Minimal Achievable Sufficient Statistic Learning

Milan Cvitkovic
California Institute of Technology/Amazon Web Services

Günther Koliander
Austrian Academy of Sciences

Goal: Find representation Z of X that is useful for producing Y

$$X \rightarrow Z \rightarrow Y$$

Goal: Find representation Z of X that is useful for producing Y

Hypothesis: Z should be a ***minimal sufficient statistic*** of X for Y [1, 2]

$$X \rightarrow Z \rightarrow Y$$

[1] Adragni, et al. "Sufficient dimension reduction and prediction in regression" 2009
[2] Shamir et al. "Learning and generalization with the information bottleneck" 2010

Hypothesis: Z should be a ***minimal sufficient statistic*** of X for Y [1, 2]

$$X \rightarrow Z \rightarrow Y$$

Statistic: $Z = f(X)$

Sufficient: $p(X \mid Y, Z) = p(X \mid Z)$

Minimal: g(Z) isn't sufficient for any non-invertible g

[1] Adragni, et al. "Sufficient dimension reduction and prediction in regression" 2009
[2] Shamir et al. "Learning and generalization with the information bottleneck" 2010

Hypothesis: Z should be a ***minimal sufficient statistic*** of X for Y [1, 2]

$$X \to Z \to Y$$

Statistic: $Z = f(X)$

Sufficient: $p(X \mid Y, Z) = p(X \mid Z)$

Minimal: g(Z) isn't sufficient for any non-invertible g

A little strict for ML

[1] Adragni, et al. "Sufficient dimension reduction and prediction in regression" 2009
[2] Shamir et al. "Learning and generalization with the information bottleneck" 2010

Hypothesis: Z should be a ***minimal achievable sufficient statistic*** of X for Y

$$X \rightarrow Z \rightarrow Y$$

Statistic: $Z = f(X)$ for f in ***F***

Sufficient: $p(X \mid Y, Z) = p(X \mid Z)$

Minimal Achievable:

g(Z) isn't sufficient for any Lipschitz, non-invertible g where $g \circ f$ in ***F***

How do we find a minimal (achievable) sufficient statistic?

X discrete: ✔

$$f \in \arg\min_{S \in \mathcal{F}} \; I(X, S(X))$$

$$s.t. \;\; I(S(X), Y) = \max_{S' \in \mathcal{F}} I(S'(X), Y)$$

X, S continuous: ❌

How do we find a minimal (achievable) sufficient statistic?

X discrete: ✔

$$f \in \arg\min_{S \in \mathcal{F}} \; I(X, S(X))$$

$$s.t. \;\; I(S(X), Y) = \max_{S' \in \mathcal{F}} I(S'(X), Y)$$

X, S continuous: ❌

I(X, S(X)) = ∞

This is a generally problematic issue in machine learning [3, 4].

[3] Bell and Sejnowski "An information- maximization approach to blind separation and blind deconvolution" 1995
[4] Goldfeld et al. "Estimating Information Flow in Neural Networks" 2018

Solution: Conserved Differential Information

$$C(X, f(X)) := H(f(X)) - \mathbb{E}_X \left[\log\left(J_f(X)\right)\right]$$

Differential entropy of the output...

...but discount any scaling

where

$$J_f(x) = \sqrt{\det\left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}}\left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}}\right)^{\mathrm{T}}\right)}$$

Solution: Conserved Differential Information

$$C(X, f(X)) := H(f(X)) - \mathbb{E}_X \left[\log \left(J_f(X)\right)\right]$$

See paper for connections to change of variables formula/Normalizing Flows/NVP

...but discount any scaling

where

$$J_f(x) = \sqrt{\det \left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}} \left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}} \right)^{\mathrm{T}} \right)}$$

How do we find a minimal (achievable) sufficient statistic?

X discrete: ✓

$$f \in \arg\min_{S \in \mathcal{F}} \; I(X, S(X))$$

$$s.t. \;\; I(S(X), Y) = \max_{S' \in \mathcal{F}} I(S'(X), Y)$$

X, S continuous: ✓

$$f \in \arg\min_{S \in \mathcal{F}} \; C(X, S(X))$$

$$s.t. \;\; I(S(X), Y) = \max_{S'} I(S'(X), Y)$$

MASS Learning

Minimize:

$$\mathcal{L}_{MASS}(f) := H(Y|f(X)) + \beta H(f(X)) - \beta \mathbb{E}_X[\log J_f(X)]$$

Make the output predictable from the representation...

...while keeping the representation as low-entropy as possible...

...but scaling things doesn't count

Results

Same accuracy as standard training and VIB

CIFAR-10, ResNet20, 4 trials

Method	Training Set Size		
	2500	10,000	40,000
SoftmaxCE	$\mathbf{50.0 \pm 0.7}$	$\mathbf{67.5 \pm 0.8}$	$\mathbf{81.7 \pm 0.3}$
VIB, β=1e−3	$\mathbf{49.5 \pm 1.1}$	$\mathbf{66.9 \pm 1.0}$	81.0 ± 0.3
VIB, β=1e−4	49.4 ± 1.0	66.4 ± 0.5	81.2 ± 0.4
VIB, β=1e−5	$\mathbf{50.0 \pm 1.1}$	$\mathbf{67.9 \pm 0.8}$	80.9 ± 0.5
VIB, β=0	$\mathbf{50.6 \pm 0.8}$	$\mathbf{67.1 \pm 1.0}$	$\mathbf{81.5 \pm 0.2}$
MASS, β=1e−3	38.2 ± 0.7	59.6 ± 0.8	75.8 ± 0.5
MASS, β=1e−4	$\mathbf{49.9 \pm 1.0}$	66.6 ± 0.4	80.6 ± 0.5
MASS, β=1e−5	$\mathbf{50.1 \pm 0.5}$	$\mathbf{67.4 \pm 1.0}$	$\mathbf{81.6 \pm 0.4}$
MASS, β=0	$\mathbf{50.2 \pm 1.0}$	$\mathbf{67.4 \pm 0.3}$	$\mathbf{81.5 \pm 0.2}$

Improved Uncertainty Quantification

CIFAR-10, ResNet20, 4 trials

Method	Test Accuracy	Entropy	NLL	Brier Score
SoftmaxCE	81.7 ± 0.3	0.087 ± 0.002	1.45 ± 0.04	$\mathbf{0.0324 \pm 0.0005}$
VIB, β=1e−3	81.0 ± 0.3	0.089 ± 0.003	1.51 ± 0.04	0.0334 ± 0.0005
VIB, β=1e−4	81.2 ± 0.4	0.092 ± 0.002	1.46 ± 0.05	0.0331 ± 0.0007
VIB, β=1e−5	80.9 ± 0.5	0.087 ± 0.005	1.58 ± 0.08	0.0339 ± 0.0008
VIB, β=0	81.5 ± 0.2	0.079 ± 0.001	1.70 ± 0.06	0.0331 ± 0.0007
MASS, β=1e−3	75.8 ± 0.5	0.139 ± 0.003	1.66 ± 0.07	0.0417 ± 0.0011
MASS, β=1e−4	80.6 ± 0.5	0.109 ± 0.002	$\mathbf{1.33 \pm 0.02}$	0.0337 ± 0.0008
MASS, β=1e−5	81.6 ± 0.4	0.095 ± 0.003	$\mathbf{1.36 \pm 0.03}$	$\mathbf{0.0320 \pm 0.0005}$
MASS, β=0	81.5 ± 0.2	0.092 ± 0.000	1.43 ± 0.04	0.0325 ± 0.0004

Caveat: current implementation expensive, but ample room for improvement

More results at our poster.
Code available at github.com/mwcvitkovic/MASS-Learning.