Dataset	Train	Dev	Test	Classes	Type
MR	9500	-	1100	2	review
SST-1	8544	1101	2210	5	sentiment
SST-2	6920	872	1821	2	sentiment
Subj	9000	-	1000	2	subjectivity
TREC	5900	-	500	6	question
AG's	120k	-	7600	4	news

Table 1: Statistics for six datasets

Dataset	MR	SST-1	SST-2	Subj	TREC	AG's
LSTM	75.9	45.9	80.6	89.3	86.8	86.1
BiLSTM	79.3	46.2	83.2	90.5	89.6	88.2
LR-LSTM	81.5	48.2	87.5	89.9	-	-
VD-CNN	-	-	-	-	-	91.3
DCNN	-	48.5	86.8	-	**93.0**	-
CNN-MC	81.1	47.4	**88.1**	93.2	92.2	-
CapsNet-1	**81.5**	48.1	86.4	**93.3**	91.8	91.1
CapsNet-2	**82.4**	**48.7**	87.8	**93.6**	92.9	**92.3**
- Orphan	81.9	48.3	87.2	93.4	92.6	91.7

Table 2: Single-task results. Row "- Orphan category" denotes a variant of CapsNet-2 without orphan category

each tasks. And the whole network is optimized in a stochastic way with multi-task training (Section 3.5).

Implement Details For word embedding, we use the word vectors in *Word2Vec* (Mikolov et al., 2013), which is 300-dimensional and has 3M vocabularies. And all the routing logits $b_{ij}^{(k)}$ is initialized to zero, so that all the capsules in adjacent layers $(\hat{\mathbf{u}}_{j|i}, \mathbf{v}_j)$ are connected with equal possibility c_{ij}. The coupling coefficients are updated by routing with 3 iterations, which performs best for our approach. For training, we use Adam optimizer (Kingma and Ba, 2014) with exponentially decaying learning rate. Moreover, we use mini-batch with size of 8 for all the datasets.

4 Experiment

We test our capsule-based models on six datasets in both single-task and multi-task scenarios to demonstrate the effectiveness of our approaches. We also in this section conduct some investigations like ablation study and visualization to give a comprehensive understanding to the characteristics of our models.

4.1 Datasets

For both single-task and multi-task scenarios, we conduct extensive experiments on six benchmarks: movie reviews (MR) (Bo and Lee, 2005), Stanford Sentiment Treebank (SST-1 and SST-2) (Socher et al., 2013), subjectivity classification (Subj) (Pang et al., 2004), question dataset (TREC) (Li and Roth, 2002), AG's news corpus (Mousa et al., 2017). These datasets cover a wide range of text classification tasks, which can fully test a model and the details are listed in Table 1.

4.2 Competitors

To demonstrate the effectiveness of our capsule network, we compare the single-task architectures with several state-of-the-art models, involving LSTM/BiLSTM (Cho et al., 2014), LSTM regularized by linguistic knowledge (LR-LSTM) (Qian et al., 2016), very deep network (VD-CNN) (Conneau et al., 2016), dynamic CNN (DCNN) (Kalchbrenner et al., 2014), CNN with multiple channels (CNN-MC) (Kim, 2014). Also, we compare the multi-task architecture (Figure 2) with several strong baselines of multi-task learning, including a general architecture for multi-task learning (MT-GRNN) (Zhang et al., 2017), recurrent neural network based multi-task learning (MT-RNN) (Liu et al., 2016), convolutional neural network with multi-task learning (MT-DNN) (Collobert and Weston, 2008), deep neural network with multi-task learning (MT-CNN) (Liu et al., 2015).

4.3 Single-Task Learning Results

We first test our approach on six datasets for text classification under the scheme of single-task. As Table 2 shows, our single-task network enhanced by capsules is already a strong model. CapsNet-1 that has one kernel size obtains the best accuracy on 2 out of 6 datasets, and gets competitive results on the others. And CapsNet-2 with multiple kernel sizes further improves the performance and get best accuracy on 4 datasets. This proves our capsule networks are effective for text. Particularly, our capsule network outperforms conventional CNNs like DCNN, CNN-MC and VD-CNN with a large margin (by average 1.1%, 0.7% and 1.0% respectively), which shows the advantages of capsule network over conventional CNNs for clustering features and leveraging the position information.

Routing Iteration The coupling coefficients c_{ij} are updated by dynamic routing algorithm, which determines the connections between the capsules. To find the best updating iteration for coupling co-