Group Sparse CNNs for Question Classification with Answer Sets

Mingbo Ma and Liang Huang (Oregon State Univ.)
Bowen Zhou and Bing Xiang (IBM T. J. Watson)

Typical Sentence Classification

✓ Sentimental Classification: (movie review)

i) Review: "I don't like this movie at all! ..." (Negative)

ii) Two categories (Positive or Negative)

iii) No overlaps between categories and single label

✓ Question Type Classification: (TREC datasets)

iv) Q: "What is Oregon's state flower?" (Entity)

v) Six categories (Entity, Location, Number ....)

vi) Still no overlaps between categories and single label

However, there are overlapping categories in questions classification ...

Question Classification

Hey, Siri, what would be the best thing to do in New York City?

Attraction	Ans. 1: Go to a museum?
Sports	Ans. 2: Watch a Yankees game?
Dining	Ans. 3: Exploring various restaurant?
...	...

There are multiple answers which come from different categories!

Examples: NY-DMV FAQs

✓ New York State DMV FAQs:

i) 8 Top level categories and 47 sub-categories

ii) 537 questions (only 388 unique sentences)

✓ FAQs Examples:

iii) Driver License/Permit/Non-Driver ID

a. Apply for original (49 questions)
b. Renew or replace (24 questions)
c. Where is my photo document? (15 questions)

...

iv) Vehicle Registrations and Insurance

a. Buy, sell, or transfer a vehicle (22 questions)
b. Reg. and title requirements (42 questions)

...

v) Driving Record / Tickets / Points

Motivations

✓ Question classification different from general sentence modeling:

i) Question categories have hierarchical and overlapping structures

- Each question often belongs to multiple categories (multi-labeled)
- Question categories often have hierarchical structures
- Question categories often have overlaps

Motivations

✓ Question classification different from general sentence modeling:

i) Questions or question categories have well-prepared answer sets

- These answer sets generally cover a larger vocabulary (than the questions themselves) and provide richer information for each class
- We believe there is a great potential to enhance question representation with extra information from corresponding answer sets

Group Sparse CNNs

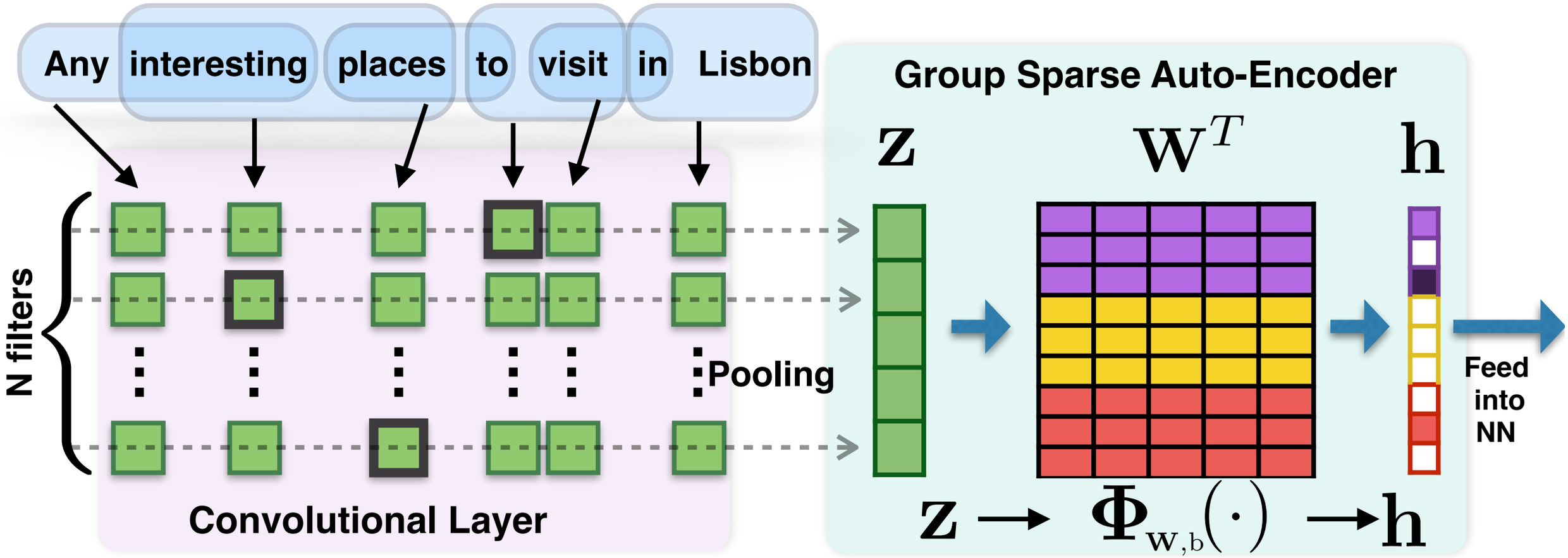
✓ Why do we need group sparse?

i) Explore the shared information

- within categories (sparse constraint)
- between categories (group sparse constraint)

ii) uses information from answers as dictionary to build more informative sentence representation

Group Sparse CNNs



- ✓ W is the projection matrix (functions as a dictionary)
- ✓ Darker colors in h mean larger values and white means zero
- ✓ h is the sparse representation of z, we apply different inter- and intra- sparse constraints on h.

Group Sparse CNNs

Our proposed Group Sparse Constrains:

$$J_{\text{groupsparse}}(\rho, \eta) = J + \alpha \sum_{j=1}^{s} KL(\rho \| \hat{\rho}_j) + \beta \sum_{p=1}^{G} KL(\eta \| \hat{\eta}_p)$$

Sparse **Group Sparse**

where

Sparse $KL(\rho \| \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}$ where $\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^{m} h_j^i$

Group Sparse $KL(\eta \| \hat{\eta}_p) = \eta \log \frac{\eta}{\hat{\eta}_p} + (1-\eta) \log \frac{1-\eta}{1-\hat{\eta}_p}$ where $\hat{\eta}_p = \frac{1}{mg} \sum_{i=1}^{m} \sum_{l=1}^{g} \| h_{p,l}^i \|_2$

Datasets

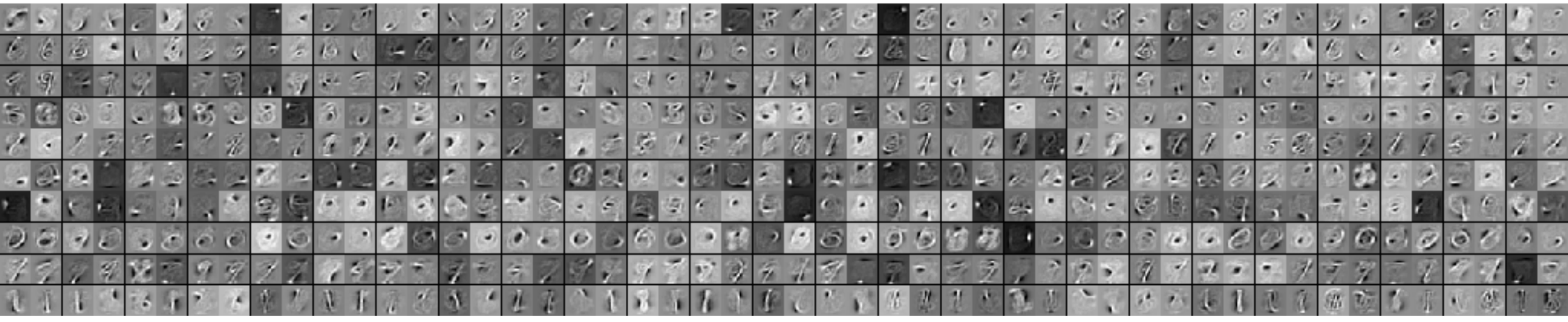
- ✓ Digits visualization
  - i) only used for visualize the group sparse AE's performance. to show the idea of group works.
- ✓ TREC dataset
  - i) single label, small data and w/o answer set
  - ii) this is well-know data, just give the reviewers the idea of our performance compared state of art performance.

Datasets

- ✓ NY-DMV dataset (self collected)
  - i) multi-label, small data and w/ answer data
  - ii) more realistic, public accessible dataset
- ✓ Yahoo dataset
  - i) Single-label, big data and w/ answer set
  - ii) The only problem is this is not multi-label problem
- ✓ Insurance dataset (private, IBM customer)
  - i) Multi-label, small data and w/ answer set
  - ii) this is ideal data set but data is too small

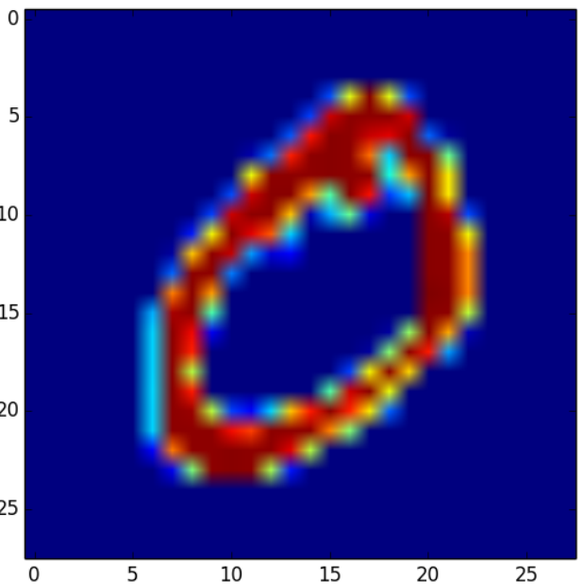
Sparsity Visualization

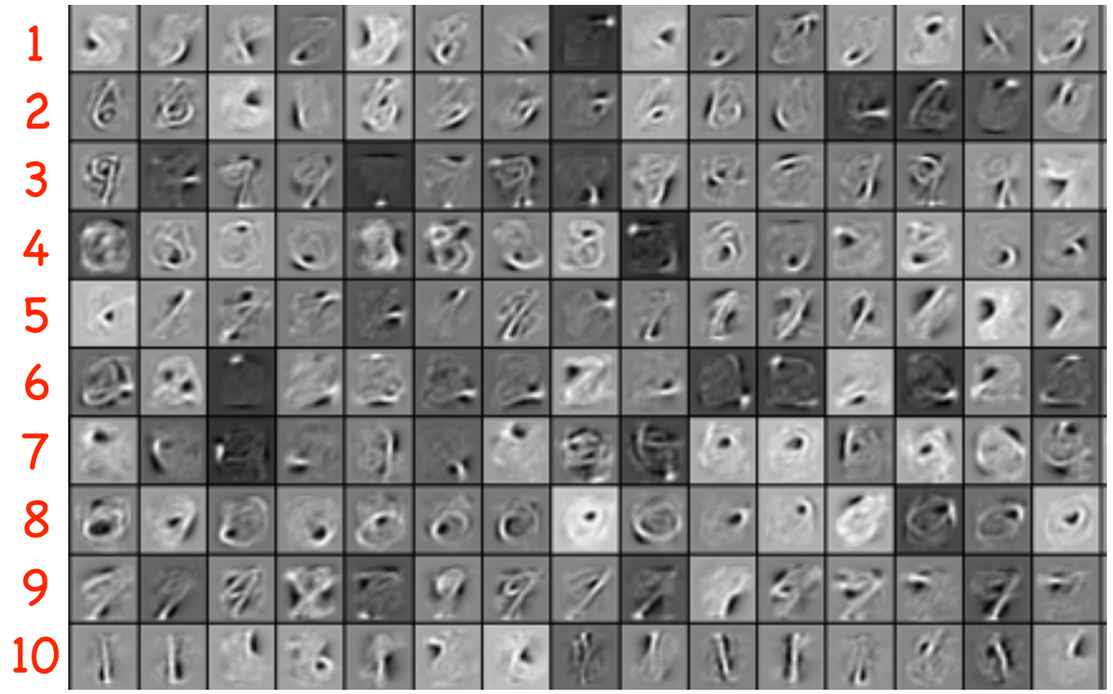
- ✓ Group sparse on MINIST dataset (handwritten digits) in order to get meaningful visualization.
- ✓ There are 10 groups, for each group there are 50 centroids
- ✓ The following are the visualization for the dictionary
  - ✓ 10 groups (row direction)
  - ✓ 50 centroids (column direction)
- ✓ This dictionary was initialize by clustering the dataset and trained by group sparse AE.

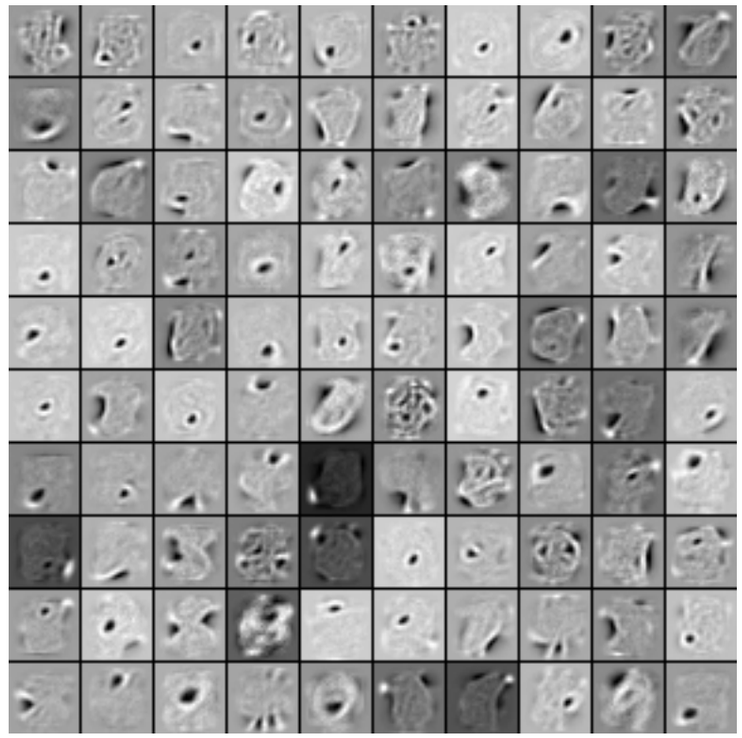


Sparsity Constraint: group sparsity

This is the visualization of W for hand written digit "0"



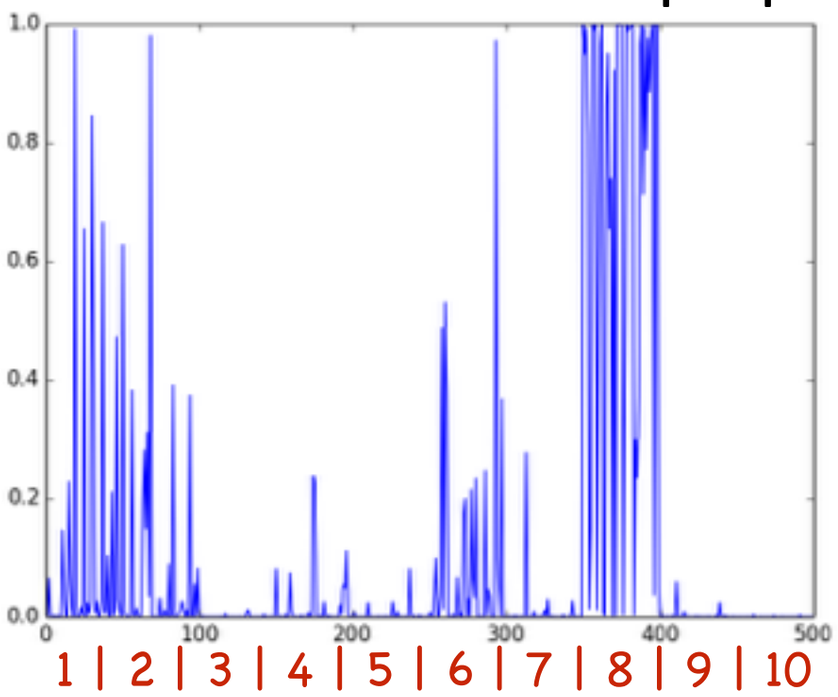


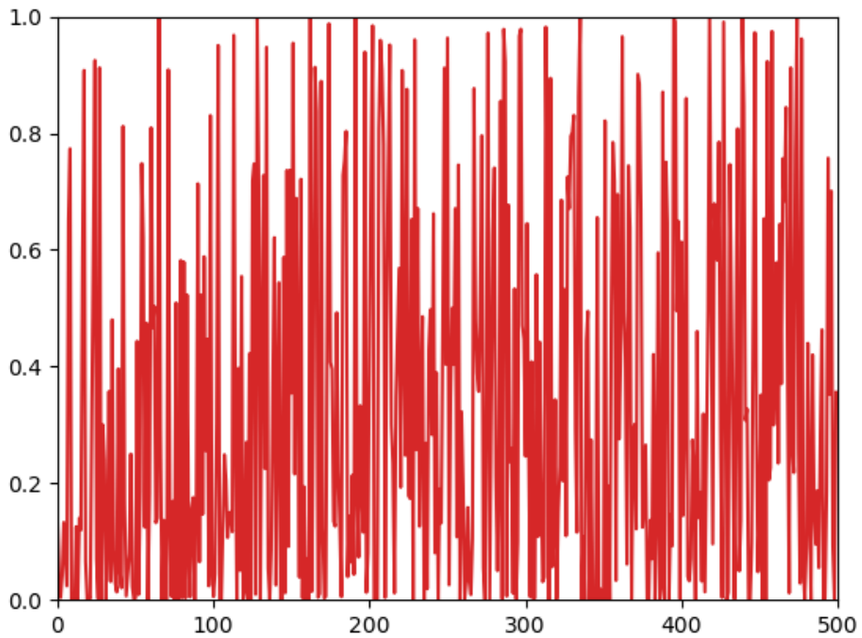


input image

Group Sparse Auto-Encoder

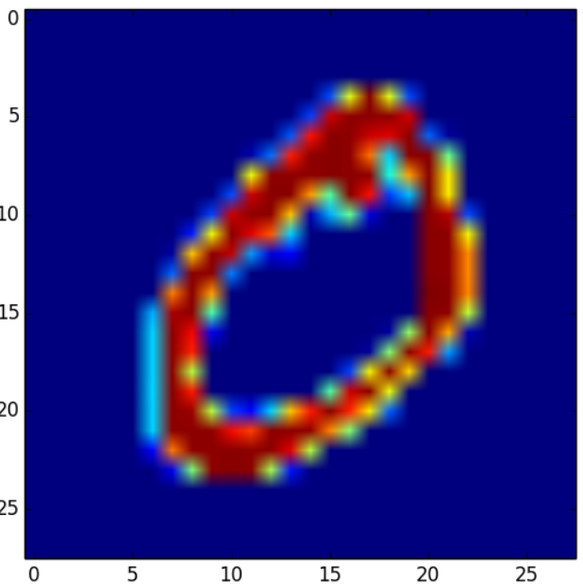
Conventional Auto-Encoder



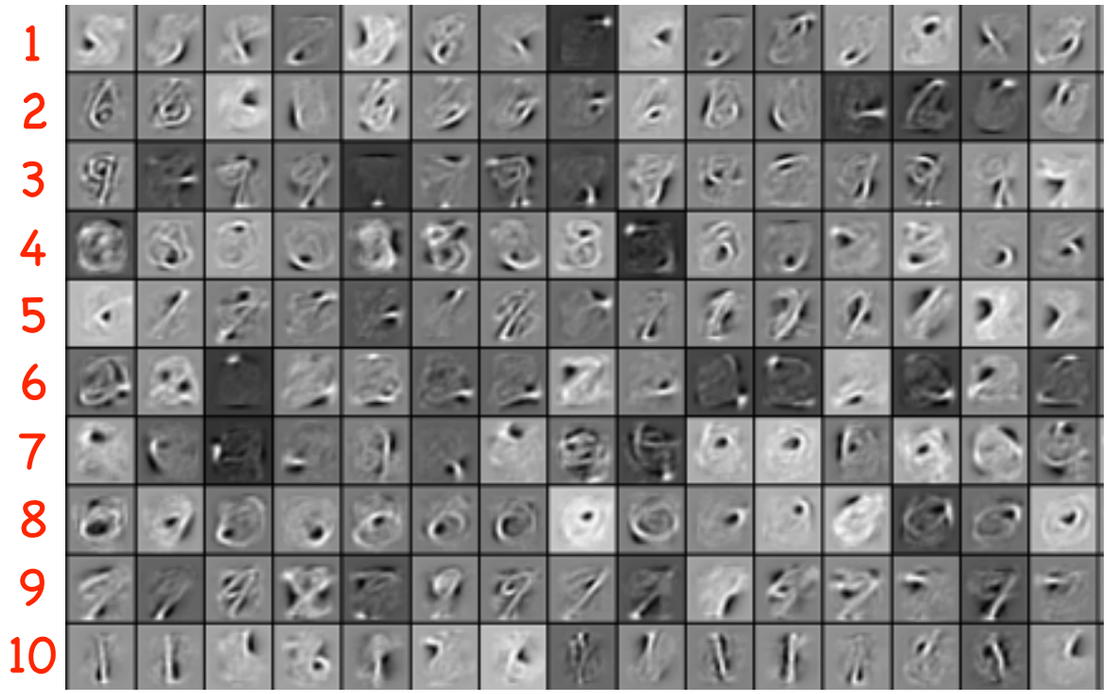


Sparsity Constraint: group sparsity

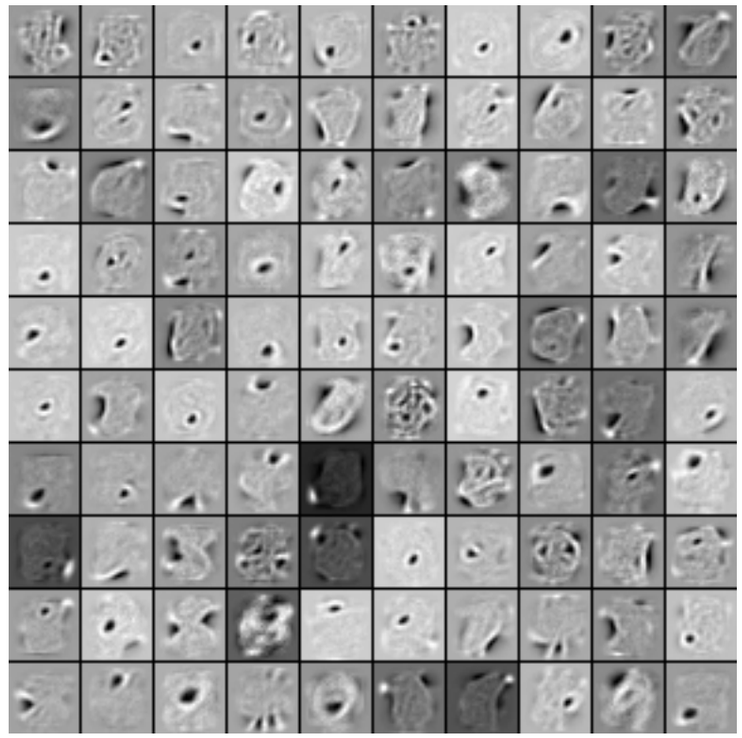
This is the visualization of W for hand written digit "0"



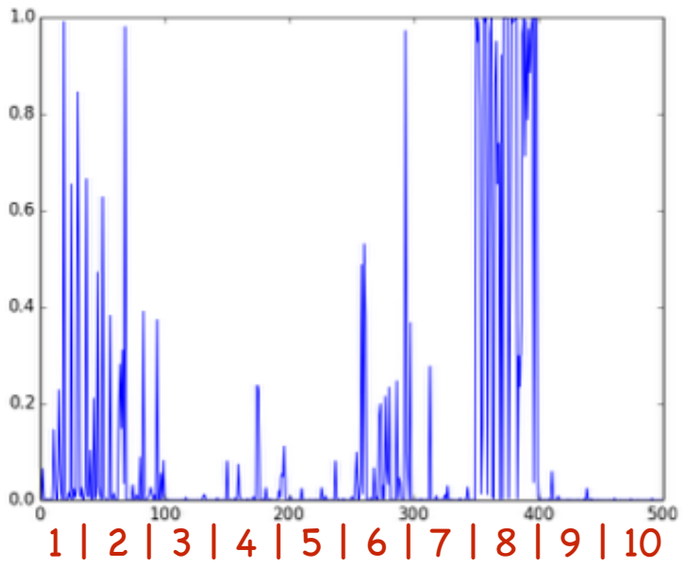
input image



Group Sparse Auto-Encoder



Conventional Auto-Encoder



- ✓ most of the responses are in group 1,2,6,8
- ✓ the results are reasonable: the groups which are similar to "0" get higher responses
- ✓ we can tell the patterns from each row (group)

Experiments

	TREC	INSUR	DMV	YAHOO dataset		
				sub	top	unseen
CNN	93.6	51.2	60	20.8	53.9	47
+sparsity	93.2	51.4	62	20.2	54.2	46
Random init. Weight	93.8	53.5	62	21.8	54.5	48
Question init. Weight	**94.2**	53.8	64	22.1	54.1	48
Ans. Init. Weight	-	**55.4**	**66**	**22.2**	**55.8**	**53**

Conclusions

- ✓ We have good improvement on Insurance and DMV dataset
- ✓ A little improvement for TREC and YAHOO dataset.
  - i) the question sentences are very short for these two datasets
- ✓ Our model perform well on unseen subcategories