REVE: REGULARIZING DEEP LEARNING WITH VARIATIONAL ENTROPY BOUND

Tuesday 24th September, 2019

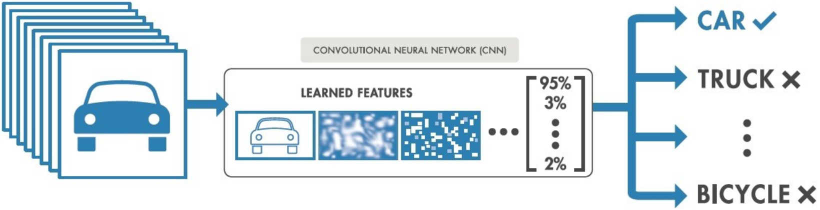
Antoine Saporta, Yifu Chen, Michael Blot, Matthieu Cord

SCIENCES SORBONNE UNIVERSITÉ

LIP6

MLIA Machine Learning & Deep Learning for Information Access

Image Classification



Source: MathWorks (https://goo.gl/zondfq)

Notations

- → X is the input image from $\mathcal{X}$
- → C is the class label from $\mathcal{C}$
- → Y is an intermediate representation of X from which is determined the predicted class $\hat{C}$

Regularization?

Problem

- → Huge number of parameters compared to the number of training samples
- → Deep networks prone to overfitting
- → Regularization: way to mitigate this gap and improve generalization

Common strategies

- → Weight decay
- → Dropout
- → Batch normalization

Information-based Regularization Criteria

Information Bottleneck (Tishby et al., 1999)

→ Principle: minimize $I(Y, X)$ at optimal $I(Y, C)$

SHADE (Blot et al., 2018)

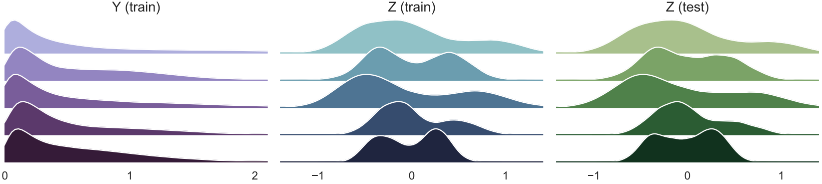
→ Principle: minimize $H(Y \mid C)$

REVE Contribution

→ Identify a *new* variable Z better suited for regularization

→ Develop a variational bound over the criterion $H(Z \mid C)$

REVE Variable



Definition

- → Linear decoder: $\boldsymbol{W}_d Y$
- → Unique decomposition: $Y = Z + Y^{\mathrm{ker}}$ where $Y^{\mathrm{ker}} \in \ker \boldsymbol{W}_d$ and $Z \in (\ker \boldsymbol{W}_d)^{\perp}$. Thus, $\boldsymbol{W}_d Y = \boldsymbol{W}_d Z$
- → REVE Variable: Z, the part of Y effectively used for prediction

REVE Criterion

→ The conditional entropy can be written:

$$H(Z \mid C) = H(Z) + H(C \mid Z) - H(C)$$

with $H(C)$ entirely determined by the problem

→ Objective Function:

$$\mathcal{L}_{REVE} = H(Z) + H(C \mid Z)$$

→ For any $q(Z)$ and $r(C \mid Z)$ variational approximations of $p(Z)$ and $p(C \mid Z)$, resp.:

$$\mathcal{L}_{\text{REVE}} \leq -\int_{\mathcal{Z}} p(\mathbf{z}) \log q(\mathbf{z}) \mathrm{d}\mathbf{z} - \iint_{\mathcal{Z}\,\mathcal{C}} p(\mathbf{z}, \mathbf{c}) \log r(\mathbf{c} \mid \mathbf{z}) \mathrm{d}\mathbf{z} \mathrm{d}\mathbf{c}$$

REVE Instantiation

Stochastic Encoding

A stochastic encoding is needed for computing entropies:

$$Y = h(\boldsymbol{W}_e, X) + \boldsymbol{\varepsilon} \text{ with } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Computing Z

- → Compact Singular Value Decomposition of $\boldsymbol{W}_d$: $\boldsymbol{W}_d = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top$ where the r column vectors of $\boldsymbol{U}$ and $\boldsymbol{V}$ correspond to the non-zero singular values of $\boldsymbol{W}_d$
- → Computation of Z:

$$Z = \boldsymbol{V} \boldsymbol{V}^\top Y$$

REVE Loss Function

Using Bayes' Theorem and DNN Markov-Chain hypothesis $C \leftrightarrow X \leftrightarrow Z$:

$$\mathcal{L}_{\text{REVE}} \leq -\iiint_{\mathcal{X}\,\mathcal{Z}\,\mathcal{C}} p(\boldsymbol{x})p(\boldsymbol{c} \mid \boldsymbol{x})p(\boldsymbol{z} \mid \boldsymbol{x})\big(\log q(\boldsymbol{z}) + \log r(\boldsymbol{c} \mid \boldsymbol{z})\big)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{c}$$

Thus, applying Monte-Carlo methods using the empirical distribution of (X, C) and the sampling of Y resp., we obtain the upperbound to minimize:

$$\Omega_{\text{REVE}}\big((\boldsymbol{x}_n, \boldsymbol{c}_n); \boldsymbol{W}_e; \boldsymbol{W}_d\big) = -\frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S}\left(\log q(\boldsymbol{z}_{n,s}) + \log r(\boldsymbol{c}_n \mid \boldsymbol{z}_{n,s})\right)$$

Approximation $r(\boldsymbol{c}|\boldsymbol{z})$

$$\Omega_{\text{REVE}} = -\frac{1}{NS} \sum_{n=1}^{N} \sum_{s=1}^{S} \left(\log q(\boldsymbol{z}_{n,s}) + \log r(\boldsymbol{c}_n \mid \boldsymbol{z}_{n,s}) \right)$$

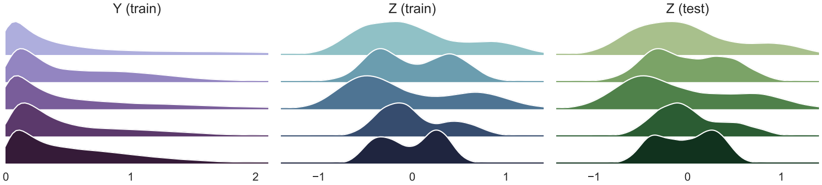
Variational approximation $r(\boldsymbol{c} \mid \boldsymbol{y})$: we use the learned classifier:

$$r(\boldsymbol{c} \mid \boldsymbol{z}) = \mathcal{S}(\boldsymbol{W}_d \boldsymbol{z} + \boldsymbol{b})_{\boldsymbol{c}}$$

Approximation $q(\boldsymbol{z})$

$$\Omega_{\mathrm{REVE}} = -\frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S}\left(\log q(\boldsymbol{z}_{n,s}) + \log r(\boldsymbol{c}_n \mid \boldsymbol{z}_{n,s})\right)$$

Variational approximation $q(\boldsymbol{z})$: how to model Z?



Bimodal Approximation

Model

- Independence between coordinates:

$$q(\boldsymbol{z}) = \prod_{i=1}^{\dim(\mathcal{Z})} q(z_i)$$

- Bimodal approximation:

$$q(z_i) = \alpha_i \mathcal{N}(z_i \mid \mu_{1,i}, \sigma_{1,i}^2) + (1 - \alpha_i)\mathcal{N}(z_i \mid \mu_{0,i}, \sigma_{0,i}^2).$$

- α_i: probability of the semantic attribute being present.
- $\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2, \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0^2$ to determine.

Compute the Parameters

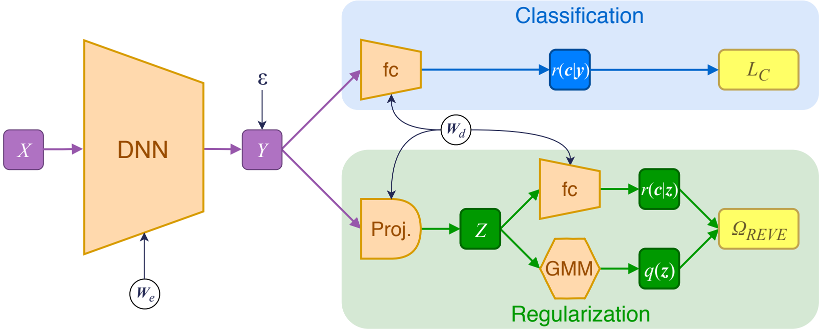
Expectation Maximization for two Gaussian Mixture Model?

- → Expensive
- → The size of the mini-batches is in general too small for obtaining a coherent model

Mini-batch Computation

- → We assume $P(M_i = 1 \mid z_i) = \sigma(z_i)$ (*e.g.* σ the sigmoid function)
- → On the mini-batch, $\alpha_i = \sum_n \sigma(z_i^{(n)})$, $\mu_{1,i} = \sum_n \frac{\sigma(z_i^{(n)})}{\alpha_i} z_i^{(n)}$,...

REVE Loss



$$\Omega_{\text{REVE}}\big((x_n, c_n); W_e; W_d\big) = -\frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S}\Bigg[\log \mathcal{S}(W_d\, z_{n,s} + b)_{c_n}$$

$$+\sum_{i=1}^{\dim(\mathcal{Z})} \log\Big(\alpha_i \mathcal{N}(z_{n,s,i} \mid \mu_{1,i}, \sigma_{1,i}^2) + (1-\alpha_i)\mathcal{N}(z_{n,s,i} \mid \mu_{0,i}, \sigma_{0,i}^2)\Big)\Bigg].$$

Performance Analysis

Reve Performance Analysis. Classification error (%) results on CIFAR-10 and CIFAR-100 test sets.

	CIFAR-10		CIFAR-100	
	AlexNet	Inception	AlexNet	Inception
Baseline	15.62	6.10	48.29	27.36
SGM Reve	14.24	6.17	-	-
KDE Reve	13.86	6.04	-	-
Reve	13.92	5.92	48.07	26.94
Reve + DO	**12.54**	**5.78**	**41.13**	**26.02**

Results

Classification error (%) results on CIFAR test sets.

	CIFAR-10			CIFAR-100		
	AlexNet	Inception Net	ResNet	AlexNet	Inception Net	ResNet
Baseline	15.62	6.10	4.08	48.29	27.36	20.70
Dropout	12.63	6.04	3.93	41.32	27.26	20.16
Information DO	14.97	6.04	NC	47.97	27.34	NC
Shade + DO	13.93	5.90	4.30	41.25	26.99	20.37
Reve + DO	**12.54**	**5.78**	**3.88**	**41.13**	**26.02**	**20.05**

Classification error (%) results on SVHN test set.

	SVHN		
	AlexNet	Inception Net	ResNet
Baseline	7.68	3.78	3.40
Reve	**6.55**	**3.29**	**3.11**

Conclusion

→ REVE is a tractable regularization loss for image classification

→ Identifies the part of the representation orthogonal to the kernel of the classifier as the variable to constrain

→ Penalizes the conditional entropy of the REVE variable given the class

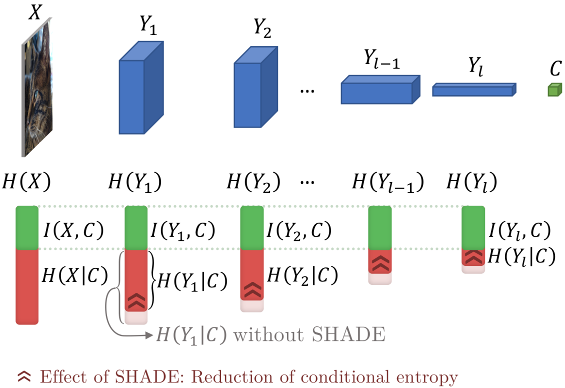
→ REVE shows consistent positive results on multiple architectures and datasets

Questions?

SCIENCES SORBONNE UNIVERSITÉ

Thank you for your attention!

SHADE



→ Layer-wise criterion: $\Omega_{\text{SHADE}} = \sum_{l=1}^{L} \sum_{i=1}^{D_l} H(Y_{l,i} \mid C)$

→ Uses a latent Bernoulli variable B as minimal sufficient statistic of C for Y: $I(Y, C) = I(Y, B)$