Group Sparse CNNs for Question Classification with Answer Sets

Mingbo Ma
EECS
Oregon State University
mam@oregonstate.edu

Liang Huang
EECS
Oregon State University
liang.huang@oregonstate.edu

Bing Xiang
IBM Watson Group
T. J. Watson Research Center
bingxia@us.ibm.com

Bowen Zhou
IBM Watson Group
T. J. Watson Research Center
zhou@us.ibm.com

Abstract

Question classification is an important task with wide applications. However, traditional techniques treat questions as general sentences, ignoring the corresponding answer data. In order to consider answer information into question modeling, we first introduce novel group sparse autoencoders which refine question representation by utilizing group information in the answer set. We then propose novel group sparse CNNs which naturally learn question representation with respect to their answers by implanting group sparse autoencoders into traditional CNNs. The proposed model significantly outperform strong baselines on four datasets.

Motivation

General sentence modeling frameworks neglect two unique properties of question classification:

1. Question categories have hierarchical and overlapping structures
   - Question categories often have hierarchical structures
   - Question categories often have overlaps
   - Each question often belongs to multiple categories (multi-labeled)
2. 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.

Examples from NYDMV FAQs

There are 8 top-level categories, 47 sub-categories, and 537 questions (among them 388 are *unique*; many questions fall into multiple categories

- 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)

  ...
- Vehicle Registrations and Insurance
  - a: *Buy, sell, or transfer a vehicle* (22 questions)
  - b: *Reg. and title requirements* (42 questions)

  ...
- Driving Record / Tickets / Points

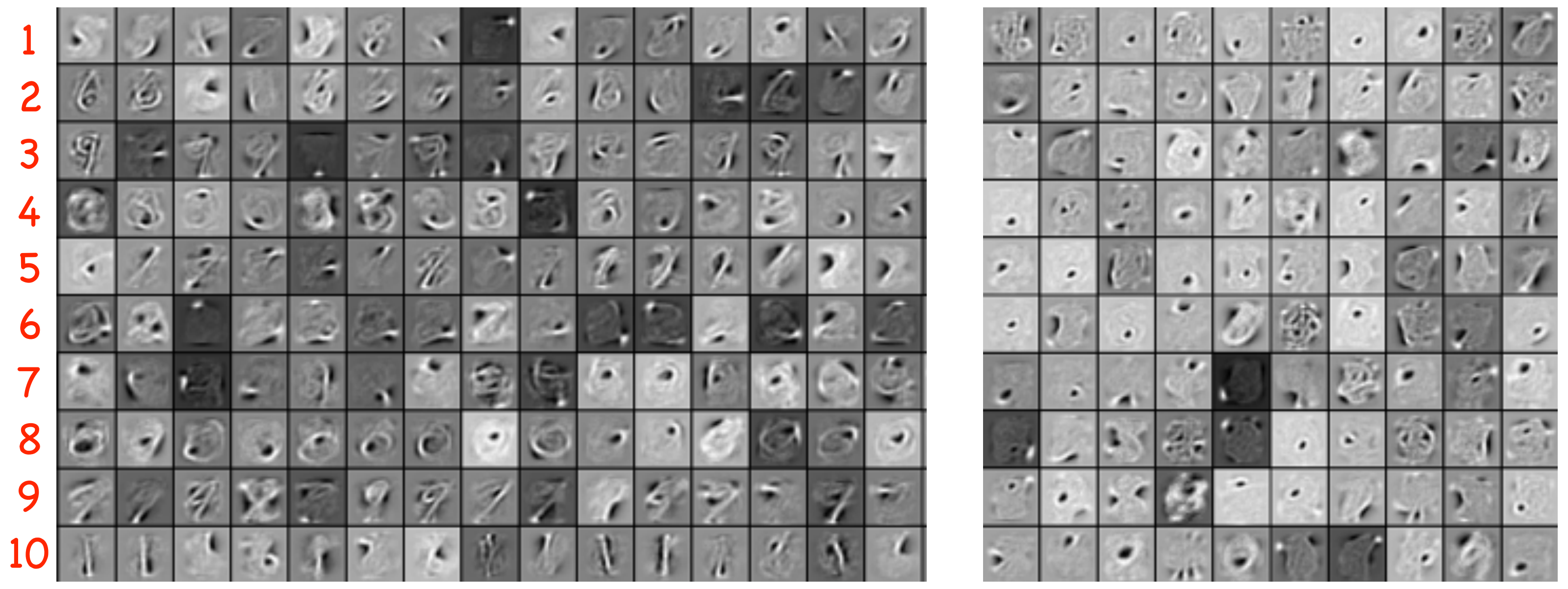
  ...

Why Group Sparse

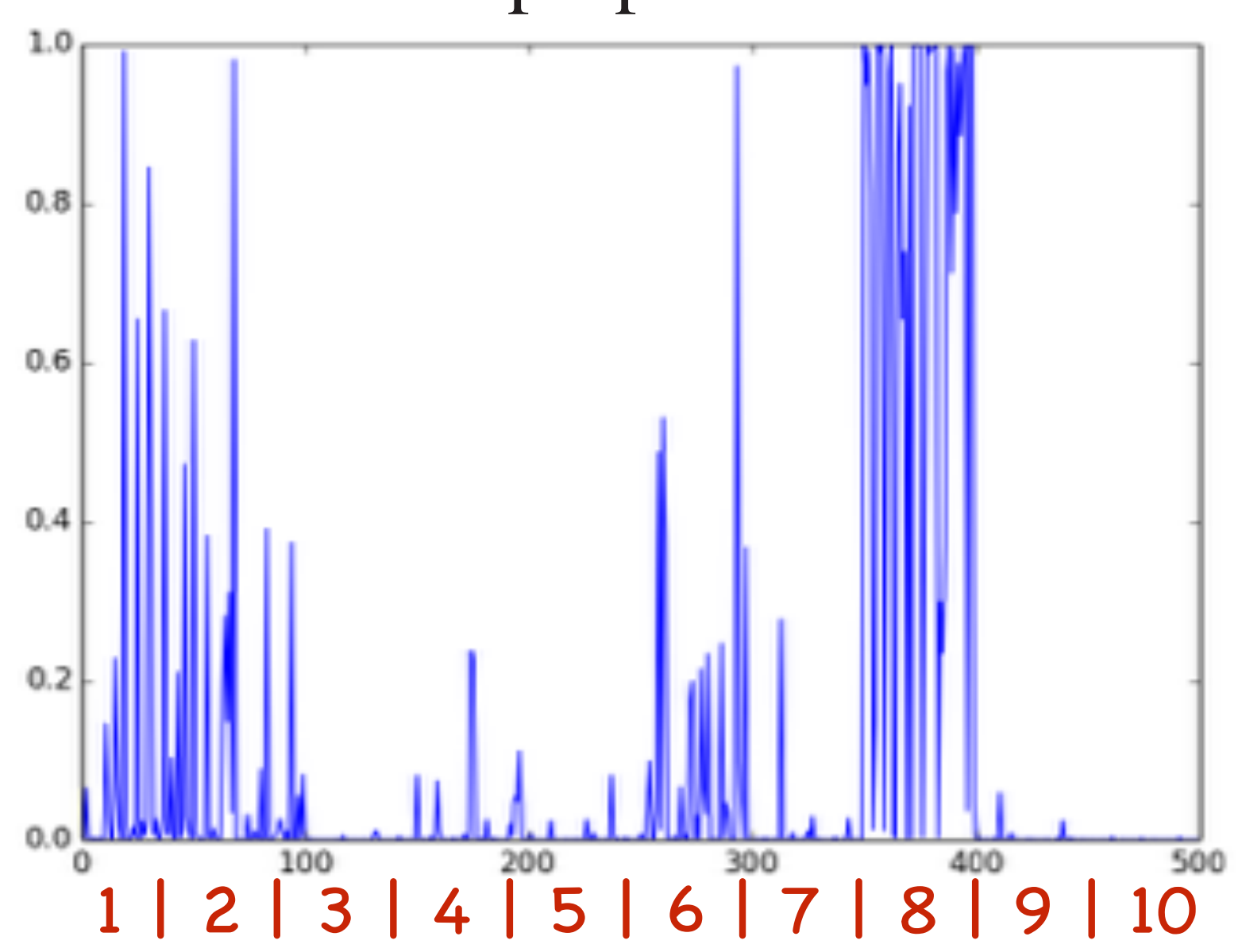
▷ Advantages of Group Sparse

- exploits the hierarchical and overlapping categories structures
- uses information from answers as dictionary (Rubinstein et al., 2010)

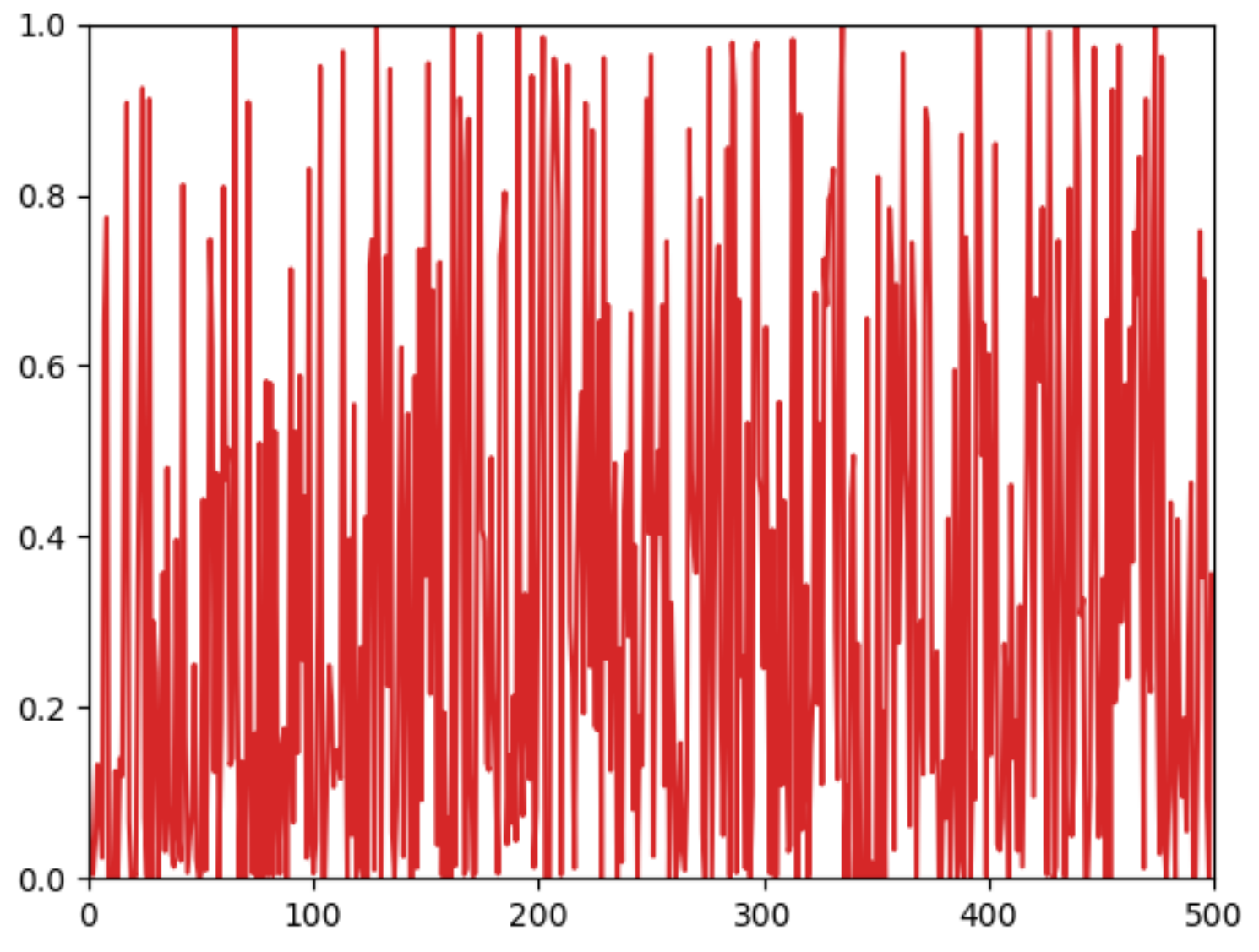
Visualization of trained projection matrix $\mathbf{W}$ on MNIST dataset (Left) and projection matrix from basic autoencoders (Right)



h in Group Sparse model



h in basic model



Our proposed Group Sparse Constrains

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

where $KL(\rho\|\hat{\rho}_j)$ and $KL(\eta\|\hat{\eta}_p)$ are group sparse constraints which are defined as follows:

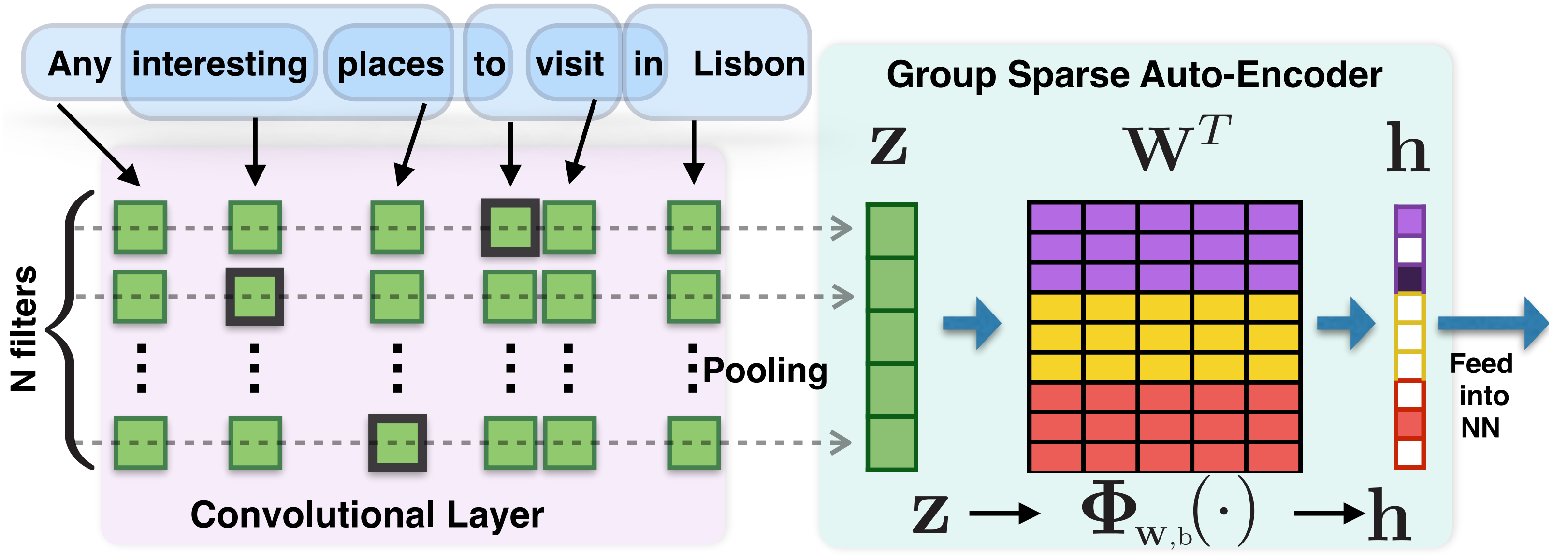
$$KL(\rho\|\hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}, \quad \text{where} \quad \hat{\rho}_j = \frac{1}{m} \sum_{i=1}^{m} \mathbf{h}_j^i$$

$$KL(\eta\|\hat{\eta}_p) = \eta \log \frac{\eta}{\hat{\eta}_p} + (1-\eta) \log \frac{1-\eta}{1-\hat{\eta}_p}, \quad \text{where} \quad \hat{\eta}_p = \frac{1}{mg} \sum_{i=1}^{m} \sum_{l=1}^{g} \|\mathbf{h}_{p,l}^i\|$$

where ρ and η are constant scalars which are our target sparsity and group-sparsity levels, resp. When α is set to zero, GSA only considers the structure sparsity between difference groups (Simon et al., 2013; Yuan and Lin, 2006). When β is set to zero, GSA is reduced to Sparse Autoencoders (Ng, 2011).

Group Sparse CNNs

We propose Group Sparse Convolutional Neural Networks (GSCNNs) by placing one extra layer between the convolutional and the classification layers. This extra layer mimics the functionality of Group Sparse Autoencoders



After the conventional convolutional layer, we get the feature map **z** for each sentence. Instead of directly feeding it into a fully connected neural network for classification, we enforce the group sparse constraint on **z** in a way similar to the group sparse constraints on hidden layer in GSA.

Experiments

Experimental results. Baselines: †sequential CNNs ($\alpha = \beta = 0$), ‡CNNs with global sparsity ($\beta = 0$). $\mathbf{W}_R$: randomly initialized projection matrix. $\mathbf{W}_Q$: question-initialized projection matrix. $\mathbf{W}_A$: answer set-initialized projection matrix. There are three different classification settings for Yahoo: subcategory, top-level category, and top-level accuracies on unseen sub-labels.

	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
$\mathbf{W}_R$	93.8	53.5	62	21.8	54.5	48
$\mathbf{W}_Q$	**94.2**	53.8	64	22.1	54.1	48
$\mathbf{W}_A$	-	**55.4**	**66**	**22.2**	**55.8**	**53**

In the unseen experiments, there are a few sub-category labels that are not included in the training data. However, we still hope that our model could still return the correct parent category for these unseen subcategories at test time.

Key References

Andrew Ng. 2011. Sparse autoencoder. In *CS294A Lecture notes*, page 72. Stanford University.

R. Rubinstein, A. M. Bruckstein, and M. Elad. 2010. Dictionaries for sparse representation modeling. In *Neural Computation*.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2013. A sparse-group lasso. In *Journal of Computational and Graphical Statistics*.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. In *Journal of the Royal Statistical Society*, volume 68, pages 49–67.