SPATIAL ENSEMBLE KERNEL LEARNING FOR SCENE CLASSIFICATION

Zhang lei

Guangdong University of petrochemical technology

Guangdong, China

Scene classification

- **hand-crafted features:** follow the bag-of-word (BoW)/vector of locally aggregated descriptor (VLAD)/Fish vector structure + local descriptors
- **convolutional neural networks (CNNs)**
  - ✓ generic CNN features: FC7/FC8 from pretrained model as VGG
  - ✓ generate a set of high-quality patches potentially containing objects, and then apply a pre-trained CNN to extract generic deep features
  - ✓ generate a set of high-quality patches potentially containing objects, and then apply a pre-trained CNN to extract generic deep features
  - ✓ deep discriminative and shareable feature learning (DDSFL)----hierarchically learn feature transformation filter banks
  - ✓ factor analyzers and fisher vector (MFA-FV)----a MFA-FV Layer on CNN to build MFAFVNet
- **new model** on Place dataset-similar structure as CNN

Shortcoming in Scene classification

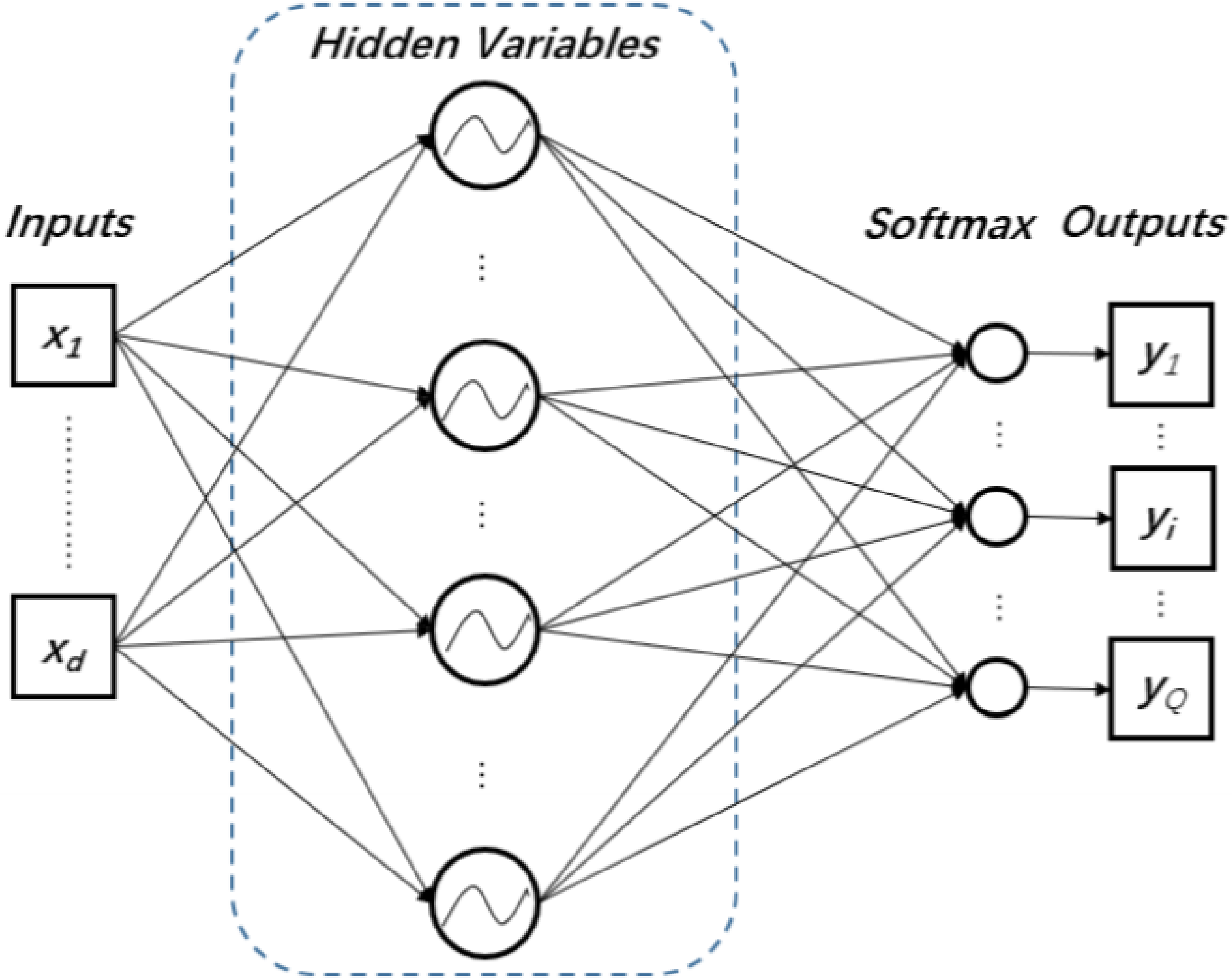
Missing of spatial layout information

Traditional CNNs pay close attention to holistic structure while still lacking spatial information. spatial layout carries the crucial cue for discriminative representation, especially for scene classification task.

Weaknesses of fusion ability

Scene information steps from diverse aspects, which is different from object classification

CACN--
Cosine
Activation in
Compact
Network



Fourier Feature Embedding

- **Kernel approximation:**
  - ✓In kernel approaches, in most cases, it is no need to explicitly define the mapping function
  - ✓With the increasing of dataset scale and considering of the calculating complexity, it is desired explicitly mapping the data to a low-dimensional Euclidean inner product space using a randomized feature map

$$\kappa(x_1, x_2) = < \phi(x_1), \phi(x_2) > \approx \Phi(x_1)^T \Phi(x_2)$$

Cosine Activation

Cosine activation

Theorem 1 *(**Bochner** [16]) A continuous function $g : \mathbb{R}^d \rightarrow \mathbb{C}$ is positive definite on $\mathbb{R}^d$ if only if it is the Fourier transformation of a finite non-negative Borel measurement $\mu(\boldsymbol{\omega})$ on $\mathbb{R}^d$, i.e.,*

$$g(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-j\boldsymbol{\omega}^\top \mathbf{x}} d\mu(\boldsymbol{\omega}), \quad \forall x \in \mathbb{R}^d \tag{3}$$

where j denotes the imaginary unit.

$$\begin{aligned} \kappa(x_1, x_2) = k(x_1 - x_2) &= \int_{\mathbb{R}^d} e^{j\omega^T(x_1 - x_2)} d\mu(\omega) \\ &= \int_{\mathbb{R}^d} \xi_\omega(x_1)\overline{\xi_\omega(x_2)} d\mu(\omega) \end{aligned} \tag{4}$$

$$\boxed{z_{\omega,b}(x) = \sqrt{2}\cos(\omega^T x + b)}$$

$$\kappa(x_1, x_2) \approx \xi_\omega(x_1)\overline{\xi_\omega(x_2)} = \Phi(x_1)^T \Phi(x_2)$$

$$\boxed{\Phi(x) = \sqrt{(2/M)}(z_{\omega_1,b_1}(x), \cdots, z_{\omega_M,b_M}(x))}$$

Optimization of CACN

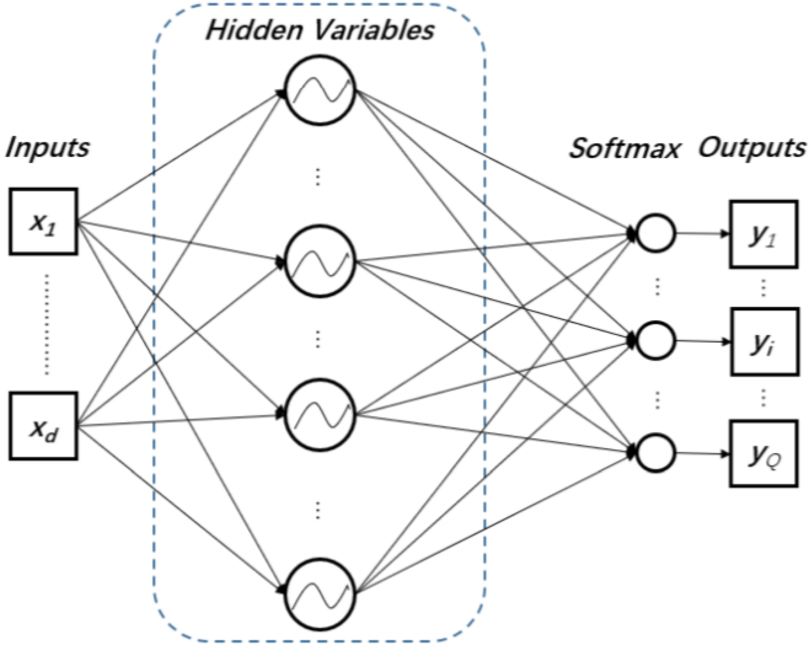
- general formula

$$\mathbf{y}_i = f(S, g(W, \mathbf{x}_i))$$

- predicted output

$$\hat{\mathbf{y}}_i = f(S, g(W, \mathbf{x}_i)) = \frac{\exp(S \times \mathbf{z}_i)}{\mathbf{1}_C^T \times exp(S \times \mathbf{z}_i)}$$

$$\mathbf{z}_i = g(W, \mathbf{x}_i)) = \sqrt{(2/M)}\cos(W\mathbf{x}_i + \mathbf{b})$$



Optimization of CACN

- Objective function

$$J = L(\mathbf{y}_i, f(S, g(W, \mathbf{x}_i))) + \lambda \Omega_g(W) + \beta \Omega_f(S)$$

- Cross entropy loss function

$$L = \frac{1}{N} \sum_i -\mathbf{y}_i^T \log(f(S, g(W, \mathbf{x}_i)))$$

$$\Omega_g(W) = ||W^T||_{2,1}$$

$$\Omega_f(S) = ||S||_* + ||S||_F$$

$$\frac{\partial J_1}{\partial S} = \sum_i tr((\frac{\partial \hat{\mathbf{y}}_i}{\partial (S \times \mathbf{z}_i)} \frac{\partial J}{\partial \hat{\mathbf{y}}_i})^T \frac{\partial (S \times \mathbf{z}_i)}{\partial S})$$
$$= \sum_i tr(((diag(\hat{\mathbf{y}}_i) - \hat{\mathbf{y}}_i \hat{\mathbf{y}}_i^T) diag(\hat{\mathbf{y}}_i)^{-1}(-\mathbf{y}_i))^T (\mathbf{z}_i)^T)$$
$$= \sum_i \mathbf{z}_i (\mathbf{y}_i - \mathbf{1}_Q^T (\mathbf{y}_i \odot \hat{\mathbf{y}}_i))^T$$
$$= \sum_i \sqrt{(2/M)} \cos(W\mathbf{x}_i)(\mathbf{y}_i - \mathbf{1}_Q^T (\mathbf{y}_i \odot \hat{\mathbf{y}}_i))^T$$

Optimization of CACN

$$\frac{\partial J_3}{\partial S} = \beta(\hat{U}\hat{V}^T + 2S)$$

Stochastic gradient descent (SGD)—**updating S**

$$S^{t+1} = S^t - \eta_s(\frac{\partial J_1}{\partial S} + \frac{\partial J_3}{\partial S})$$

$$\underbrace{J = L(\mathbf{y}_i, f(S, g(W, \mathbf{x}_i)))}_{J_1} + \underbrace{\lambda \Omega_g(W)}_{J_2} + \underbrace{\beta \Omega_f(S)}_{J_3}$$

J_1 J_2 J_3

$$J = L(\mathbf{y}_i, f(S, g(W, \mathbf{x}_i))) + \lambda \Omega_g(W) + \beta \Omega_f(S)$$

Optimization of CACN

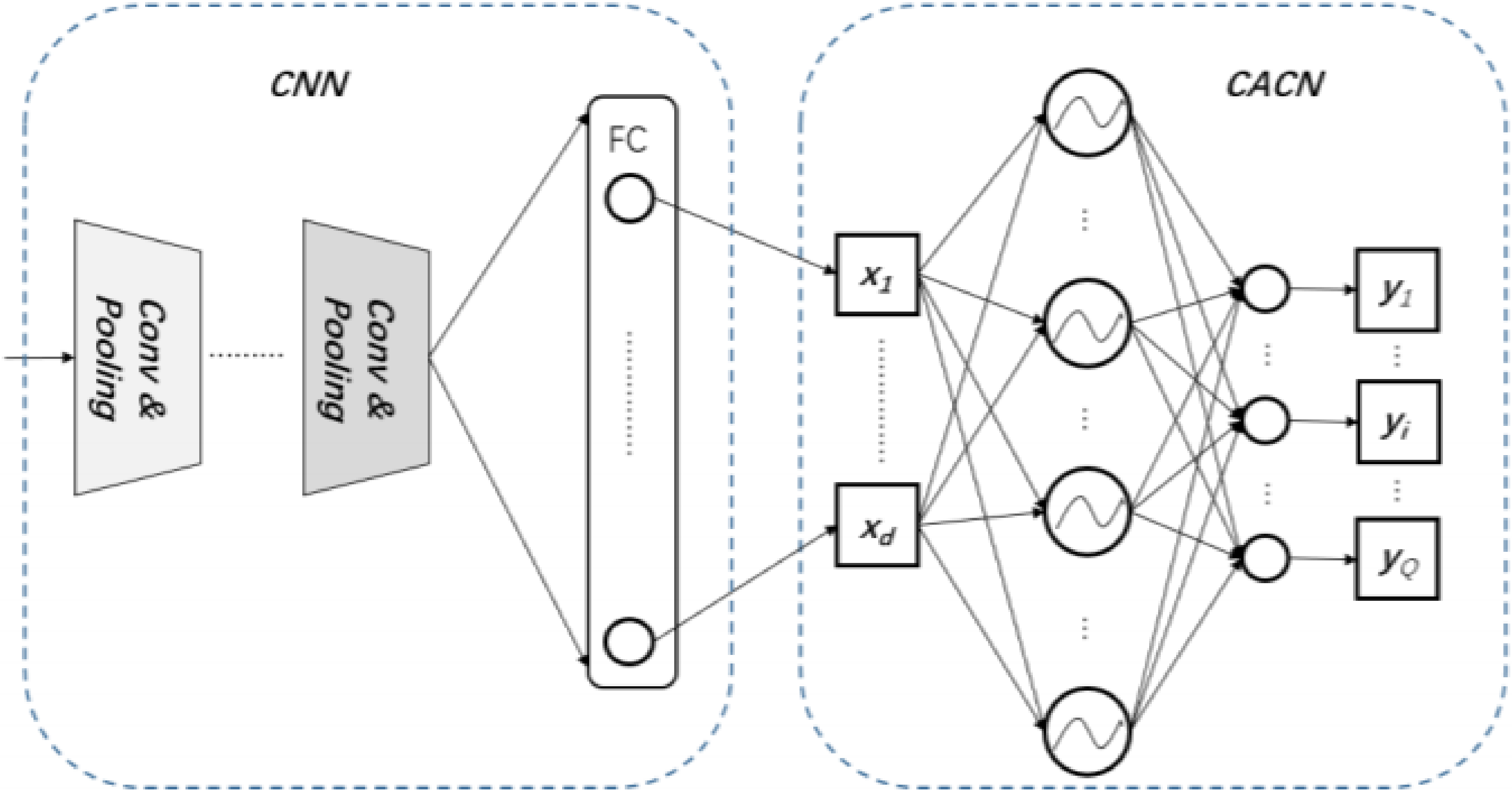
Stochastic gradient descent (SGD)—**updating W**

$$\begin{aligned}\frac{\partial J_1}{\partial W} &= \sum_i tr((\frac{\partial \hat{\mathbf{y}}_i}{\partial (S \times \mathbf{z}_i)} \frac{\partial J}{\partial \hat{\mathbf{y}}_i})^T \frac{\partial (S \times \mathbf{z}_i)}{\partial W}) \\ &= \sum_i (S^T(\mathbf{y}_i - \mathbf{1}_Q^T(\mathbf{y}_i \odot \hat{\mathbf{y}}_i))) \odot \sin(W\mathbf{x}_i) \times \mathbf{x}_i^T\end{aligned}$$

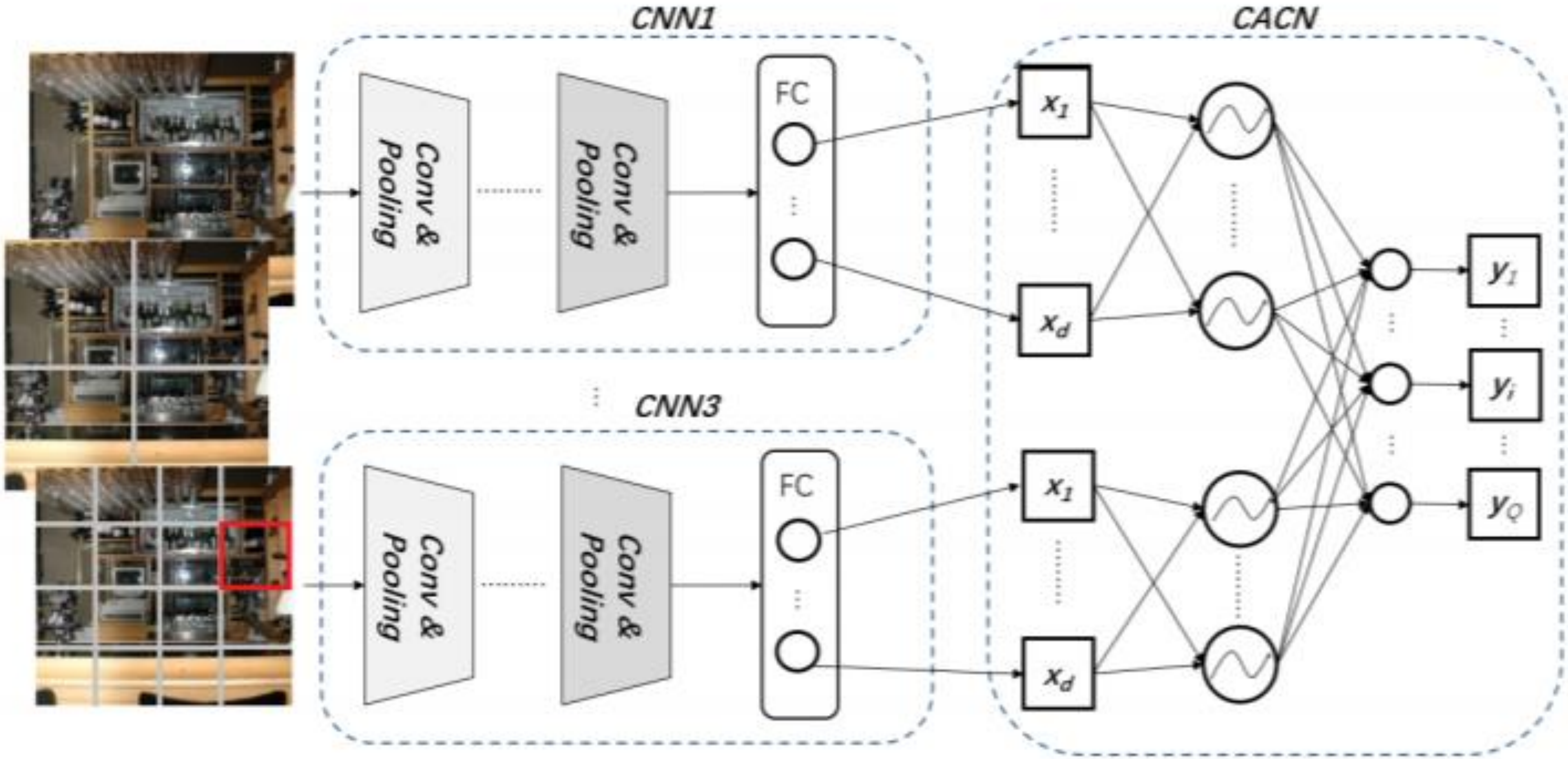
$$\frac{\partial J_2}{\partial W} = 2\lambda W \times diag(\frac{1}{2||\mathbf{w}_i||_2})$$

$$W^{t+1} = W^t - \eta_w(\frac{\partial J_1}{\partial W} + \frac{\partial J_2}{\partial W})$$

CACN+CNN



Spatial Ensemble Kernel Learning

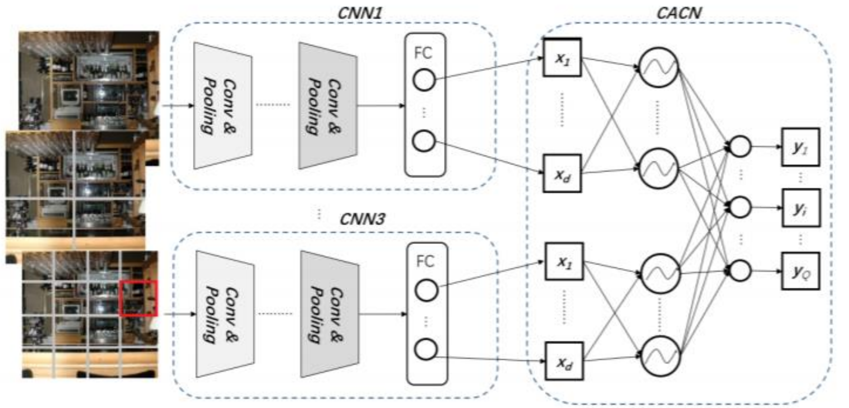


Spatial Ensemble Kernel Learning

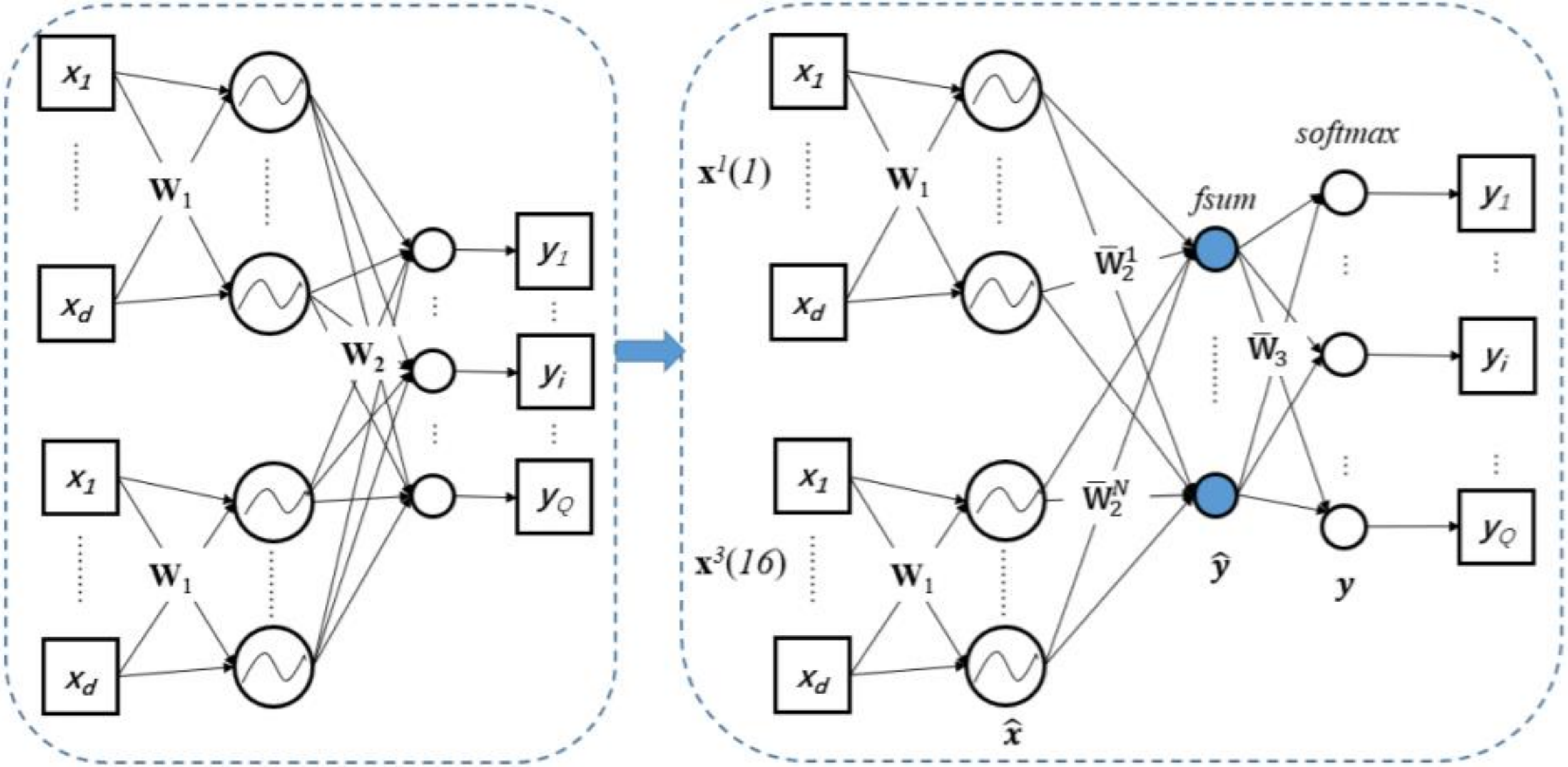
- Traditional spatial pyramid match kernel

$$\kappa(x_1, x_2) = \sum_{l=1}^{L} \sum_{i=1}^{I} \kappa(x_1^l(i), x_2^l(i))$$

- Structure expanding
  - Three different granularities are adopted in SPM and the whole image is divided into {1, 4, 16} grids separately. Each grid is fed into CNNs, which is VGG-16 model pre-learned by ImageNet.



Spatial Ensemble Kernel Learning-Deep analysis of combination



$$\tilde{x}_d^l(i) = \frac{(x_d^l(i) - \mu_{x^l(i)})}{\sigma_{x^l(i)}}$$

$$\hat{\mathbf{x}}^l(i) = \Phi(\widetilde{\mathbf{x}}^l(i)) = \frac{2}{\sqrt{D}} \cos(W_1^T \times \widetilde{\mathbf{x}}^l(i))$$

Parameter sharing

Spatial Ensemble Kernel Learning

-Deep analysis of combination

Equivalence proof

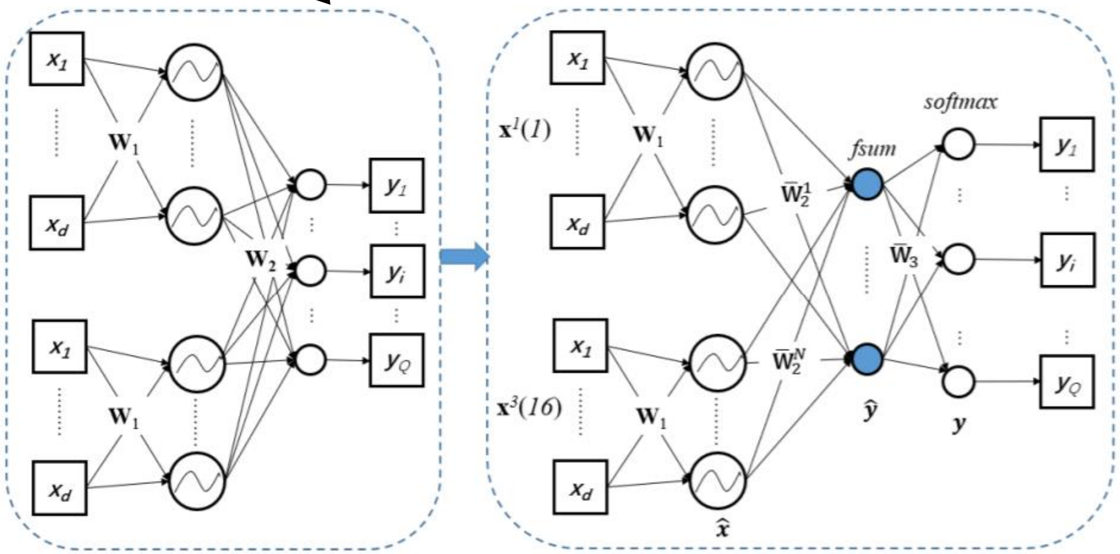
Spatial Ensemble Kernel Learning

-Deep analysis of combination

$$\mathbf{y} = \overline{W}_3^T \Big(\sum_{n=1}^{N} (\overline{W}_2^n)^T \hat{\mathbf{x}}^n\Big) + \bar{\mathbf{b}}_3 = \overline{W}_3^T \left(\begin{bmatrix} \overline{W}_2^1 \\ \overline{W}_2^2 \\ \vdots \\ \overline{W}_2^N \end{bmatrix}^T \hat{\mathbf{x}} + \bar{\mathbf{b}}_2 \right) + \bar{\mathbf{b}}_3$$

$$= \begin{bmatrix} \overline{W}_2^1 \overline{W}_3 \\ \overline{W}_2^2 \overline{W}_3 \\ \vdots \\ \overline{W}_2^N \overline{W}_3 \end{bmatrix}^T \hat{\mathbf{x}} + \overline{W}_3^T \bar{\mathbf{b}}_2 + \bar{\mathbf{b}}_3$$

$$\mathbf{y} = W_2^T \hat{\mathbf{x}} + \mathbf{b}_2$$



$$\kappa(\mathbf{x}_1, \mathbf{x}_2) = \sum_{l=1}^{L} \sum_{i=1}^{I} \mathbf{x}_1^l(i)^T \mathbf{x}_2^l(i) = \mathbf{x}_1^T \mathbf{x}_2$$

- ✓ By Φ function, it is easy to understand the combination of different level and different grid information from kernel approximation aspect

- ✓ WW^T term can be viewed as the weight of different level and different grid, which is learned by supervised way.

$$\kappa(\mathbf{y}_1, \mathbf{y}_2) = \Phi(\mathbf{x}_1)^T W W^T \Phi(\mathbf{x}_2)$$

Kernel aspect explanation

Spatial Ensemble Kernel Learning-Deep analysis of combination

Experiments and Results

Table 1 Performance on MIT indoor and SUN 397

Dataset	Method	Accuracy (%)
MIT indoor	$fc8$(VGG)+SVM	59.50
	CACN+CNN	71.89
	SEK	**75.73**
SUN 397	$fc8$(VGG)+SVM	47.15
	CACN+CNN	52.17
	SEK	**56.58**

MIT indoor: The whole number of categories is 67. The database contains 15,620 images and all images have a minimum resolution of 200 pixels in the smallest axis.

SUN 397: SUN (Scene UNderstanding) 397 dataset contains approximate 100,000 images of 397 categories. Only color images of 200 × 200 pixels or larger were kept.

Experiments and Results

Table 2. Comparison on MIT indoor.

Method	Accuracy (%)
DeCaF [23]	59.50
MOP-CNN [11]	68.88
fc8-FV [10]	72.86
MFA-FS [24]	81.43
SEK	**75.73**

Table 3. Comparison on SUN 397 dataset.

Method	Accuracy (%)
Combined 12 feature types [21]	38.00
FV (SIFT) [25]	43.30
DeCaF [23]	43.76
FV (SIFT+LCS) [25]	47.20
MOP-CNN [11]	51.98
fc8-FV [10]	54.40
MFA-FS [24]	63.31
SEK	**56.58**

[23] Decaf: A deep convolutional activation feature for generic visual recognition. CVPR 2013
[11] Multi-scale orderless pooling of deep convolutional activation features. ECCV 2014
[10] Scene classification with semantic fisher vectors. CVPR 2015
[24] Object based scene representations using fisher scores of local subspace projections. NIPS 2016
[21] Sun database: Large-scale scene recognition from abbey to zoo. CVPR 2010.
[25] Image classification with the fisher vector: Theory and practice. IJCV 2013

Conclusion

- we have presented a cosine activation compact network (CACN) and two kinds of extension in scene classification.
- spatial ensemble kernel learning approach----when combined with SPM
- Advantage: To compensate the loss of spatial layout information and the weaknesses of fusion ability from diverse aspects in scene classification while maintain the advantages of deep learning

Thanks for your attention