Master Thesis:
Dynamic Memory based Capsule Networks for Few-Shot Text Classification

Task Definition: Few-Shot Text Classification

- **What is Few-Shot Learning?**
  - Task in which a classifier must be adapted to recognize to new classes not seen in training

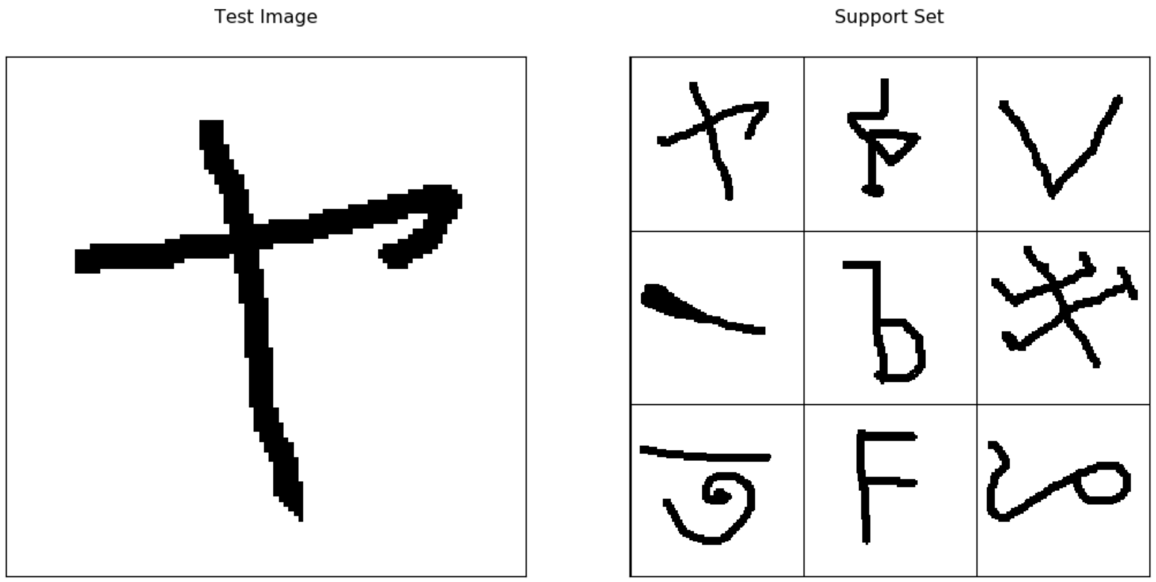

Figure 1 : 9-Way 1-Shot Task

https://sorenbouma.github.io/blog/oneshot/

How to remedy this problem?

- **The nature of the problem : Unreliable Empirical Risk Minimizer**
  - Expected Risk : Given hypothesis h, we want to minimize its expected risk R, which is the loss measured with respect to $p(x, y)$. Specifically,

$$R(h) = \int \ell(h(x), y)\, dp(x, y) = \mathbb{E}_{(x,y)\sim P(x,y)}[\ell(h(x;\theta), y)]$$

  - Empirical Risk : but $p(x, y)$ is unknown, instead we use the empirical risk

$$R_I(h) = \frac{1}{I}\sum_{i=1}^{I} \ell(h(x_i), y_i)$$

How to remedy this problem?

- **The nature of the problem : Unreliable Empirical Risk Minimizer**
  - The total error can be decomposed as

$$\mathbb{E}\left[R(h_I) - R(\hat{h})\right] = \underbrace{\mathbb{E}\left[R(h^*) - R(\hat{h})\right]}_{\mathcal{E}_{\text{app}}(\mathcal{H})} + \underbrace{\mathbb{E}\left[R(h_I) - R(h^*)\right]}_{\mathcal{E}_{\text{est}}(\mathcal{H},I)}$$

  - where
    - $\hat{h} = argmin_h R(h)$ be the function that minimizes the expected risk
    - $h^* = argmin_{h\in\mathcal{H}} R(h)$ be the function in $\mathcal{H}$ that minimizes the expected risk
    - $h_I = argmin_{h\in\mathcal{H}} R_I(h)$ be the function in $\mathcal{H}$ that minimizes the empirical risk
    - $\mathcal{E}_{\text{app}}(\mathcal{H})$: the approximation error measures how close the functions in $\mathcal{H}$ can approximate the optimal hypothesis $\hat{h}$
    - $\mathcal{E}_{\text{est}}(\mathcal{H}, I)$: the empirical error measures the effect of minimizing the empirical risk instead of the expected risk within $\mathcal{H}$
  - Learning to reduce the total error can be attempted from the perspectives of (i) data and (ii) model

How to remedy this problem?

- **The nature of the problem : Unreliable Empirical Risk Minimizer**
  - $\mathcal{E}_{\text{est}}\ (\mathcal{H}, I)$ can be reduced by having a larger number of samples
  - But, in FSL, the number of available examples I is so small that empirical risk $R_I(h)$ may then be far from being a good approximation of the expected risk $R(h)$, and the resultant empirical risk minimizer h_I overfits.

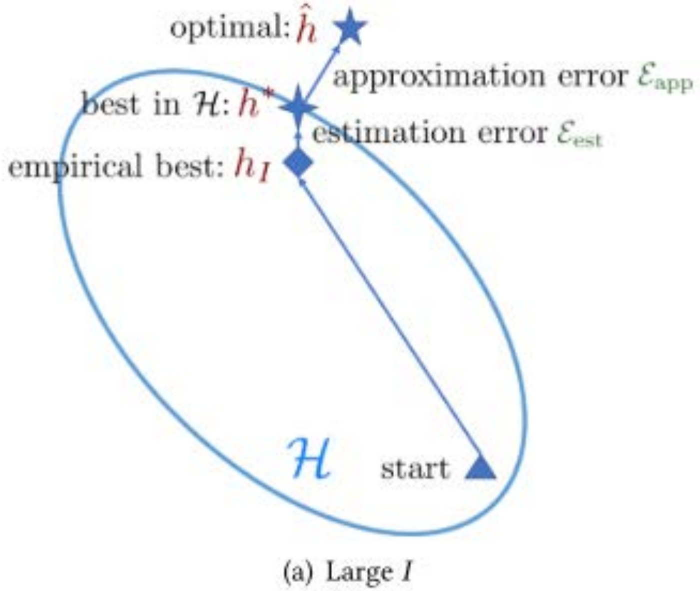

Figure 2 :Comparison of learning with sufficient and few training samples

How to remedy this problem?

- **2 Ways to leverage prior knowledge : Data, Model**
  - **Data** : methods use prior knowledge to augment D_{train}, and increase the number of samples from I to $\tilde{I}$.
  - **Model** : methods use prior knowledge to constrain the complexity of $\mathcal{H}$, which results in a smaller hypothesis space $\widetilde{\mathcal{H}}$

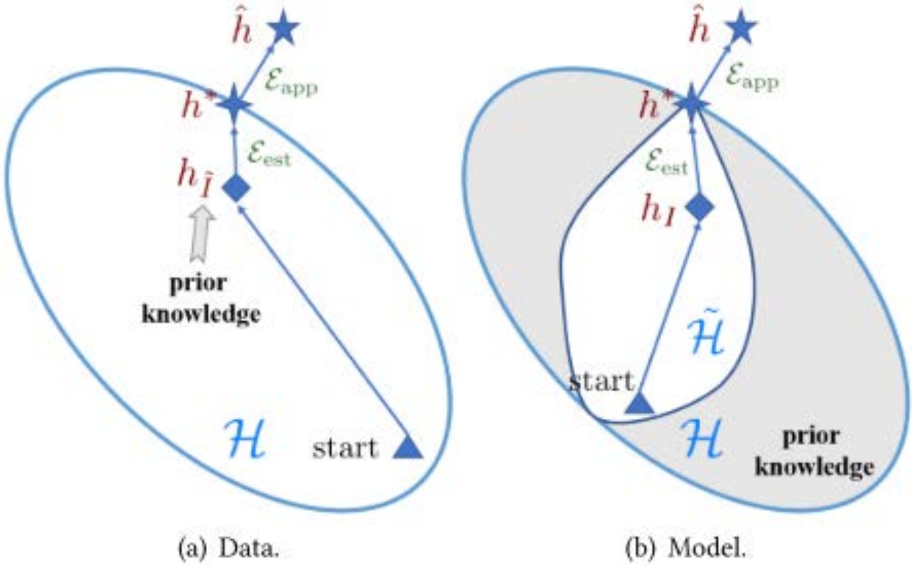

Figure 3 :Different perspectives on how FSL methods solve the few-shot problem

Wang, Yaqing, et al., "Generalizing from a few examples: A survey on few-shot learning", In ACM Computing Surveys (CSUR), Vol. 53, Issue. 3, pp. 1-34, 2020

How have researchers approached it before?

- **Method : Meta-learning**
  - Method that enhances "**model"** using **prior knowledge**
  - Meta-Learning
    - Learn how to learn
    - By episodic training, Meta-learning helps the model to learn the prior knowledge that can be applied to the new task
    - Taxonomy : Optimization-based, **Metric-based**, Model-based,
  - Episodic Training
    - A method in which the model learns general(prior) knowledge about task learning by learning from the distribution of tasks similar to the target task

Wang, Yaqing, et al., "Generalizing from a few examples: A survey on few-shot learning", In ACM Computing Surveys (CSUR), Vol. 53, Issue. 3, pp. 1-34, 2020
Hospedales, Timothy, et al., "Meta-learning in neural networks: A survey", In arXiv preprint arXiv:2004.05439, 2020
Ravi, Sachin, and Hugo Larochelle. "Optimization as a model for few-shot learning.", In the International Conference on Learning Representations(ICLR), 2016

How have researchers approached it before?

- **Episodic Training – class as task**

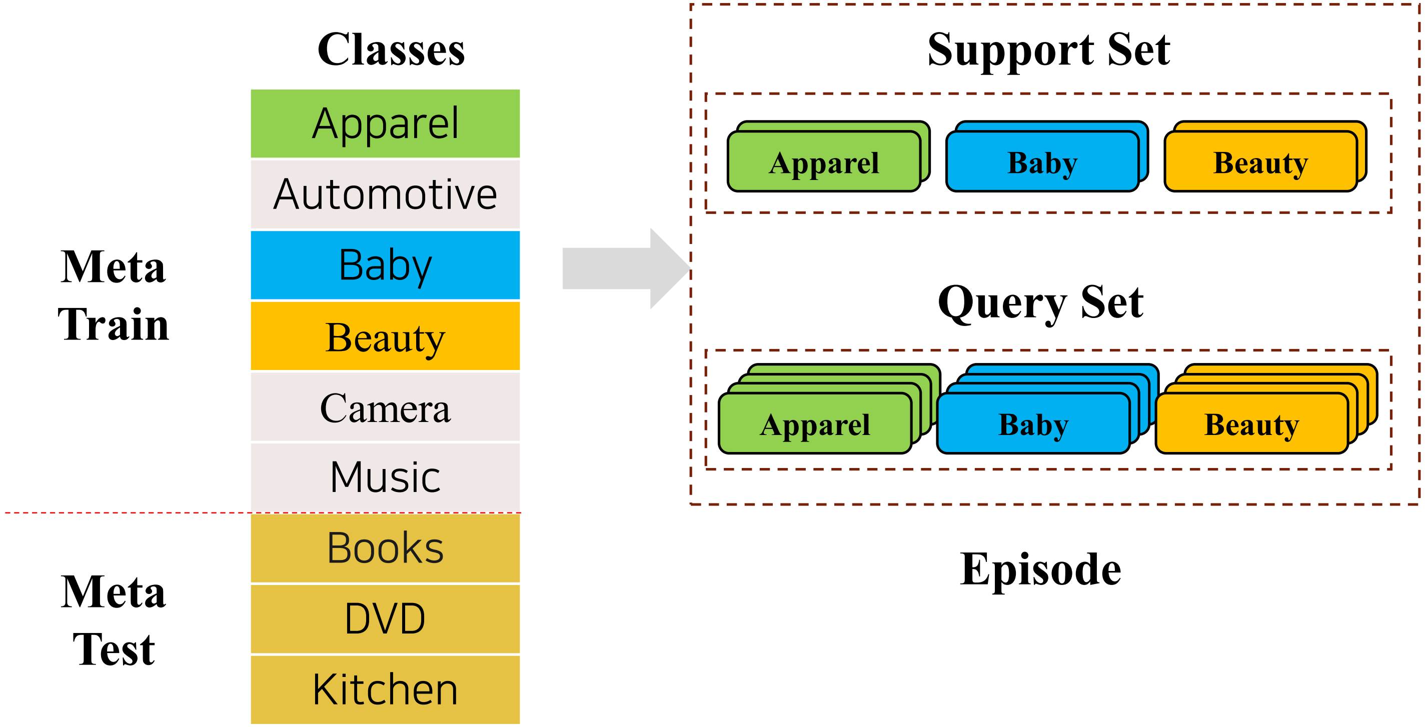

Figure 4 : Task sampling form meta-train set

- Split the dataset into Meta-Train & Meta-Test
- Select N classes
- N -way K -Shot **Support Set**
- N -way **Query Set**

Geng, Ruiying, et al., "Dynamic Memory Induction Networks for Few-Shot Text Classification", In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics(ACL), pp.1087-1094, 2020

How have researchers approached it before?

- **Episodic Training – class as task**

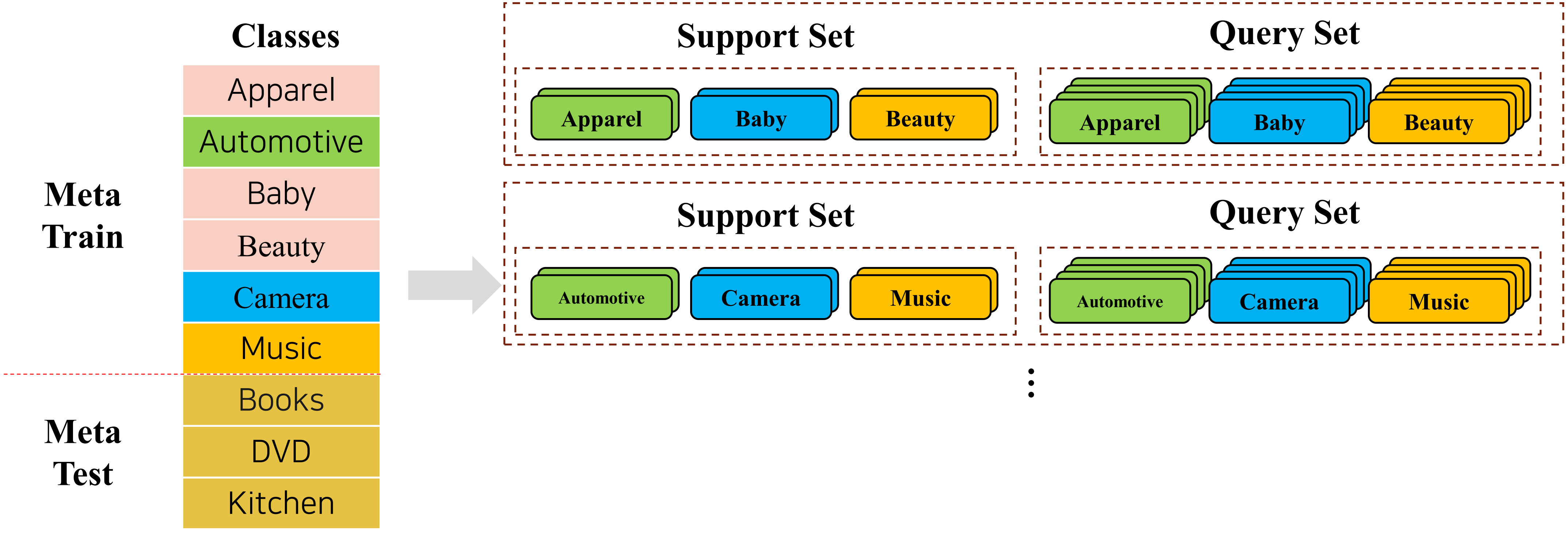

Figure 4 : Task sampling form meta-train set

Geng, Ruiying, et al., "Dynamic Memory Induction Networks for Few-Shot Text Classification", In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics(ACL), pp.1087-1094, 2020

How have researchers approached it before?

- **Episodic Training – domain as task**

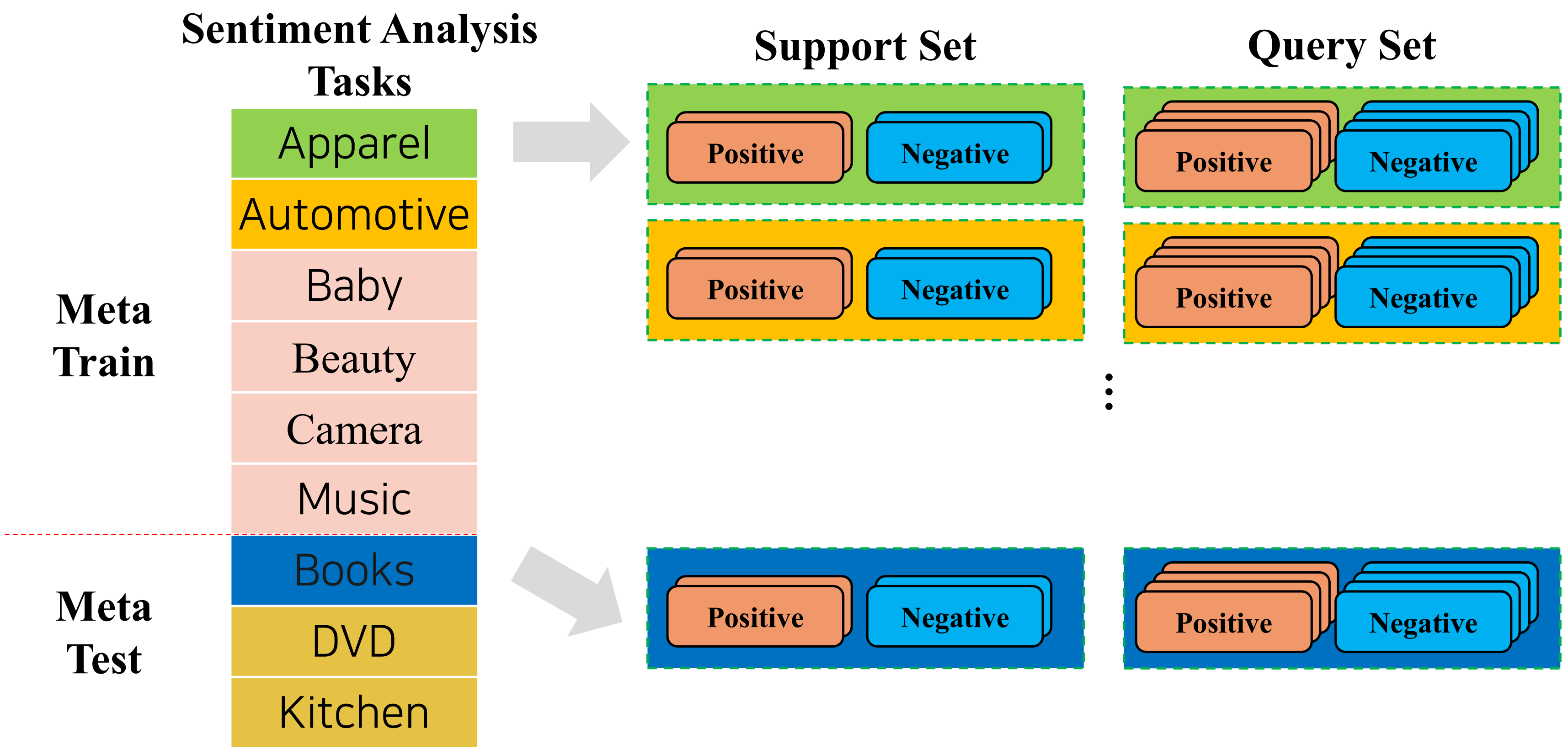

Figure 5 : Task sampling process of few-shot text classification(domain as task)

How have researchers approached it before?

- **Metric-based Meta-Learning**
  - Task-invariant Embedding Model
  - Learn a general embedding model that samples of different classes to be well distinguished by episodic learning
  - Neural Networks based weighted K-NN

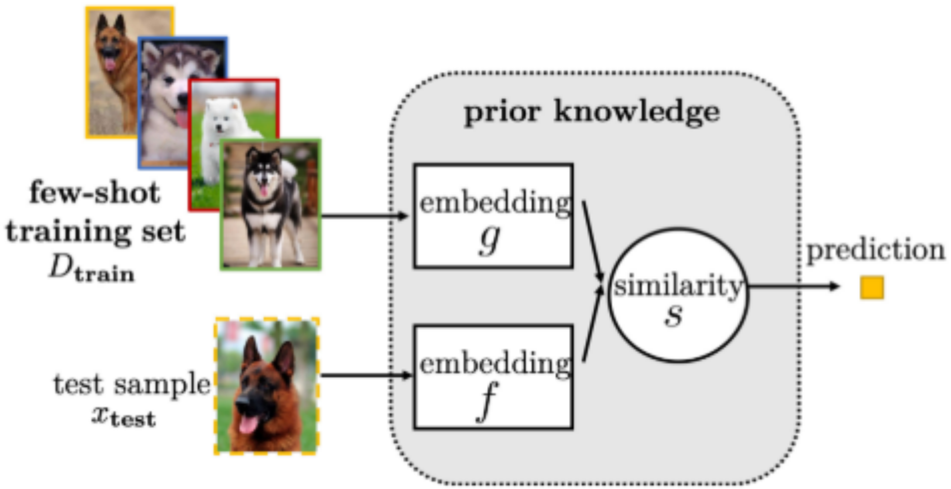

Figure 6 :Solving the FSL problem by Metric-based Meta-Learning

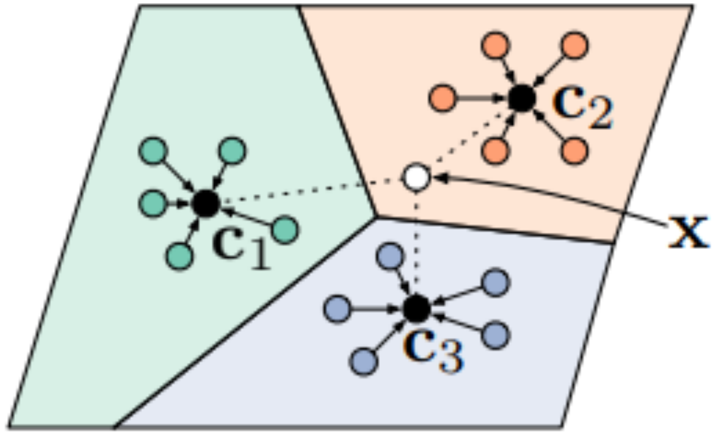

Figure 7 : Prototypical Network Architecture

Wang, Yaqing, et al., "Generalizing from a few examples: A survey on few-shot learning", In ACM Computing Surveys (CSUR), Vol. 53, Issue. 3, pp. 1-34, 2020
Snell, Jake, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning", In Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS), pp. 4080-4090, 2017

Related Works

- **Induction Networks for Few-Shot Text Classification(Geng et al., 2019)**
  - Sample-wise comparison to the support set can be disturbed by the various expression in the same class
  - To learn a generalized class-wise representation, Induction Networks use the dynamic routing algorithm

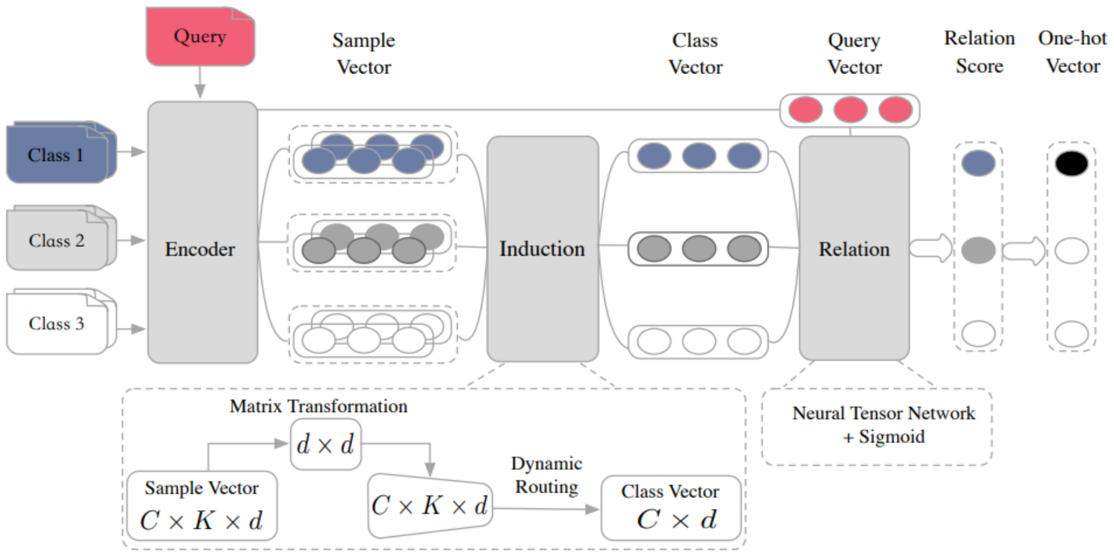

Figure 8 : Induction Networks architecture
for a *C*-way *K*-shot (*C*=3, *K*=2) problem with one query example

Geng, Ruiying, et al. "Induction networks for few-shot text classification." *arXiv preprint arXiv:1902.10482* (2019).

Related Works

- **Learning with external memory**
  - extract knowledge from training set and stores it in external memory
  - each new sample x_{test} is represented by a weighted average of contents extracted from the memory

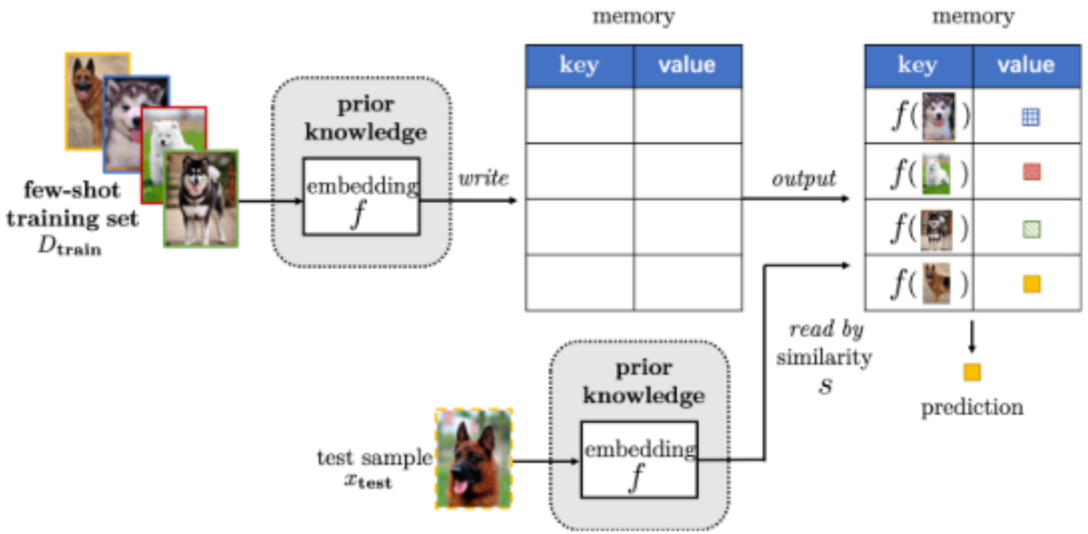

Figure 9 : Solving the FSL problem by learning with external memory

Related Works

- **Dynamic Memory Induction Networks for Few-Shot Text Classification(Geng et al., 2020)**
  - Leverage class representations acquired from fine-tuned BERT as memory(prior knowledge)

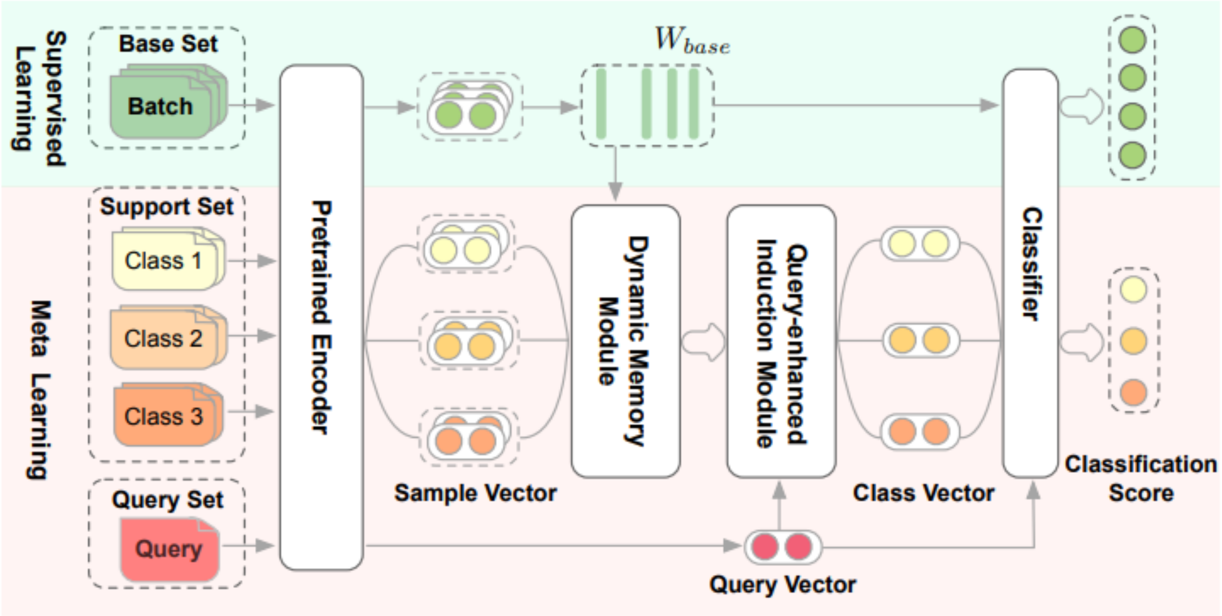

Figure 10 : An Overview of Dynamic Memory Induction Network with a 3-way 2-shot example

Geng, Ruiying, et al. "Induction networks for few-shot text classification." *arXiv preprint arXiv:1902.10482* (2019).

Limitations of related works

- **DMIN requires supervised learning phase to acquire the prior knowledge**
  - If the size of the training set is small then, the model may overfit on supervised learning phase
- **Because DMIN trains BERT together, it requires a significant amount of computing resources**
- **Not only task-invariant embedding but also task-specific embedding need to be considered to classify the query sample and reduce the task diversity error**

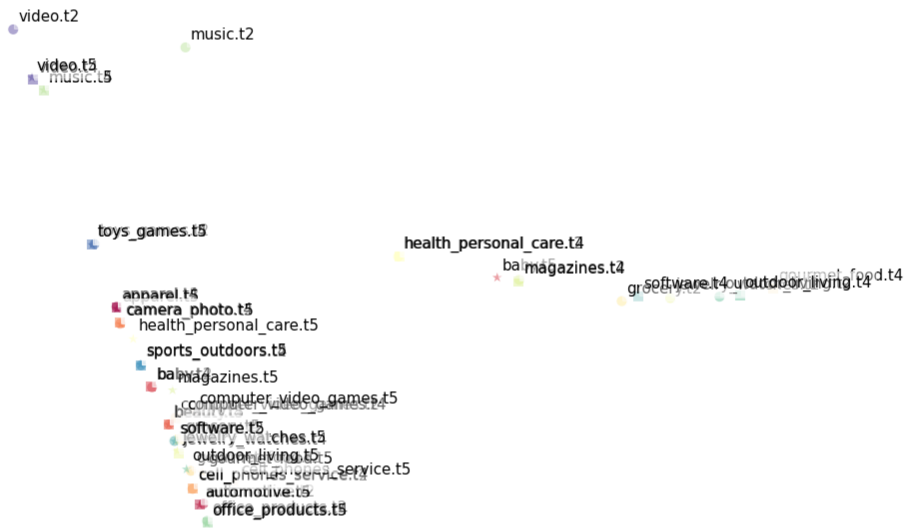

Figure 11 : Task embedding visualization of the ARSC dataset

Contributions of Proposed Method

- **Replace the supervised learning phase with the unsupervised learning phase(MLM, NSP)**

- **Enable efficient learning , by fine-tuning only a portion of the parameters of the entire model that relevant to task specific knowledge**

- **Achieve better performance than exists methods, by considering both task-specific and task-invariant embedding**

Proposed Method : Dynamic Memory based Capsule Networks(DMCN)

- **Overall Architecture**

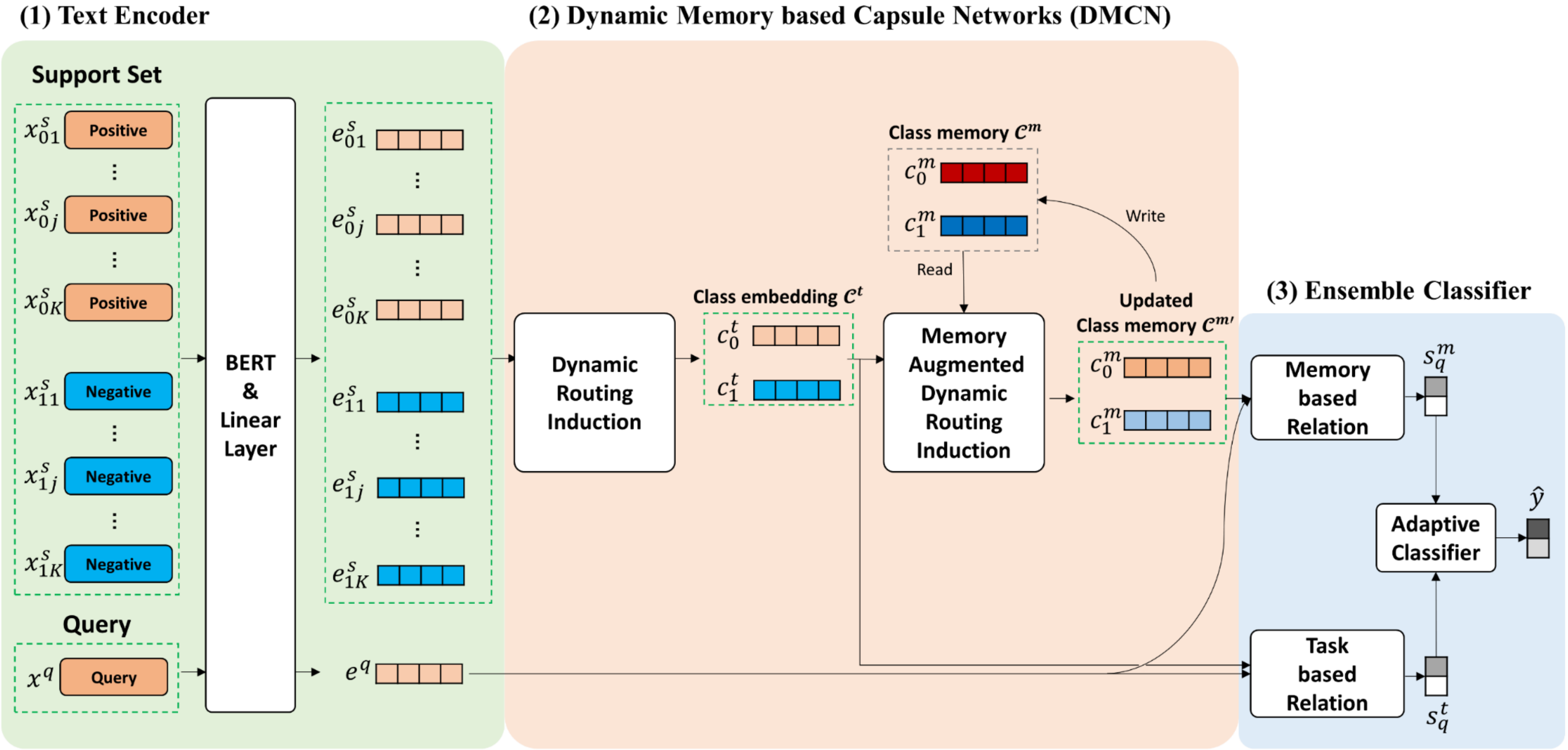

Figure 12 : Overall Architecture of the Dynamic Memory based Capsule Networks(DMCN)

Proposed Method :
Dynamic Memory based Capsule Networks(DMCN)

1. Text Encoder

- Use BERT-base and linear transformation for sentence encoding
- Average the BERT output layers instead using [CLS] token
- BERT is further pretrained on the meta-train set
- Only last two transformer layers of the BERT are fine-tuned

(1) Text Encoder

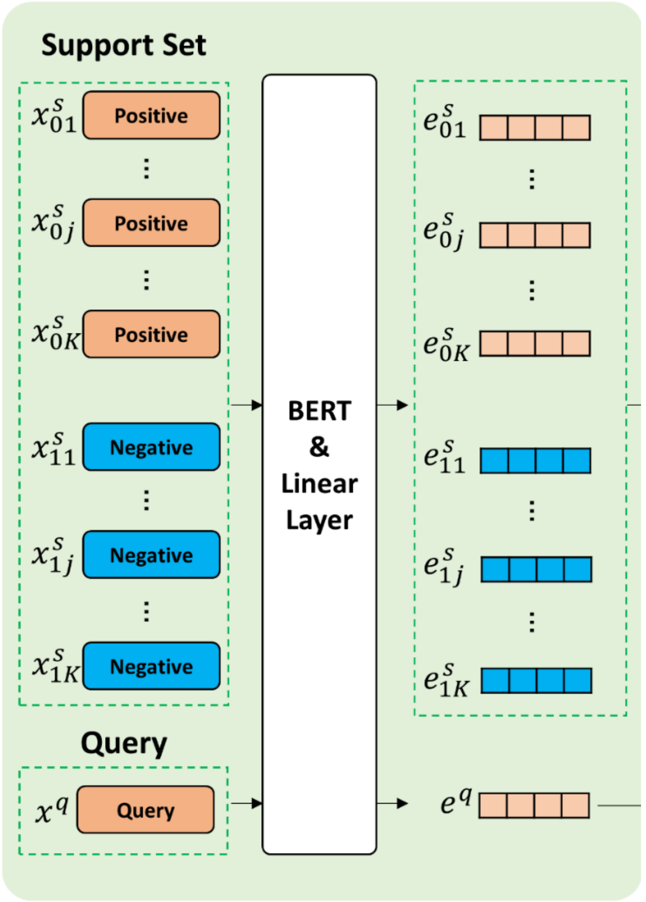

Figure 13 : Text Encoder

Proposed Method :
Dynamic Memory based Capsule Networks(DMCN)

2. DMCN

- Dynamic Routing Induction(Geng et al., 2019)
  - Extract the task-specific class embedding set C^t from sample vectors in the Support set

- Memory Augmented Dynamic Routing Induction
  - Update the task-invariant class embedding (memory) set C^m to $C^{m'}$ by integrating C^t into C^m

(2) Dynamic Memory based Capsule Networks (DMCN)

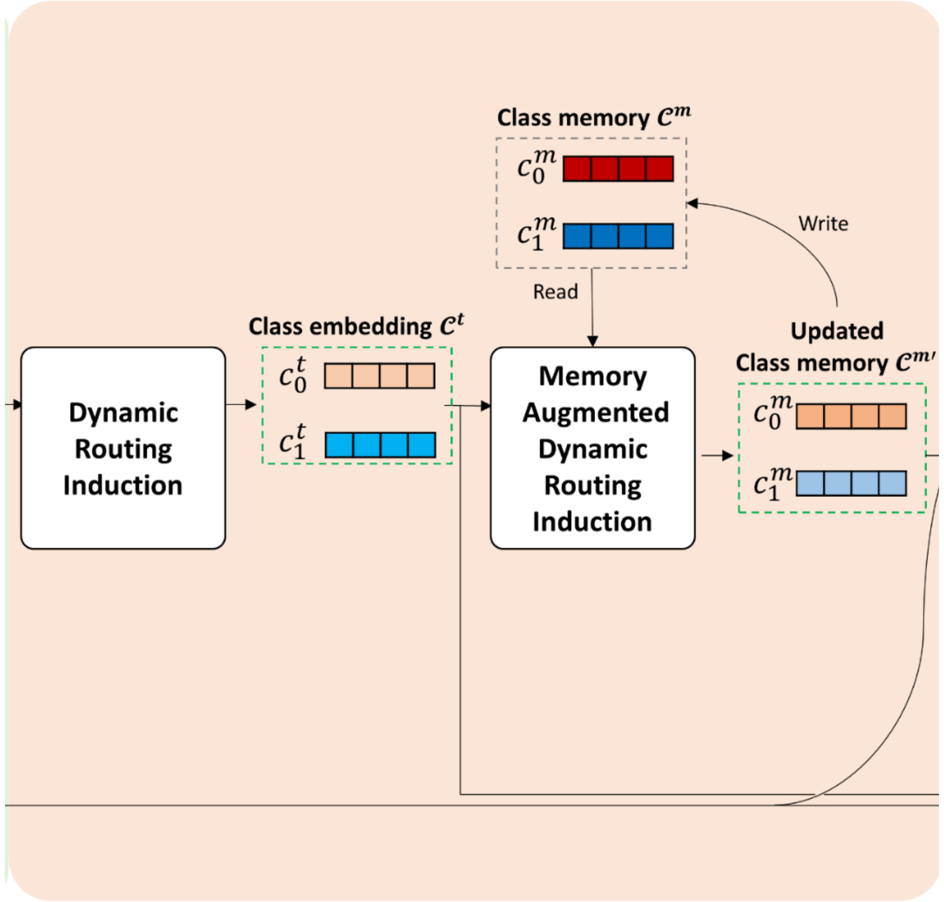

Figure 14 : DMCN

Proposed Method : Dynamic Memory based Capsule Networks(DMCN)

3. Ensemble Classifier

- It measures the correlation between each pair of query and class in two perspectives. One is task-invariant view, the other one is task-specific view.

- Memory based Relation Module
  - It takes the task-invariant embedding(memory) set C^m as input, and outputs the relation score s_q^m between the each element of C^m and the query q

- Task based Relation Module
  - Do the same process on the task-specific embedding set C^t

- Adaptive Classifier
  - Outputs the probability distribution that the query corresponds to each class by scaling the two scores according to the values of the two scores(s_q^m, s_q^t)

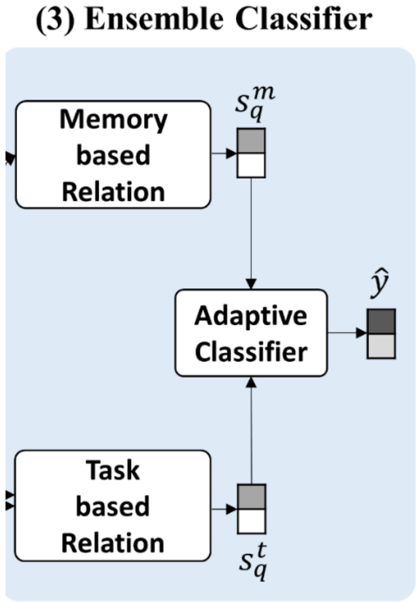

Figure 13 : Ensemble Classifier

Experiment:
Dataset & Evaluation Method

- **Dataset : Amazon Review Sentiment Classification(ARSC)**
  - Benchmark Dataset for Few-Shot Text Classification
  - 23 categories(domains) of products
  - Each category has 3 sentiment classification tasks
    - the thresholds of polarity are 2, 4, 5
    - e.g. the threshold = 4, then 4, 5 → pos / 1, 2, 3 → neg
  - Total 23 x 3 = 69 tasks
  - Meta-test : 4 domains(Books, DVD, Electronics and Kitchen), 12 tasks
  - Meta-train : 19 domains(remains of meta-test),57 tasks
  - 2-way 5-shot

- **Evaluation Method : 2-way 5-shot accuracy**

Results

1. **Compare the performance with exist methods**
   - DMCN is the proposed method
   - Knowledge Guided Metric Learning(KGML) uses external knowledge base
   - Except KGML, it achieves the best accuracy(87.47%)

Model	Mean Accuracy (%)
Matching Networks	65.73
Prototypical Networks	68.15
Relation Networks	86.09
ROBUSTTC-FSL	83.12
Induction Networks	85.63
Knowledge Guided Metric Learning*	87.93
MAML	78.33
P-MAML	86.65
DMCN (proposed)	**87.47**

* external knowledge database

Table 1 : 2-way 5-shot mean accuracy on ARSC dataset

Results

2. Ablation Study

- To verify the effect of further pretraining, further pretraining condition DMCN are compared with basic condition MACN
- To check the memory architecture's effect, the model using "**Only Memory-based Relation**" score are compared with the model using "**Only Task-based Relation**" score

Method	Mean Accuracy (%)	
	DMCN	NP-DMCN
Only Task-based Relation	87.40	85.27
Only Memory-based Relation	85.19	85.01
Memory and Task-based Relation	**87.47**	**86.73**

Table 2 : Performance comparison according to whether further pre-learning is performed and whether memory is applied

Results

2. **Ablation Study**
   - As the result, further pretraining group get a better results,
   - Task-based relation model gets more accuracy than memory-based relation model but memory-base relation model shows the robust performance regardless of further pretraining, and the proposed model gets better result than all the others.
   - This shows that the task-invariant embedding of the memory helps the model improve performance even when the quality of the embedding is poor.

Method	Mean Accuracy (%)	
	DMCN	NP-DMCN
Only Task-based Relation	87.40	85.27
Only Memory-based Relation	85.19	85.01
Memory and Task-based Relation	**87.47**	**86.73**

Table 2 : Performance comparison according to whether further pre-learning is performed and whether memory is applied

Conclusion

- **Can be applied to the case where the dataset is insufficient to get the prior knowledge**

- **Can efficiently train the model , by fine-tuning only a portion of the BERT**

- **By using both task-specific and task-invariant embedding, Achieve better performance than exists methods**