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 $n \perp y$, then $I(n; z) = 0$
- **Maximally disentangled**: $TC(z) = KL(p(z)\|\prod_i p(z_i))$ is minimized
Information Bottleneck Lagrangian

A minimal sufficient representation is the solution to:

$$\text{minimize}_{p(z|x)} \quad I(x; z)$$

s.t. \quad H(y|z) = H(y|x)

Information Bottleneck Lagrangian:

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

cross-entropy  \quad \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,

$$l(z; n) \leq l(z; x) - l(x; y)$$

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.

Nuisance information $l(x; n)$

Only nuisance information dropped in a bottleneck (sufficiency).

Task information $l(x; y)$

Increasingly more minimal implies increasingly more invariant to nuisances.

The classifier cannot overfit to nuisances.

Stacking layers
Stacking multiple layers makes the representation increasingly minimal.
Information Dropout: a Variational Bottleneck
Creating a soft bottleneck with controlled noise

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

Nuisance information \( I(x; n) \)
Task information \( I(x; y) \)
Multiplicative noise \( \sim N(0, \alpha(x)) \)

Learning invariant representations

(Achille and Soatto, 2017)

Deeper layers filter increasingly more nuisances

Stronger bottleneck = more filtering

Only informative part of the image

Other information is discarded