CS 103: Representation Learning, Information Theory and Control

Lecture 5, Feb 8, 2019

Representation Learning and Information Bottleneck

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

Sufficient  Minimal

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:

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

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. (A. 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

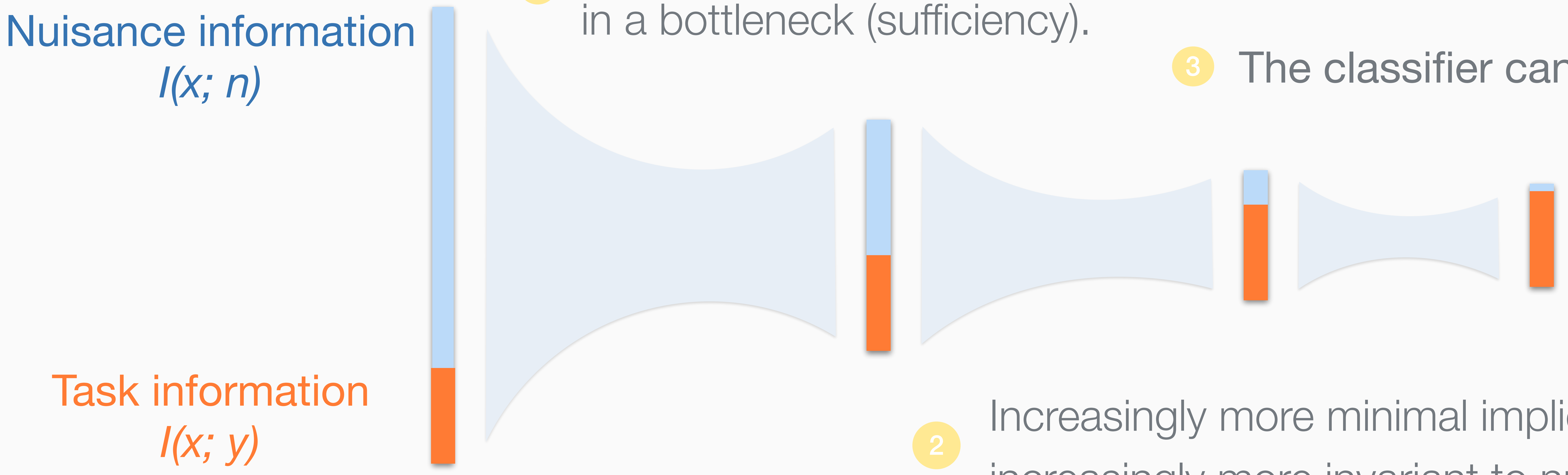
The standard architecture alone already promotes invariant representations

Regularization by architecture

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

1 Only nuisance information dropped in a bottleneck (sufficiency).

3 The classifier cannot overfit to nuisances.



2 Increasingly more minimal implies increasingly more invariant to nuisances.

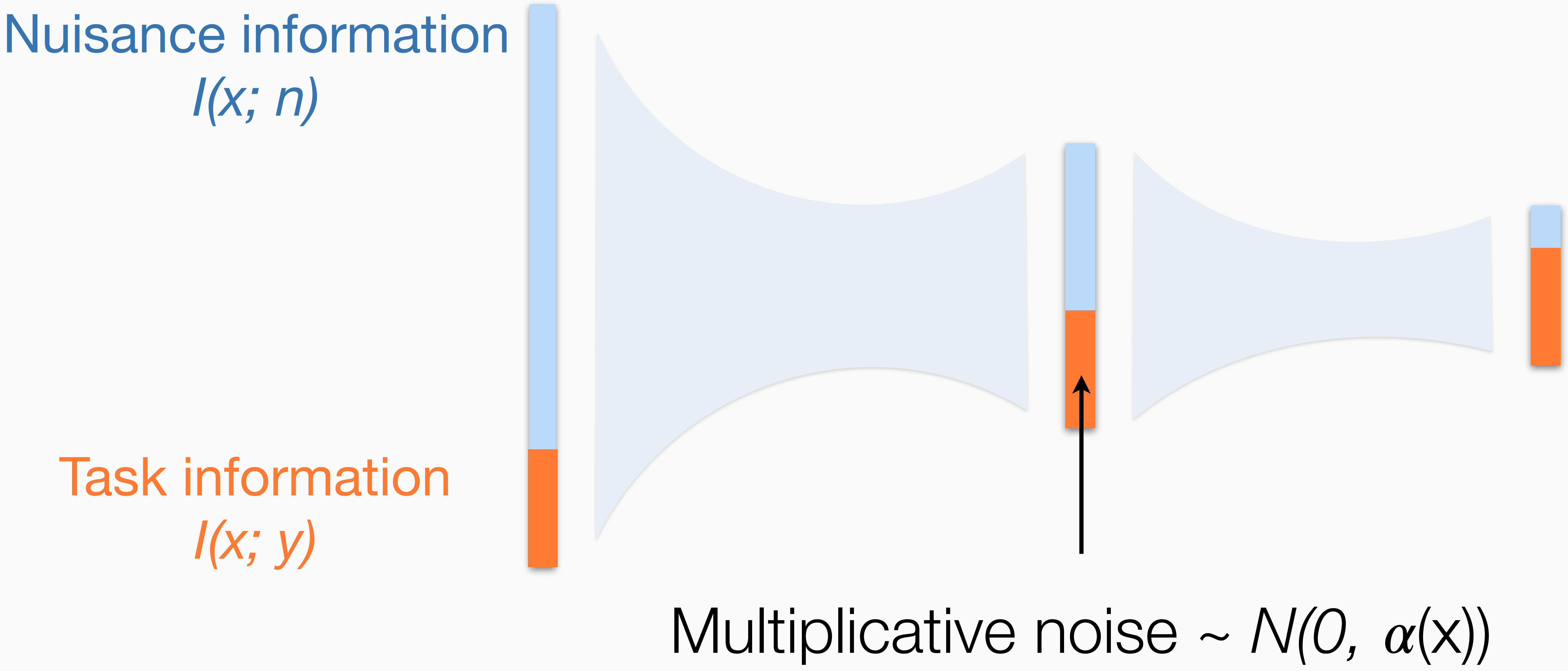
Stacking layers

Stacking multiple layers makes the representation increasingly minimal.

Information Dropout: a Variational Bottleneck

Creating a soft bottleneck with controlled noise

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

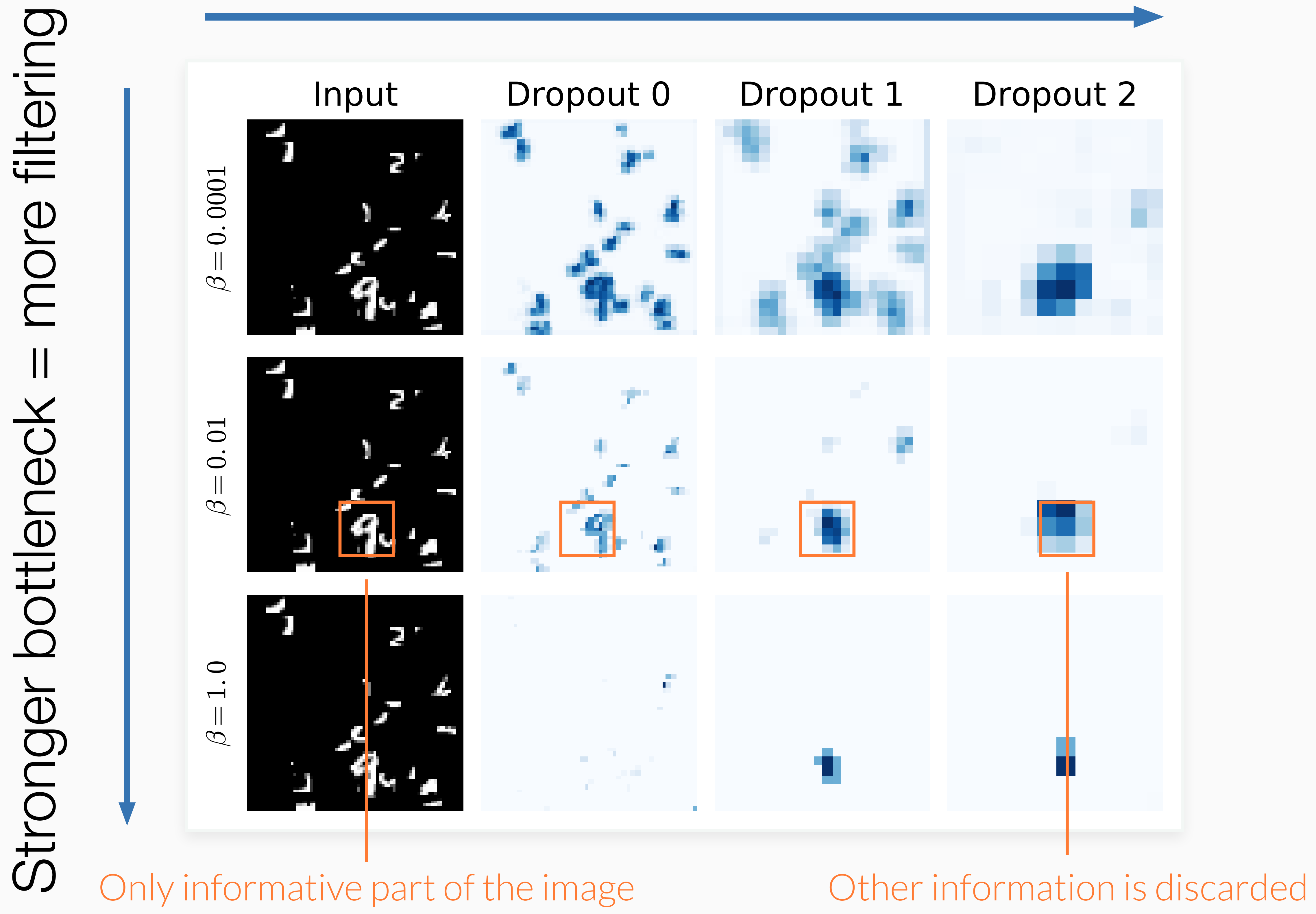


Achille and Soatto, "*Information Dropout: Learning Optimal Representations Through Noisy Computation*", PAMI 2018 (arXiv 2016)

Learning invariant representations

(Achille and Soatto, 2017)

Deeper layers filter increasingly more nuisances



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