Appendix

1 Single-Label Dataset

Dataset	Train	Dev	Test	Classes	Classification Task
MR	8.5k	1.0k	1.1k	2	review classification
SST-2	8.6k	1.0k	1.8k	2	sentiment analysis
Subj	8.1k	0.9k	1.0k	2	opinion classification
TREC	5.3k	1.1k	0.5k	6	question categorization
CR	3.1k	0.3k	0.4k	2	review classification
AG's news	108k	12k	7.6k	4	news categorization

Table A.1: Characteristics of the datasets. We picked 10 percent of training data randomly as dev data to train the best model and evaluate the model with the test data

	MR	**SST2**	**Subj**	**TREC**	**CR**	**AG's**
LSTM	75.9	80.6	89.3	86.8	78.4	86.1
BiLSTM	79.3	83.2	90.5	89.6	82.1	88.2
Tree-LSTM	80.7	85.7	91.3	91.8	83.2	90.1
LR-LSTM	81.5	87.5	89.9	-	82.5	-
CNN-rand	76.1	82.7	89.6	91.2	79.8	92.2
CNN-static	81.0	86.8	93.0	92.8	84.7	91.4
CNN-non-static	81.5	87.2	93.4	93.6	84.3	92.3
CL-CNN	-	-	88.4	85.7	-	92.3
VD-CNN	-	-	88.2	85.4	-	91.3
Capsule-A	81.3	86.4	93.3	91.8	83.8	92.1
Capsule-B	82.3*	86.8	93.8	92.8	85.1*	92.6

Table A.2: Comparisons of our capsule networks and baselines on six text classification benchmarks with significance test. Numbers with * mean that improvement from the model is statistically significant over the baseline methods (t-test, p-value < 0.05).

	MR	**SST2**	**Subj**	**TREC**	**CR**	**AG's**
SVM	80.8	86.3	92.3	92.5	83.4	91.8
LSTM	75.9	80.6	89.3	86.8	78.4	86.1
BiLSTM	79.3	83.2	90.5	89.6	82.1	88.2
Tree-LSTM	80.7	85.7	91.3	91.8	83.2	90.1
LR-LSTM	81.5	87.5	89.9	-	82.5	-
CNN-rand	76.1	82.7	89.6	91.2	79.8	92.2
CNN-static	81.0	86.8	93.0	92.8	84.7	91.4
CNN-non-static	81.5	87.2	93.4	93.6	84.3	92.3
CL-CNN	-	-	88.4	85.7	-	92.3
VD-CNN	-	-	88.2	85.4	-	91.3
Bi-BloSAN [1]	**83.1**	**87.4**	**94.5**	**94.8**	84.8	**93.3**
Capsule-A	81.3	86.4	93.3	91.8	83.8	92.1
Capsule-B	82.3	86.8	93.8	92.8	**85.1**	92.6

Table A.3: Comparisons of our capsule networks and baselines on six text classification benchmarks. (Adding SVM and Bi-BloSAN as baselines)

2 Multi-Label Experimenet

Dataset	Train	Dev	Test	Category	Description
Reuters-Multi-label	5.8k	0.6k	0.3k	10	Remove single-label samples in test data.
Reuters-Full	5.8k	0.6k	3.4k	10	Include single- and multi-label samples in test data.

Table A.4: Characteristics of Reuters-21578 (10 categories) corpus

	Reuters-Multi-label				Reuters-Full			
	ER	Precision	Recall	F1	ER	Precision	Recall	F1
LSTM	23.3	86.7	54.7	63.5	62.5	78.6	72.6	74.0
BiLSTM	26.4	82.3	55.9	64.6	65.8	83.7	75.4	77.8
CNN-rand	22.5	88.6	56.4	67.1	63.4	78.7	71.5	73.6
CNN-static	27.1	91.1	59.1	69.7	63.3	78.5	71.2	73.3
CNN-non-static	27.4	92.0	59.7	70.4	64.1	80.6	72.7	75.0
SVM	26.3	90.7	57.6	68.2	63.2	81.2	73.3	75.1
Pruned-Sets [2]	29.6	93.1	62.2	72.5	66.1	84.2	76.5	78.2
Bi-BloSAN [1]	28.7	92.8	61.2	71.7	66.0	84.3	76.2	78.0
Capsule-A	57.2	88.2	80.1	82.0	66.0	83.9	**80.5**	80.2
Capsule-B	**60.3**	**95.4**	**82.0**	**85.8**	**67.7**	**86.4**	80.1	**81.4**

Table A.5: Comparisons of the capability for transferring from single-label to multi-label text classification on Reuters-Multi-label and Reuters-Full datasets. For fair comparison, we use margin-loss for our model and other baselines.

1. Shen T, Zhou T, Long G, et al. Bi-directional block self-attention for fast and memory-efficient sequence modeling[J]. arXiv preprint arXiv:1804.00857, 2018.

2. Read J, Pfahringer B, Holmes G. Multi-label classification using ensembles of pruned sets, Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008: 995-1000. (Label Correlation Algorithm)