Dict Layer:
A Structured Dictionary Layer

Yefei Chen, Jianbo Su

{marschen, jbsu}@sjtu.edu.cn

Shanghai Jiao Tong University, China

Content

- Goal
  - Dictionary Learing
  - Deep Learning
- Dict Layer
- Experiment
- Conclusion

Deep Learning

Adavantage

- Strong ability for **feature extraction**
- **End-to-end**

Disadvantage

- Meaning of neural units being significant is **unclear**

Research Field of Deep Learning

- Vanishing of gradient
- Over-fitting
- Training time

Road Map

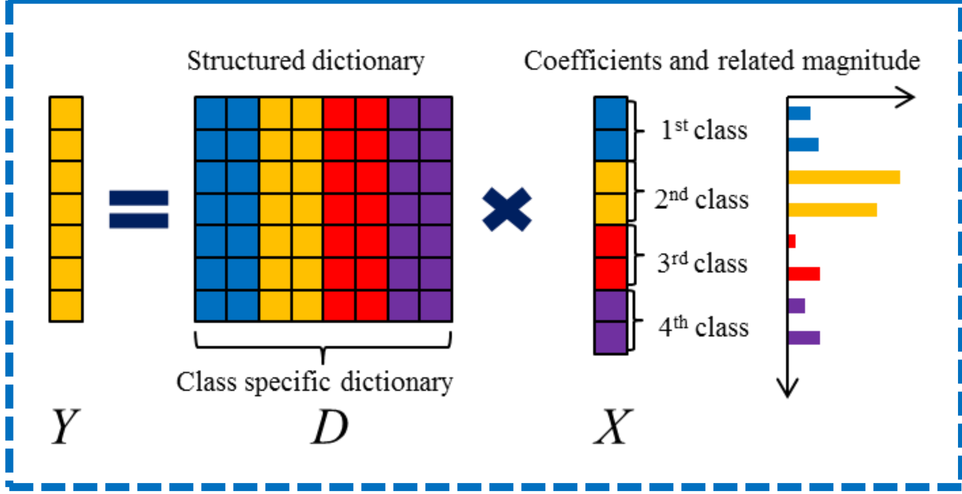
AlexNet 2013
DeepFace 2014
DeepID 2014-2016
VGGNet 2015
GoogleNet 2015
FaceNet 2015
ResNet 2016
LBCNN 2017

Dictionary Learning

Adavantage

- **Clear explanation** for representation coefficients being significant
- Coefficients can be **class specified** and helpful for classification

Dictionary Learning with Structured Dictionary



Disadvantage

- Rely on the **feature extracted**

Our goal is trying to inherit the property from Dictionary Learning and Deep Learning.

Dictionary Learning

Key Function of Dictionary Learning

$$\min_{X,D,\theta} \{r(Y,D,X,\theta) + \lambda_1 \|X\|_1 + \lambda_2 g(X) + \lambda_3 h(\theta)\}.$$

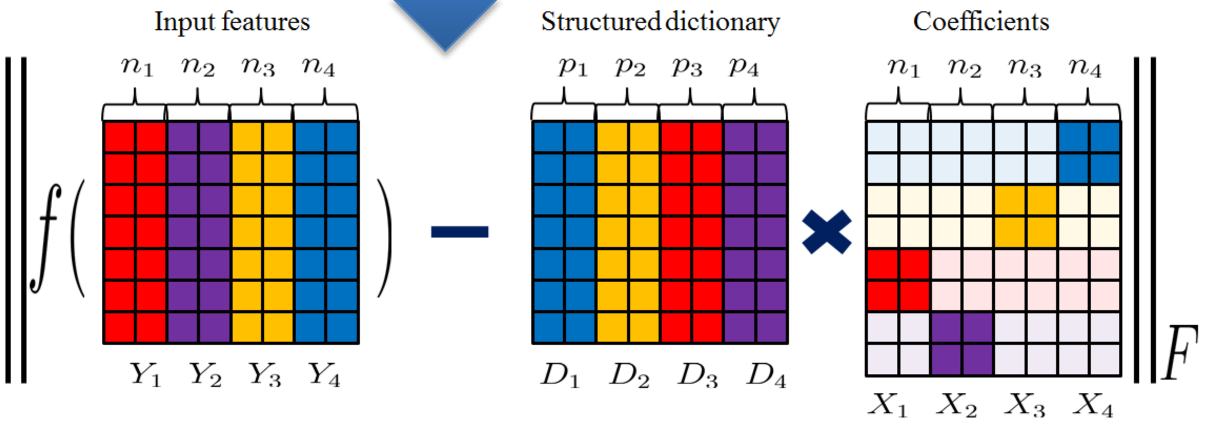
Reconstruction term

Regularization term

Loss of coefficients

Parameters of feature extraction

$$r(Y,D,X,\theta) = \|f_\theta(Y) - DX\|_F^2,$$



Reconstruction Term

Main Improvements in Dictionary Learning

- Design the loss of coefficients
- Structure of Dictionary
- Design more powerful feature extraction method

Dictionary Learning

	Road Map	Coefficients	Dictionary Structure	Feature Extraction
Michael Elad	**KSVD 2006** ,Unsupervised learning			
	Linear Classification + KSVD 2008, Supervised learning with linear classification on coefficients	√		
VGG, Oxford & INRIA	**SDL(Supervised Dictionary Learning) 2009 ,** Supervised learning with **Softmax**	√		
	DKSVD(Discriminant KSVD) 2010, First introduce the idea of structure dictionary		√	
	LC-KSVD(Label Consistent KSVD) 2011, Consistence of samples (labels)	√	√	
Jean Ponce, INRIA	**TDL(Task-driven Dictionary Learning) 2012,** Framework of Dictionary Learning	√	√	
Zhang Lei, HKPU	**MFL(Meta Face Learning) 2010 ,** Introduce structure dictionary into face recognition	√	√	
	FDDL(Fisher Discriminant Dictionary Learning) 2011, 2014, Fisher criteria	√	√	
	DGSR(Discriminative Dictionary for Group SR) 2014, Structure dictionary with group lasso	√	√	
	SDR(Sparse and Dense Hybrid Representation) 2015, Structure dictionary with low rank decomposition		√	
Rama Chelleppa, Maryland	**SE(Sparse Embedding) 2012 ,** Framework of linear feature extraction and dictionary learning		√	√
	DRSR(Dimensionality Reduction for SR) 2010, Feature extraction and dictionary learning simultaneously	√	√	√
	JDDRDL(Joint Discriminative Dimensionality Reduction and Dictionary Learning) 2013	√	√	√
Jiwen Lu, Tsinghua	**SFDL(Simultaneous Feature and Dictionary Learning) 2014,2017,** Extension to nonlinear feature extraction	√	√	√
	BDDL(Bilinear Discriminant Dictionary Learning) 2014	√	√	√
	Deep Dictionary Learning(DDL) 2016,2017, Dictionary can be stack into deep network like DBN			

Content

- Goal
- Dict Layer
  - Relationship between DLs
  - Dict Layer
- Experiment
- Conclusion

Relationship between DLs

- Both can be deep stacked
- Loss function g() can be identical
- Similar in mathematical

Deep Learning in ADMM Style

$$\min_{X^m, a^m, W_m} \{\sum_{c=1}^{C} g(X_c^M, L)\}$$
$$s.t. \quad X_c^m = W_m a_{,c}^{m-1}, \quad a_c^m = h_m(X_c^m), \ \forall m, \forall c$$

$$\min_{f, X^M} \{\sum_{c=1}^{C} g(X_c^M, L)\}$$
$$s.t. \quad X_c^M = f_M(Y_c), \ \forall c$$

Simplified function of Dictionary Learning

$$\min_{D, X, f} \{\|f(Y) - DX\|_F^2 + \lambda_1 \|X\|_F^2 + \lambda_2 g(X, L)\},$$

$$\min_{D, X_c, f} \{\sum_{c=1}^{C} (\lambda_1 \|X_c\|_F^2 + \lambda_2 g(X_c, L))\}$$
$$s.t. \quad f(Y_c) = DX_c, \ \forall c$$

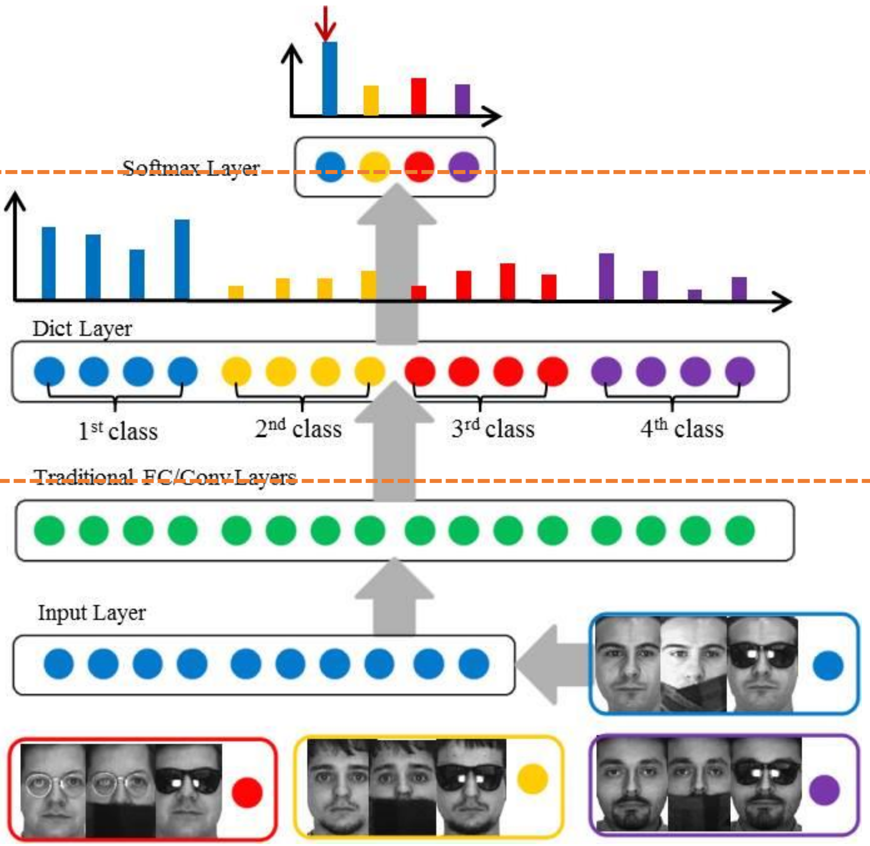
Our goal is trying to inherit the property from Dictionary Learning and Deep Learning.

Structured Dictionary in Deep Learning

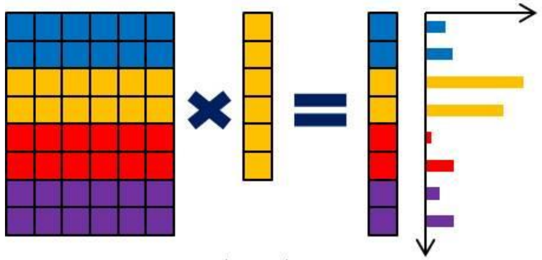
Loss function of Deep Learning with Dict Layer

$$\min_{f}\left\{\sum_{c=1}^{C}\left(g(f_M(Y_c),L)+\lambda\|W_{oc}f_m(Y_c)\|_F^2\right)\right\}.$$

Illustration of Dict Layer



Calculation in Dict Layer



$$W\cdot f(y_{c_i})=x_{c_i}$$

Derivation of Errors in Back Propagation

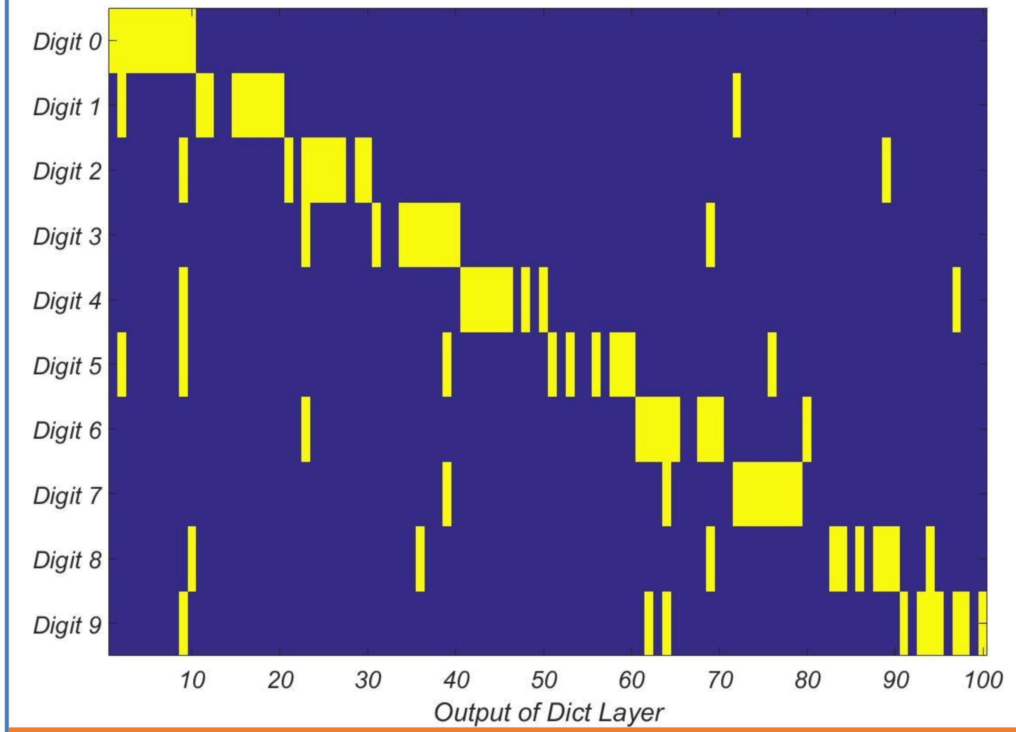
$$\begin{aligned}
\delta_\star^m &= \lambda W_{oc}f_m(Y_c)\circ h_m'(X_c^m),\\
\delta^m &= W_{m+1}^T\delta^{m+1}\circ h_m'(X_c^m)+\delta_\star^m,\\
\frac{\partial E}{\partial W_m} &= \frac{(f_{m-1}(Y_c))^T\delta^m}{n_c},
\end{aligned}$$

Content

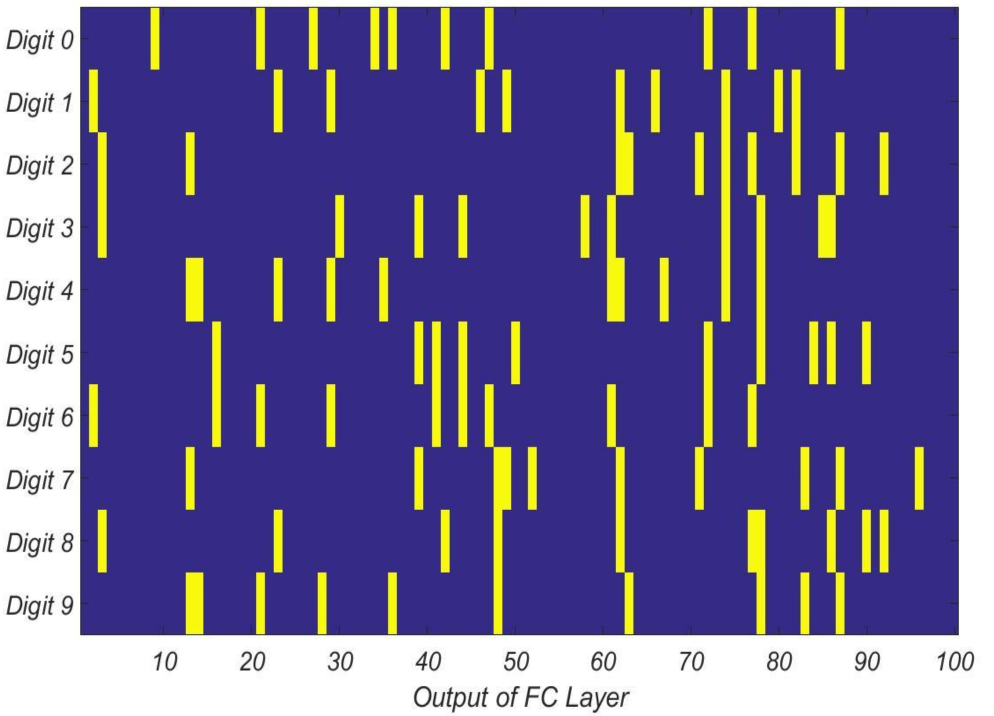
- Goal
- Dict Layer
- Experiment
  - Visualization result
  - Face recognition result
- Conclusion

Experimental Result

Hidden Neural Unit on MNIST



Neural Units on Dict Layer



Neural Units on FC Layer

Visualization of hidden neural units on Dict Layer and FC layer.
Distribution on Dict Layer is more equivalence and class specified.

Experimental Result

Comparison on Dict Layer and FC Layer with LBCNN(CVPR2017)

SVHN:

80 convolutional filters

SVHN		
	Dict Layer	FC Layer
300	**5.52%**	5.61%
400	**5.48%**	5.54%
500	5.59%	**5.55%**
600	**5.53%**	5.54%

CIFAR-10:

100 convolutional filters

Only 12500 images are used for training

CIFAR-10		
	Dict Layer	FC Layer
300	**16.65%**	17.07%
400	**16.28%**	16.92%
500	**16.21%**	16.83%
600	**15.96%**	16.87%

Experimental Result

Results on Face Recognition

AR

Dims	Dict Layer		FC Layer	
	Softmax	Cosine	Softmax	Cosine
500	**0.93±1.56%**	0.81±1.36%	5.97±1.55%	16.37±1.36%
600	**0.80±1.51%**	0.79±1.60%	5.67±1.03%	16.27±1.25%
<u>700</u>	**0.76±1.51%**	0.69±1.31%	5.63±1.23%	16.25±1.47%
800	**0.71±1.16%**	0.67±1.22%	5.35±0.73%	15.81±1.08%
900	**0.67±1.17%**	0.60±1.10%	5.17±0.90%	15.24±1.32%
1000	**0.69±1.21%**	0.59±1.06%	4.90±0.90%	15.29±1.35%

CMUPIE

Dims	Dict Layer		FC Layer	
	Softmax	Cosine	Softmax	Cosine
340	**3.61±0.18%**	3.56±0.19%	9.17±0.31%	11.04±0.39%
680	**3.62±0.18%**	3.49±0.16%	8.58±0.35%	11.21±0.49%
1020	**3.55±0.22%**	3.43±0.22%	8.24±0.37%	11.04±0.33%
<u>1360</u>	**3.67±0.18%**	3.49±0.20%	8.04±0.33%	11.20±0.42%
1700	**3.65±0.19%**	3.47±0.16%	7.86±0.39%	11.15±0.33%
2040	**3.62±0.17%**	3.45±0.20%	7.68±0.26%	11.24±0.35%

Ext.YALeB

Dims	Dict Layer		FC Layer	
	Softmax	Cosine	Softmax	Cosine
190	**1.98±0.40%**	1.75±0.41%	3.93±0.50%	4.96±0.25%
380	**1.90±0.44%**	1.75±0.57%	4.22±0.58%	5.72±0.79%
570	**2.04±0.29%**	1.87±0.38%	4.25±0.50%	6.23±0.77%
<u>760</u>	**2.13±0.36%**	1.91±0.43%	4.25±0.47%	6.68±0.61%
950	**2.21±0.36%**	1.91±0.35%	4.53±0.54%	7.10±1.16%
1140	**2.25±0.37%**	1.93±0.39%	4.70±0.56%	7.32±0.95%

Dict Layer is more powerful under cosine distance.

A good feature can make a simple classifier being powerful.

Experimental Result

Results on Face Recognition

AR

Method	Accuracy
FDDL [130]	92.2%
JDDRDL [141]	94.0%
BDDL [143]	93.6%
SFDL[145]	97.14%
FC Layer	94.37%
Dict Layer[1]	92.86%
Dict Layer[2]	**99.24%**

Ext. Yale B

Method	Accuracy
FDDL [130]	94.4%
MFL [133]	91.3%
SFDL[145]	96.79%
FC Layer	95.75%
Dict Layer	**97.87%**

CMU PIE

Method	Accuracy
FDDL [130]	92.3%
BDDL [143]	93.3%
SFDL[145]	94.1%
FC Layer	91.96%
Dict Layer	**96.33%**

Dictionary Learning is more useful on database with small category about 100.
Nonlinear extraction is more powerful than linear extraction.
Structure information is useful for Deep Learning.

Conclusion

Adavantage

- Each neural is class specified and designed for classification.
- Dict Layer can be viewed as an enhancement of FC Layer and can be used on any networks with FC Layer.
- A exemplar of introducing traditional method into Deep Learning.

Disadvantage

- Expansion of dimensionality with the growth of categories.
- Designed for situation where training categories and testing categories are identical.
- One more parameter is needed.

Thanks!

Dict Layer:
A Structured Dictionary Layer

Yefei Chen, Jianbo Su

{marschen, jbsu}@sjtu.edu.cn

Shanghai Jiao Tong University, China