Learning to Find Correlated Features by Maximizing Information Flow in Convolutional Neural Networks

Wei Shen, Fei Li, Rujie Liu
Fujitsu R&D center

The First Workshop on Statistical Deep Learning in Computer Vision 2019
SDL-CV 2019

Motivation

- Information bottleneck (IB) trade-off tries to minimize the mutual information between the input data and the hidden representation
- Useful discriminative features may be lost when multiple of them are correlated (e.g. feature cooccurrence) in conventional model training process
- The **maximization** instead of minimization of mutual information between hidden representations may provide more information to the final classifier for learning discriminative features

Method

We propose an information flow maximization (IFM) loss as a regularization term to find the discriminative correlated features. With less information loss the classifier can make predictions based on more informative features.

Mutual information estimation

Mutual information is calculated as

$$I(X,Z) = \sum_{z\in Z}\sum_{x\in X} p(x,z)\log\frac{p(x,z)}{p(x)p(z)} = KL(p(x,z)||p(x)p(z)),$$

which can be approximated by maximizing

$$F(\omega) \approx \mathbb{E}_P[\sigma(V_\omega(x))] - \log\mathbb{E}_Q[1-\sigma(V_\omega(x))]$$

P is the joint distribution p(x, z) and Q is the product of marginal distribution p(x)p(z). σ is the sigmoid activation.

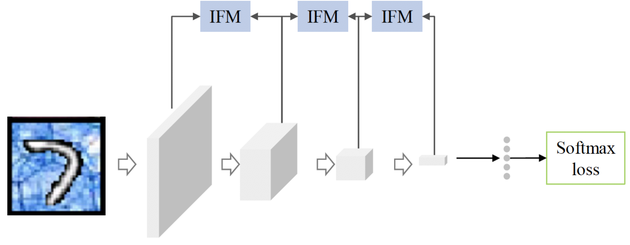

Fig. 1 The architecture of the proposed method.

Constructing sample pairs

- Upsampling (l+1)-th feature maps to the size of the l-th feature maps
- Sampling feature vectors

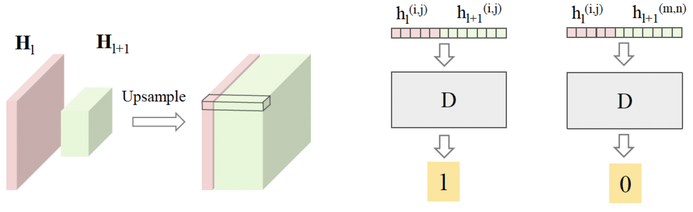

Fig. 2 Constructing sample pairs from the joint distribution and the product of marginal distributions.

Information flow maximization

$$L = L_{clf} - \sum_{l=1}^{M} F_l(\omega)$$

L_{clf} is the classification loss (e.g. the softmax loss) and M is the number of layers that used to calculate the information flow.

Experiments

Dataset -- shiftedMNIST

(a) Each digit is associated with a fixed type of texture

(b) Each digit is associated with a random type of texture

Fig.3. Some training examples (a) and test samples (b) from the shiftedMNIST dataset.

Task

- Train a 10-way classification model on images with fixed pairs of digit and texture type
- Test the learned model on images with random pairs of digit and texture type

Results

Model	acc (digit)	acc (texture)
Baseline$_{digit}$	12.44%	95.07%
Baseline$_{texture}$	12.05%	96.44%
iCE fi-RevNet	40.01%	-
Ours$_{digit}$	54.54 %	40.41%
Ours$_{texture}$	31.78 %	69.00%

- Baseline model is sensitive to texture features and ignores digit features for the 10-way classification task
- Information maximization flow helps model learn both digit and texture features for the classification task