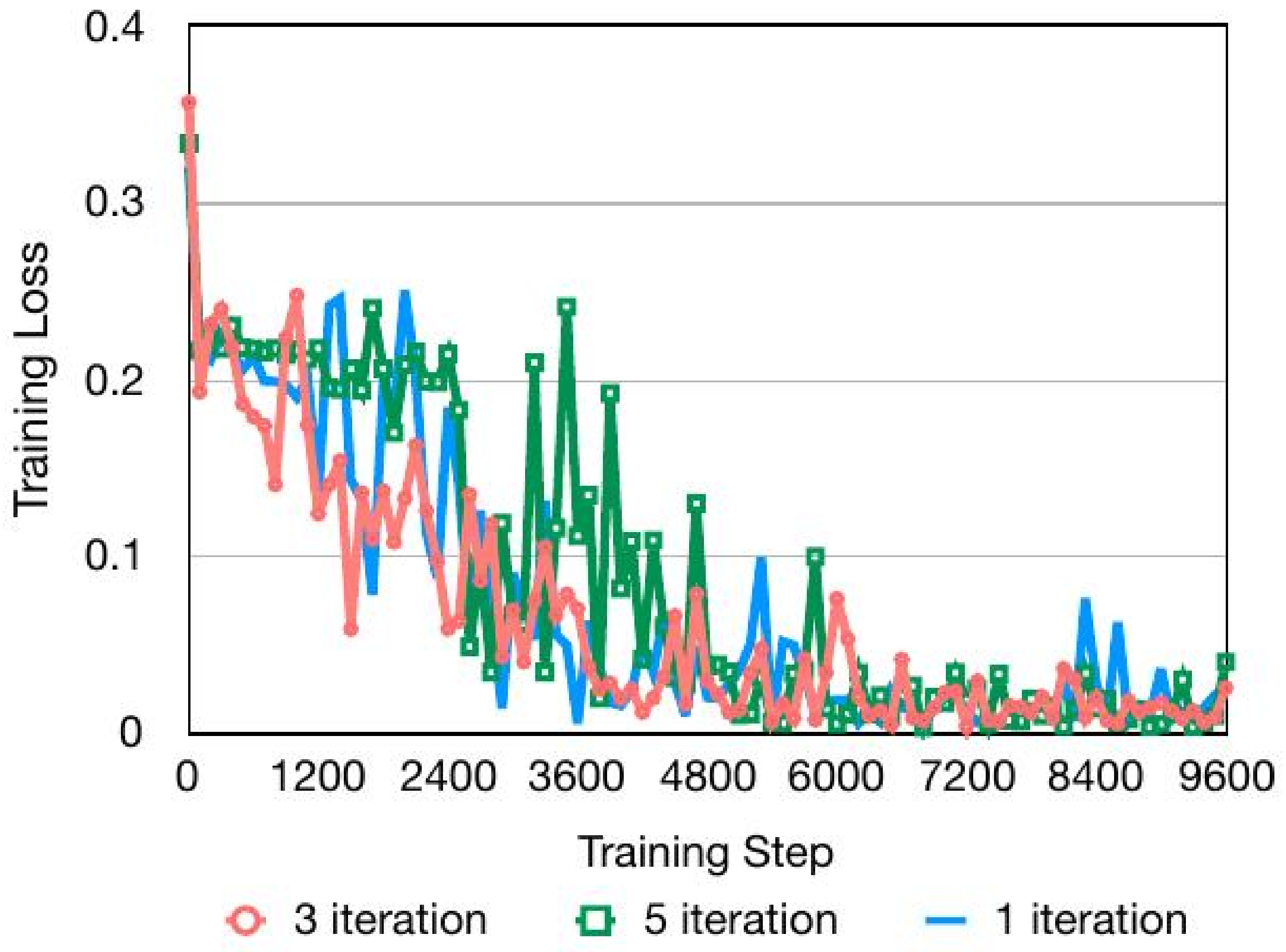

Figure 3: Influence of routing iteration

efficients, we test the CapsNet-2 with a series of iterations (1, 3 and 5) on MR dataset. As shown in Figure 3, network with 3 iterations convergences fast and performs best, which stays in line with the conclusion in (Sabour et al., 2017). So we utilize 3 iterations in all our experiments.

Ablation Study on Orphan Category Orphan category in class capsule layer helps collect the noise capsules that contain the 'background' information like stop words, punctuations or any unrelated words. We conduct the ablation experiment on orphan category, and result (Table 2) shows that network with orphan category perform better than the without one by 0.4%. This demonstrates the effectiveness of orphan category.

4.4 Multi-Task Learning Results

Up to now, we have obtained an optimized single-task architecture. In this section, we equip CapsNet-2 with the *task routing* and multi-task training procedure, namely the model MCapsNet, so that this capsule based architecture can learn several datasets in a unified network. Extensive experiments are conducted in this section to demonstrate the effectiveness of our multi-task learning architecture, as well as its ability for feature clustering.

Multi-Task Performance
We simultaneously train our model McapsNet on six tasks in Table 1 and compare it with single-task scenario (Table 3). We can see that our multi-task architecture clearly improves the performance over the single task models, which demonstrates the benefits of our multi-task architecture.

Dataset	MR	SST-1	SST-2	Subj	TREC	AG's	Avg.△
BiLSTM	79.3	46.2	83.2	90.5	89.6	88.2	+0
MT-GRNN		49.2	87.7	94.3	93.8		+2.6
MT-RNN		49.6	87.9	94.1	91.8		+3.5
MT-DNN	82.1	48.1	87.3	93.9	92.2	91.6	+2.9
MT-CNN	81.6	49.0	86.9	93.6	91.8	91.9	+3.0
CapsNet-1	81.5	48.1	86.4	93.3	91.8	91.1	+2.5
CapsNet-2	82.4	48.7	87.8	93.6	92.9	92.3	+3.3
MCapsNet	83.5	49.7	88.6	94.5	94.2	93.8	+4.6

Table 3: Multi-task results of MCapsNet. In column Avg.△, we use BiLSTM as baseline and calculate the average improvements over it.

As Table 3 shows, MCapsNet also outperforms the state-of-the-art multi-task learning models by at least 1.1%. This shows the advantages of our task routing algorithm, which can cluster the features for each task, instead of freely sharing the features among tasks.

4.5 Routing Visualization

To show the mechanism how capsule benefits the multi-task learning, we visualize the coupling coefficient $c_{ij}^{(k)} \in [0,1]$ between primary and class capsules. We use kernel with size 1 for primary capsule layer so that every capsule represents only one 3-gram phrase. The strength of these connections indicates the importance of these 3-grams to their corresponding task and class.

We feed a random sample from the dataset MR into MCapsNet. In the first row of Table 4, we show the most important 3-gram phrases for two tasks MR and Subj (two classes for each) with word cloud. The sizes of the grams represent the weights of coupling coefficients. We can see that task routing algorithm helps lead the grams into the most related tasks, which allows each task only consider the helpful features for them. In another word, task routing builds a feature space for each task and avoids they contaminate each other. This demonstrates that MCapsNet has the ability of feature clustering, which can benefit MTL by reducing the interference.

We also illustrate the coupling coefficients sequentially for each task. The height of the blue and gray lines represents the polarity of positivity and subjectivity respectively. It is clear that MCapsNet can focus on the appropriate positions for each task, which helps make the final correct predication for every task.