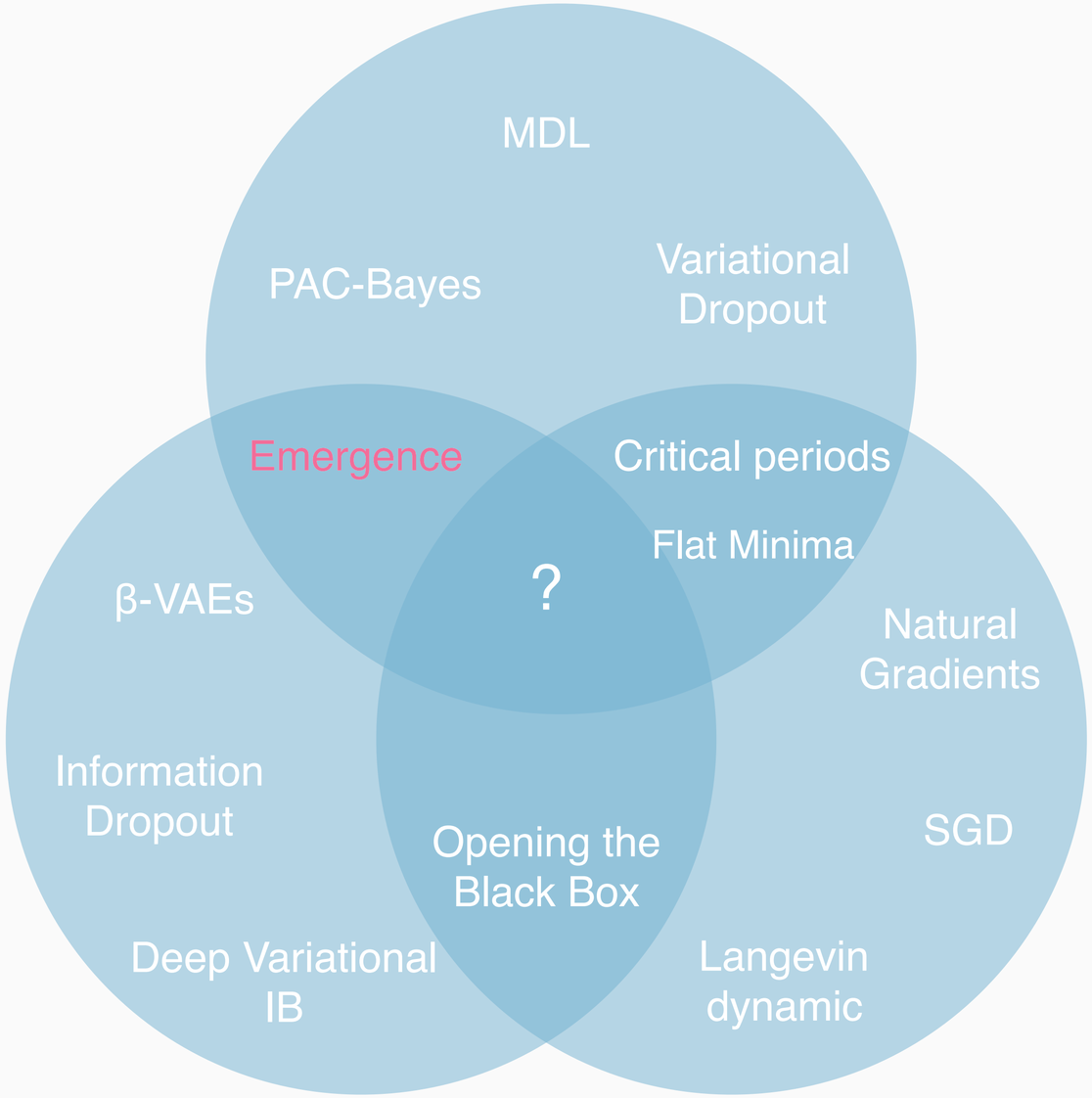
For random labels, at $\beta = 1$ (the VLBO value) there is a phase transition between overfitting and underfitting.





Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018

Information in weights

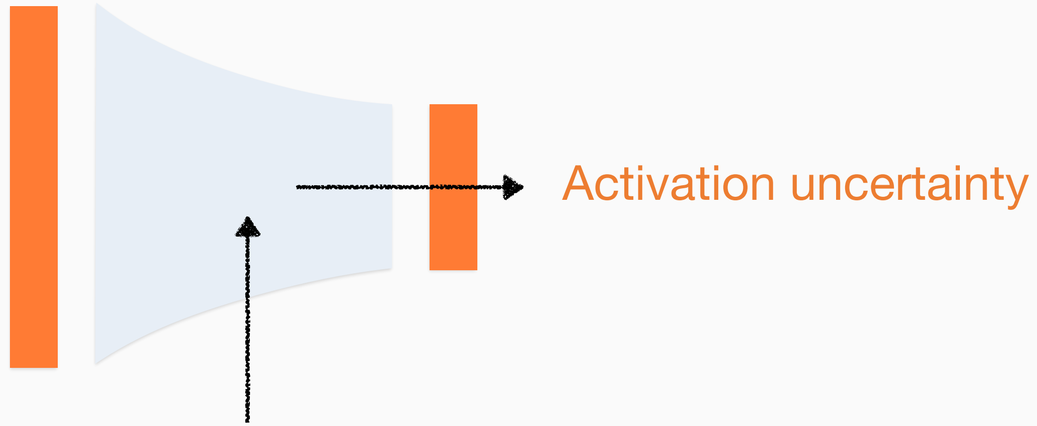


Information in activations

Optimization

Information in Activations and Information Weights

Compressed weights ⇒ Compressed activations



The Emergence Bound. For a single layer z = w x, we have the tight bound

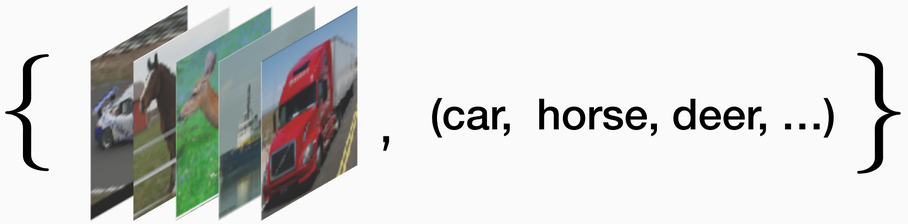
$$I(z;x) + TC(z) \leq g(\, I(w;\mathcal{D})\,) + O(1/\dim(x)),$$

where *g* is a strictly increasing function.

Corollary. Less information in the *weights* increases invariance and disentanglement of the learned representation.

Compression of the **weights** biases toward invariant and disentangled **representations**.

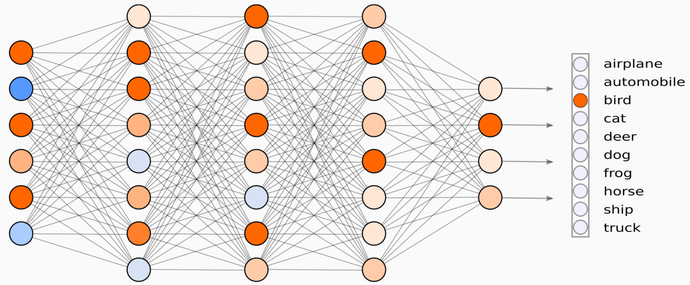
Training data



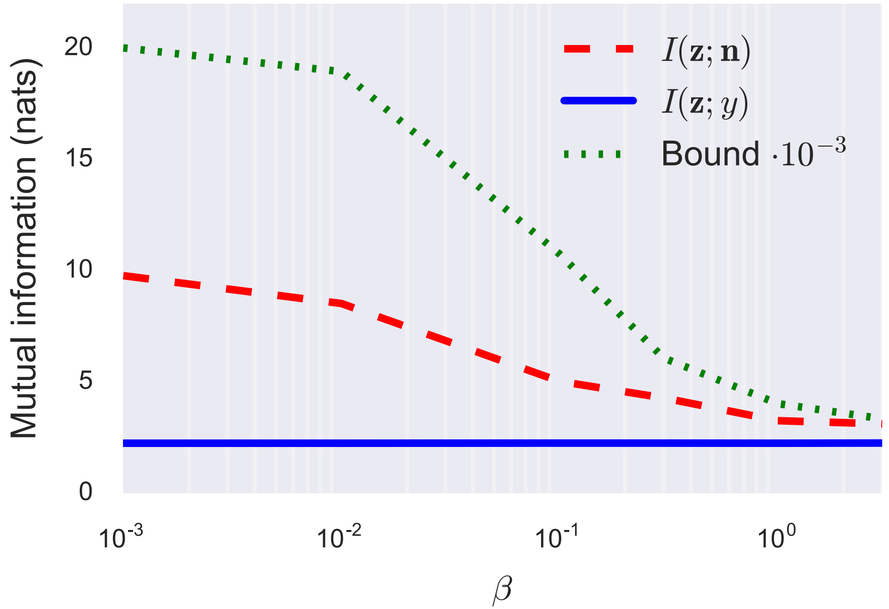
Weights

Invariant representation

Testing



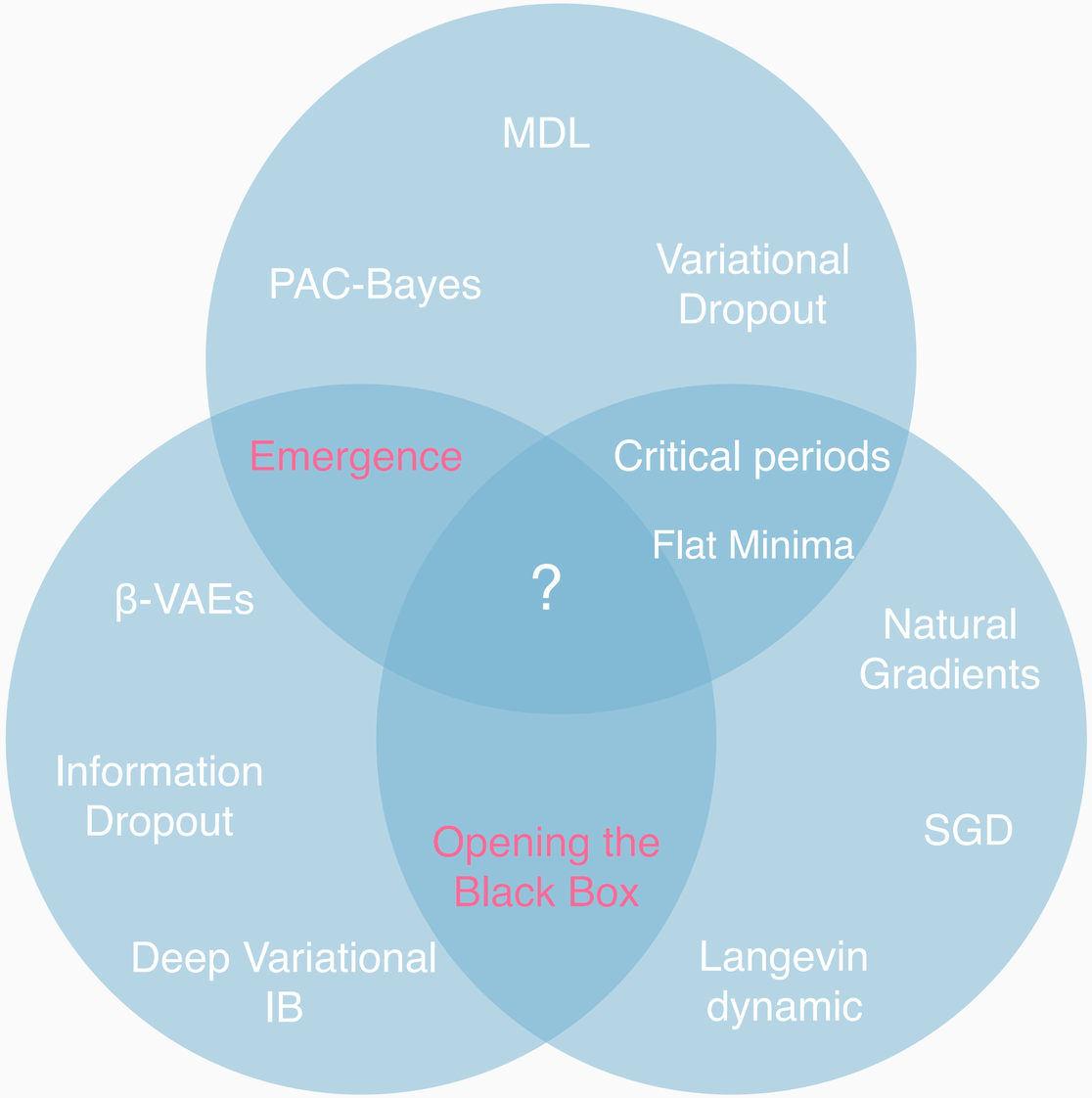
Learning occlusion invariance



Increasing IB weight regularization, the learned representation is increasingly more invariant.

Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018

Information in weights

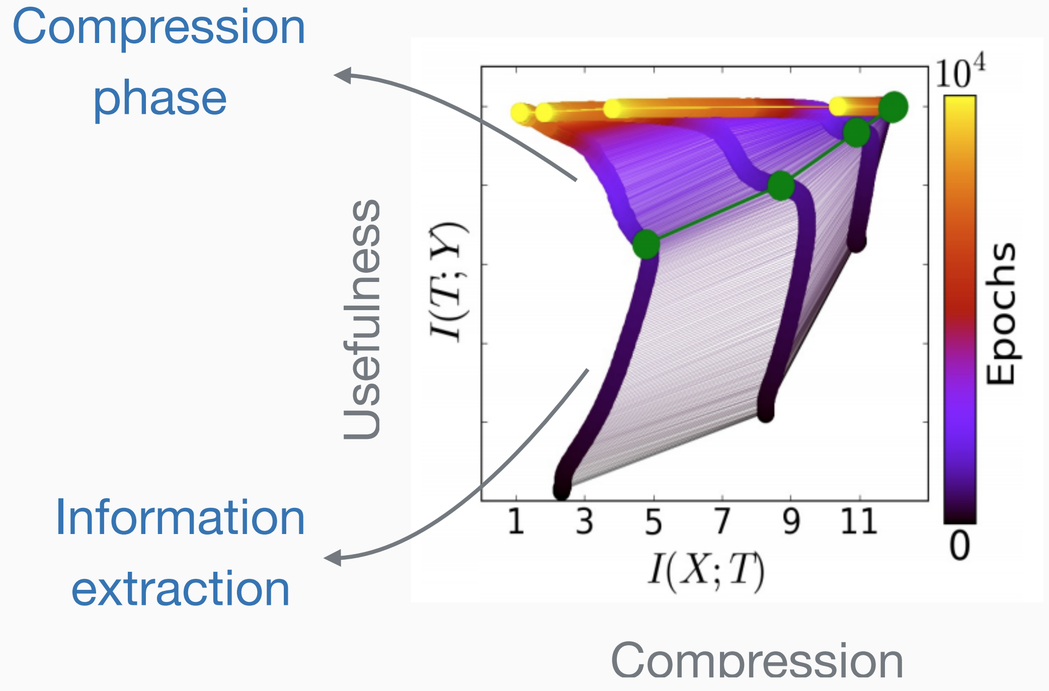


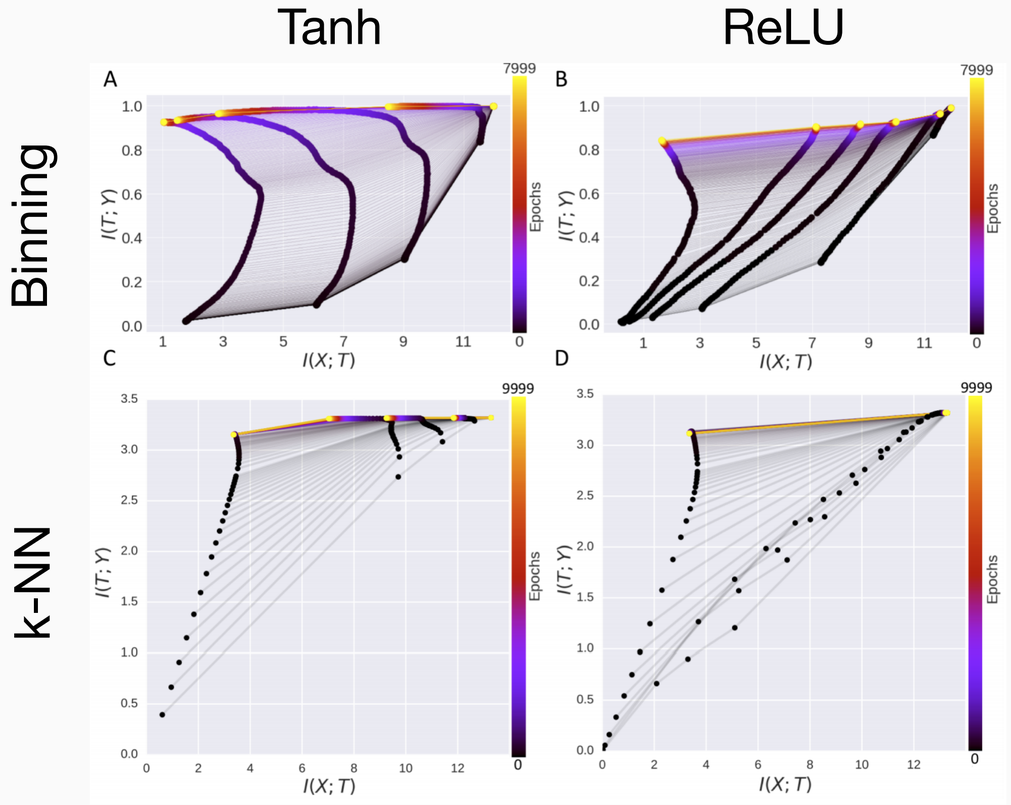
Information in activations

Optimization

Information in the activations and SGD

Shwartz-Ziv and Tishby, 2017

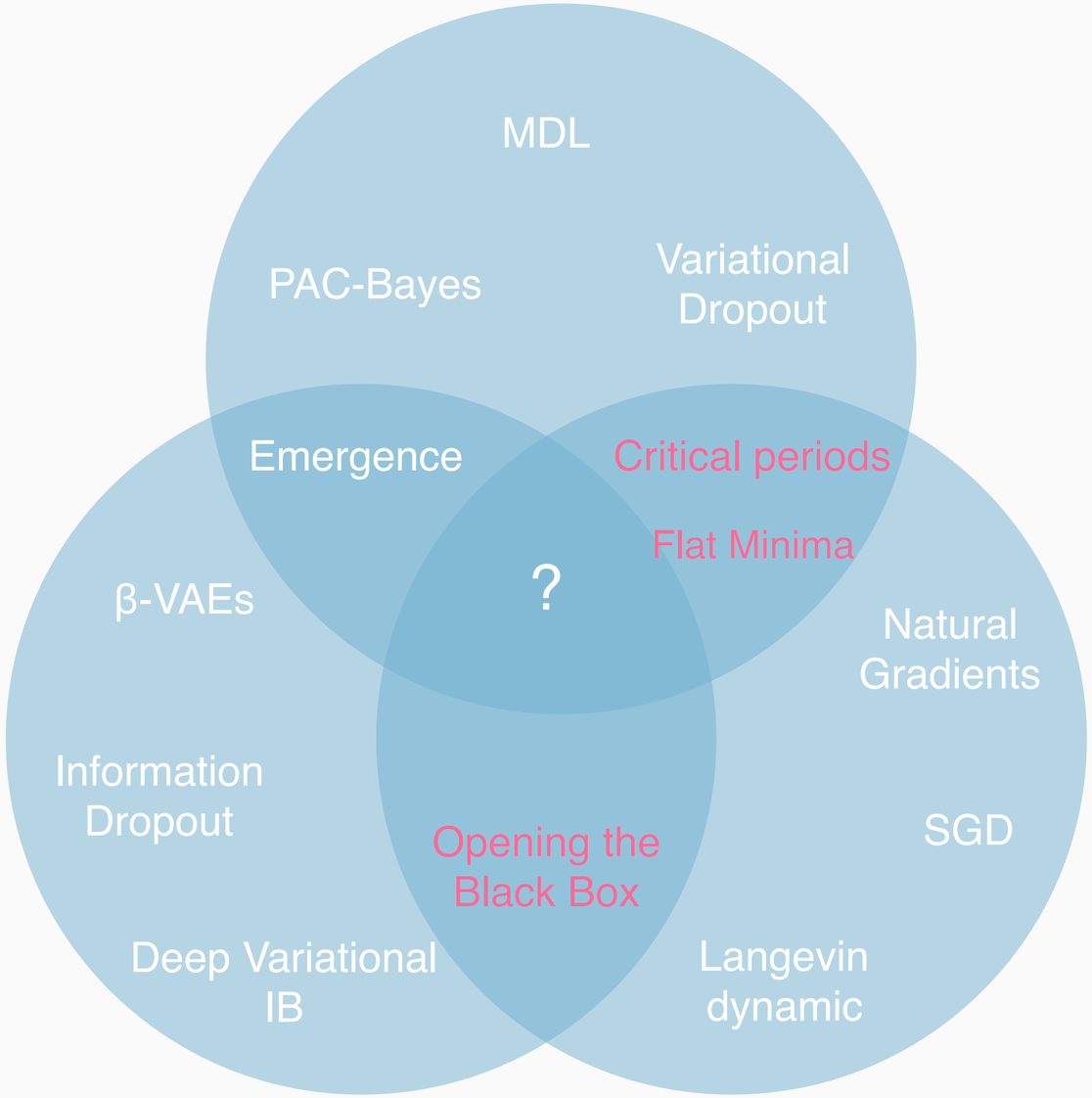




Shwartz-Ziv and Tishby, *Opening the black box …*, 2017

Saxe et al., *On the Information Bottleneck Theory of Deep Learning*, 2018

Information in weights



Information in activations

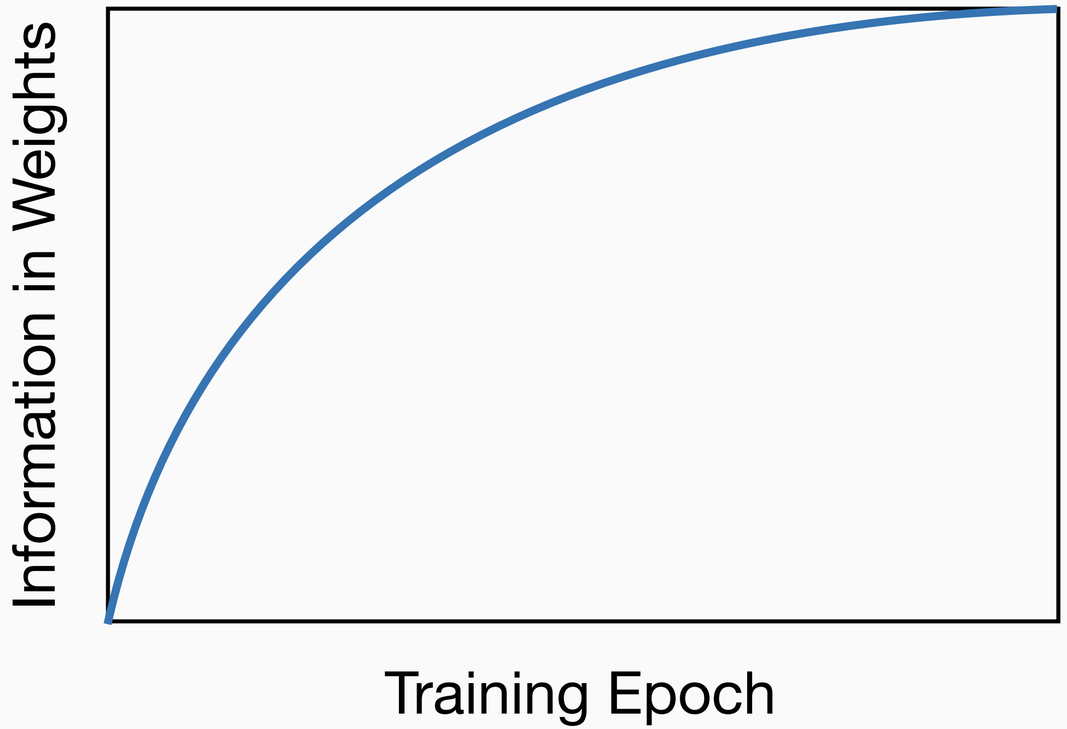
Optimization

The Dynamics of Learning: Critical learning periods

Information in Weights during training

What should we expect from the information in the weights during training?

Maybe something like this?



Information in Weights and the Fisher Information Matrix

Let w^* be a local minimum. Then for a standard normal prior and choosing the optimal posterior $q(w|D)$ centered in w^* we have:

$$KL(q(w|\mathcal{D})\|p(w)) = \frac{\|w^*\|^2}{\lambda^2} + \log|2\lambda^2 NF(w^*) + I|$$

where $F(w^*)$ is the Fisher Information Matrix computed in w^*.

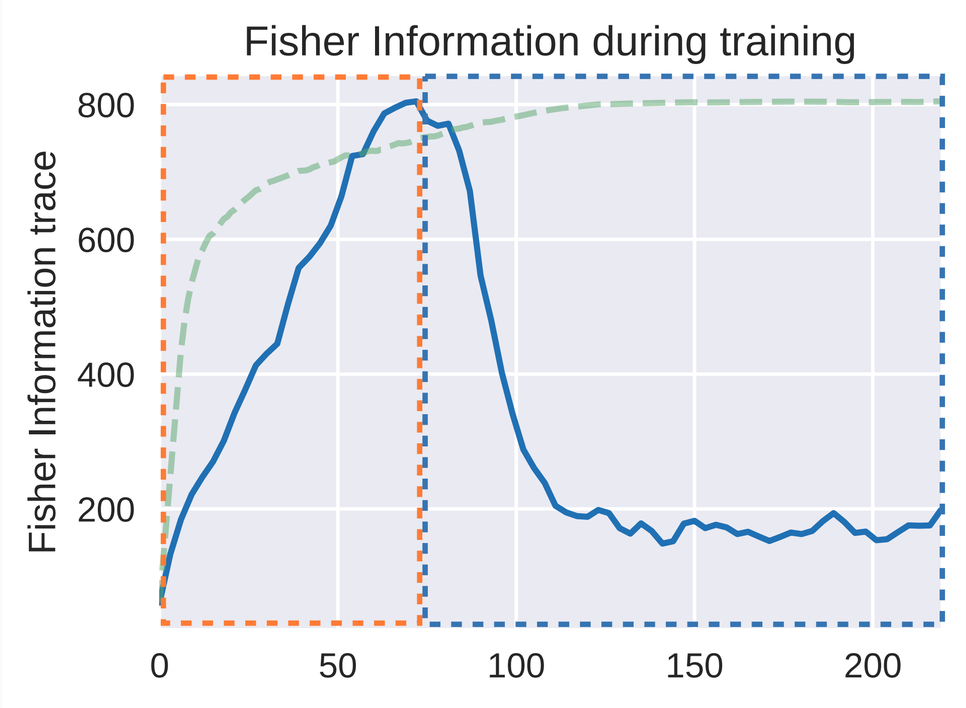
Note: The FIM computed in a local minimum is equal to the Hessian (Martens, 2014)

Corollary: Flat minima have low information in the weights.

Information during training

Information extraction

Information consolidation

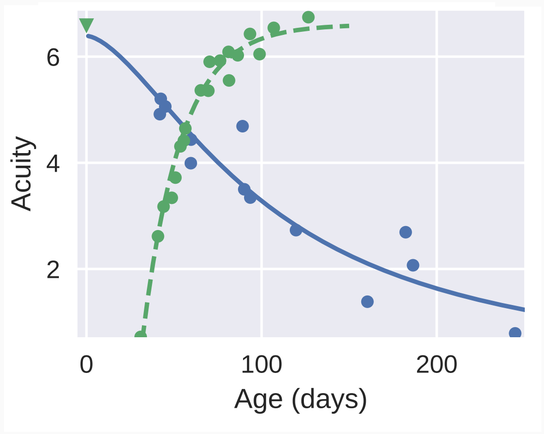


Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

Critical periods

Critical periods: A time-period in early development where sensory deficits can permanently impair the acquisition of a skill

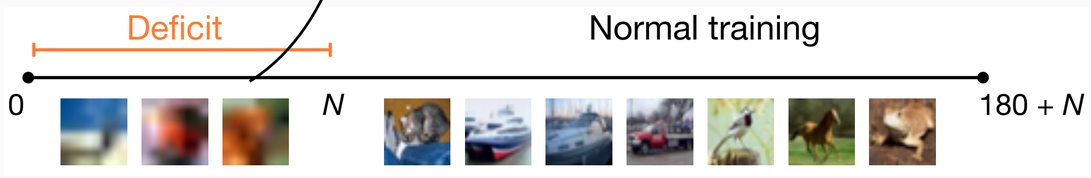
Examples: monocular deprivation, cataracts, imprinting, language acquisition

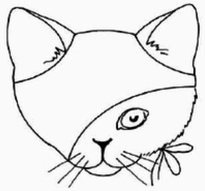


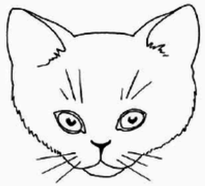
Kitten does not recover vision in covered eye

Critical periods in Deep Networks

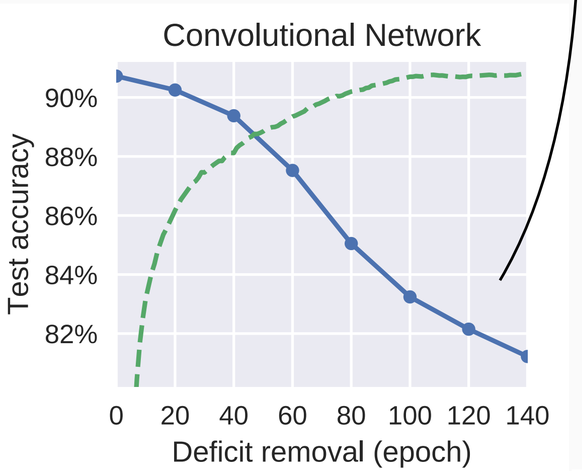
Show network blurred images to simulate cataract

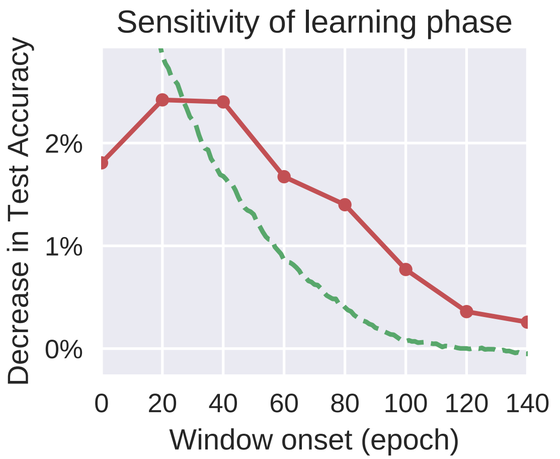


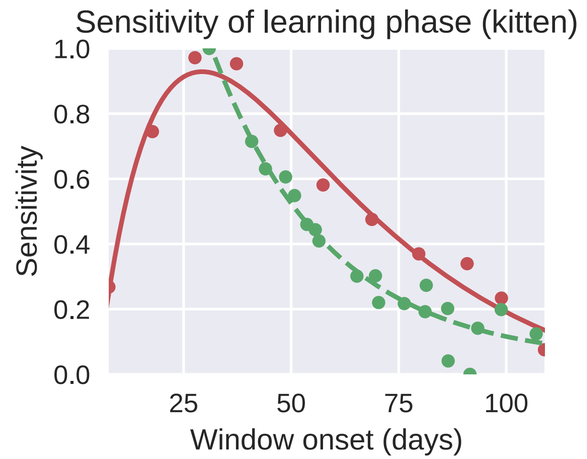




The network does not classify correctly if the deficit is removed to late





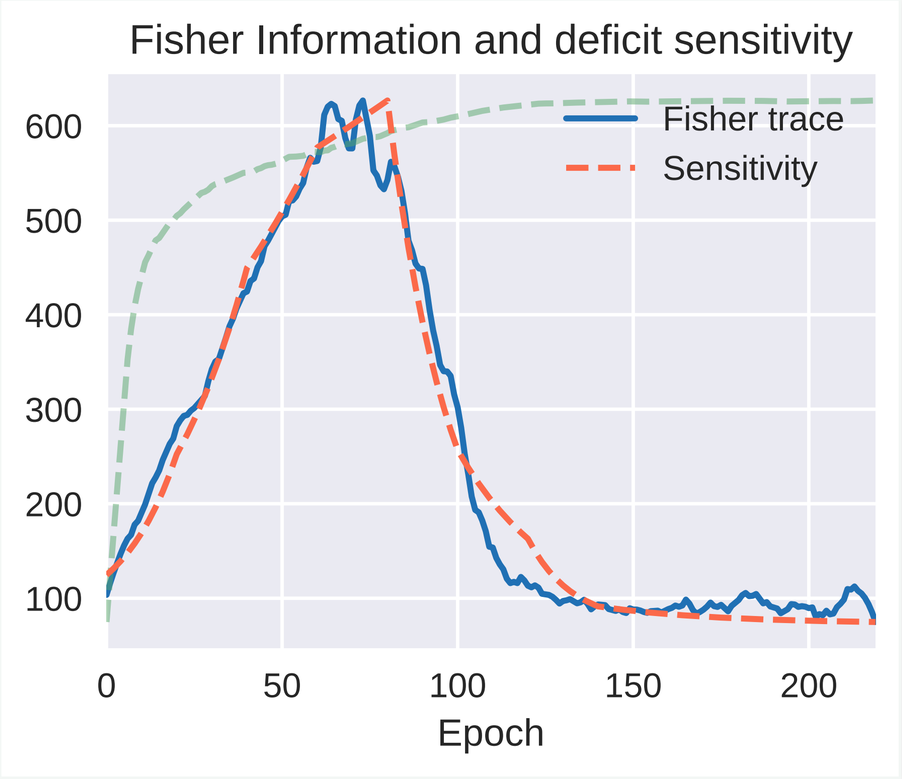


A short deficit at epoch ~40 is enough to permanently damage the network!

Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

Critical learning periods and Information in Weights

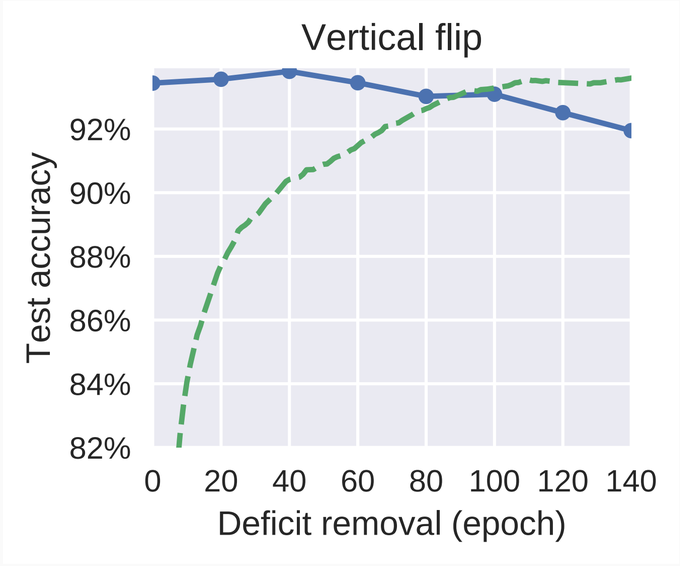
Sensitivity to deficits peaks when network is absorbing information.
Is minimal when the network is consolidating information.

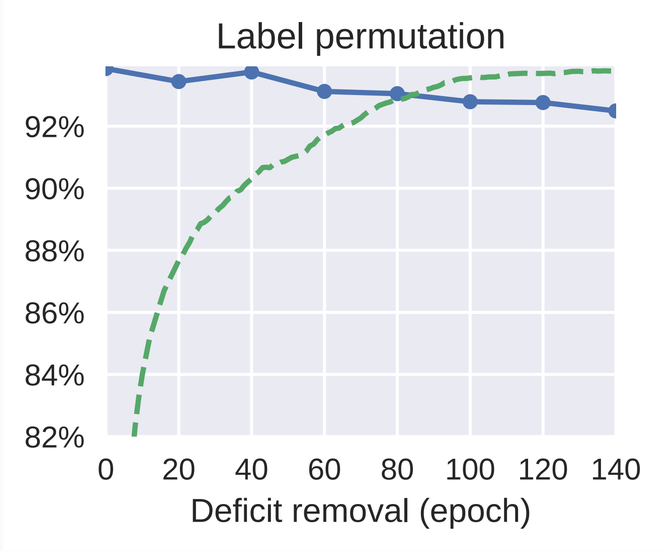


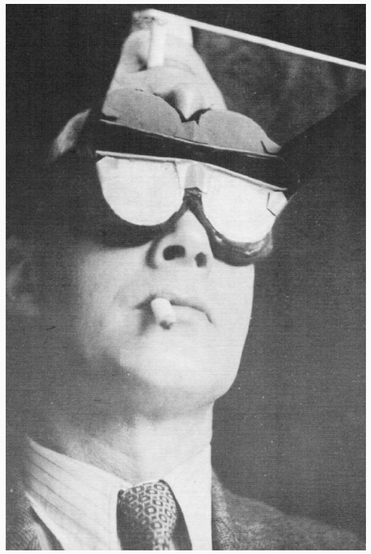
Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

High-level deficits do not have a critical period

Deficits that only change high-level statistics of the data do not show a critical period.



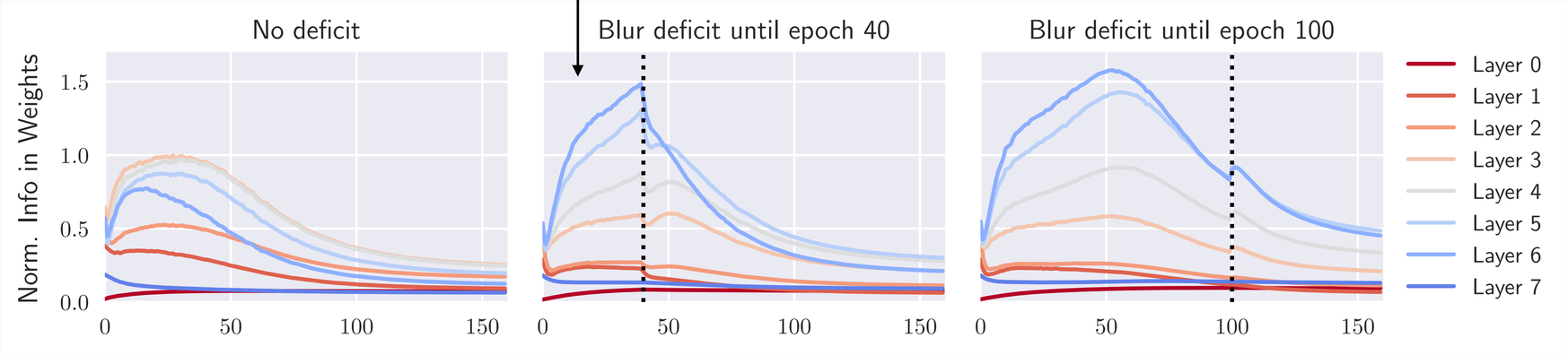


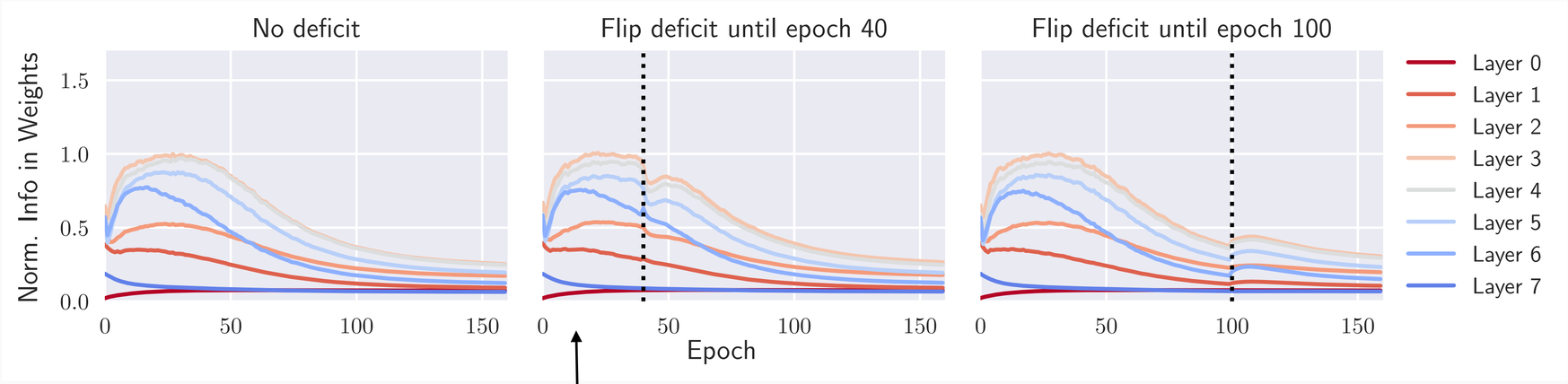


Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*
Picture from *"The world is upside down" — The Innsbruck Goggle Experiments of Theodor Erismann and Ivo Kohler,* Sachse et al.

Critical periods and changes of connectivity

Introducing a blur deficit
changes layer organization





High-level deficit do not change
layer organization

Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

Summary

Two types of compression: of activations and of weights.

- Compression of activations = invariance, disentanglement
- Compression of weights = generalization

Compression of weights implies compression of activations (Emergence bound)

Compression can happen explicitly adding noise or implicitly through optimization.

The initial transient phase of SGD is far from trivial!