CS 502 - Project report: Extending Few-Shot Learning Benchmark with Relation Network Algorithm

Marija Zelic Elena Mrdja

Abstract

The scarcity of labeled samples available for some problems poses a great challenge to potential deep-learning solutions. This limitation is especially prevalent when dealing with biomedical data. Few-shot learning refers to the class of models that can learn the underlying pattern in the data from only a few training samples. Our approach expands upon the few-shot learning benchmark created by the Brbic Lab. Toward this goal, we incorporated Relation Network and evaluated its performance on the Swiss-Prot and Tabula Muris datasets.

1. Introduction

The rapid progress of deep learning has enabled automation in various fields, yet many models require abundant labeled data for accurate predictions, limiting their applicability in scenarios with scarce data, such as rare diseases or drug discovery. Few-shot learning addresses this by training models to classify new samples with minimal labeled examples. Inspired by humans' ability to learn from few examples, Meta-Learning episodically trains models on diverse tasks with limited samples per class. For biomedical datasets, overcoming data limitations is crucial. This project extends the Brbic Lab benchmark (Brb) by incorporating the Relation Network model (Sung et al., 2018) and evaluating its performance on benchmark datasets.

2. Methods and Datasets

2.1. Datasets

The goal of the benchmark is to evaluate few-shot learning algorithms of the different biomedicine-related tasks: protein function prediction based on Gene Ontology labels and cell-type annotation across different tissues. For that purpose, the Swiss-Prot and Tabula Muris datasets were selected, respectively.

Swiss-Prot The benchmark included the Swiss-Prot dataset, a labeled protein sequence database developed by the University of Geneva's Department of Medical Biochemistry. To align with benchmark requirements, Swiss-Prot amino acid sequences were transformed into Evolutionary Scale Modeling (ESM-2) embeddings, derived from a state-of-the-art protein model. These embeddings facilitate further fine-tuning of deep learning models (Lin et al., 2022). Gene Ontology (GO) labels were used for predictions, describing gene roles, locations, and functions across species. Labeling challenges arose from selecting unique labels for proteins with multiple valid labels, addressed by the Brbic Lab by choosing the most specific label.

Tabula Muris Tabula Muris is a single-cell transcriptomic dataset encompassing 105,960 cells that represent 124 cell types collected across 23 organs of mouse model organism. The features correspond to the gene expression profiles of cells. Out of the 23,341 genes, 2,866 genes with high standardized log dispersion given their mean were selected to reduce the dimensionality of the data (Cao et al., 2020).

2.2. Implemented algorithms

Multiple few-shot learning approaches were implemented in the benchmark, namely transfer learning algorithms such as Baseline and Baseline++ (Chen et al., 2019), and meta-learning algorithms, both Metric-based such as MatchingNet (Vinyals et al.) and ProtoNet (Snell et al., 2017), and Optimization-based such as Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017).

Transfer learning algorithms: Transfer learning utilizes knowledge acquired from solving one task to enhance performance on a related task by transferring learned features or representations. This accelerates learning in new domains and improves model efficiency. Both Baseline and Baseline++ adopt this approach, training the feature extractor and classifier with abundant pre-training data. During fine-tuning, the network parameters in the feature extractor are fixed, and a new classifier is trained with labeled examples in a few-shot setup. In contrast to the Baseline model, Baseline++ employs cosine distances between input features and class weight vectors to reduce intra-class variations. This strategy in Baseline++ demonstrates competitive performance with more sophisticated few-shot learning methods in certain scenarios (Chen et al., 2019).

Meta-Learning algorithms: Meta-learning algorithms

seek to enable models to "learn how to learn" by undergoing a two-phase process: meta-training and meta-testing. This process aims to efficiently leverage knowledge gained from previous tasks to address novel tasks with a shared underlying structure.

During both meta-training and meta-testing, datasets comprise disjoint support (training) and query (testing) sets with a shared label space. The few-shot problem is defined as K-shot N-way, where the learning process involves K instances per N distinct classes in the support set.

Meta-Learning is rooted in episodic training, where each iteration involves the creation of an episode. This entails the random selection of N classes, each with K labeled samples, to form the support set. Additionally, a subset of the remaining samples from these N classes is chosen to constitute the query set, with no requirement for equal numbers of support and query examples.

Meta-Learning algorithms generally adhere to either the *metric-based approach* or the *optimization-based approach*. In the metric-based paradigm, the primary emphasis is on acquiring a suitable similarity metric or embedding space. This facilitates effective generalization and adaptation to new tasks by evaluating the distance or similarity between instances in the feature space. Examples of methods within this category include MatchingNet, ProtoNet, and RelationNet. In contrast, the optimization-based approach concentrates on learning a favorable initialization, enabling the network to be readily fine-tuned for a target task with minimal gradient steps. Examples of such methods encompass MAML, Reptile, LEO, and others.

For MatchingNet (Vinyals et al.) and ProtoNet (Snell et al., 2017), both algorithms evaluate examples in the query set by measuring the distance between the query feature and the support feature for each class. MatchingNet uses cosine distance, computing the average across classes, while ProtoNet employs Euclidean distance, comparing query features to the class mean of support features, referred to as the prototype.

The MAML (Finn et al., 2017) optimizes parameters for specific tasks by fine-tuning them with support sets through a limited number of gradient updates. In the meta-training phase, the loss of the fine-tuned model on the query set adjusts the parameters of the pre-trained model.

As a metric-based methodology, RelationNet (Sung et al., 2018) aligns with MatchingNet and ProtoNet in its approach. However, a distinctive feature is its methodology, where it replaces the conventional distance assessment between query and support set samples with a trainable module known as a relation module. After processing samples from both the support and query set through the shared embedding module, two distinct approaches are employed based on whether the problem involves a one-shot or K-shot learning setup. In *one-shot* learning, embeddings are combined through the concatenation of each query set sample with every support set sample. For K*-shot* learning, an extra step involves element-wise summation over the embedding module outputs of all support set samples per class to generate the class-specific feature map.

The combined feature maps of the support and query sets are fed into the relation module, producing a scalar in the range of 0 to 1 representing the similarity between each pair of support and query samples. This scalar termed the relation score, is generated for each pair, resulting in a total of N relation scores for a single query, irrespective of the one-shot or K-shot learning setting.

The architecture of the RelationNet model is founded on the fully connected network backbone by stacking linear blocks that are built out of the elements displayed in the 1, to mimic convolutional architectures in the original paper, adapted to our non-image datasets.

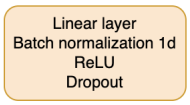

Figure 1. Linear block architecture for Metric-based methods.

3. Experiments

To address the performance of each of the models in the benchmark, several different experiments were explored. Their results provide valuable insights into the models' behavior in diverse settings of hyperparameters (network architectures in terms of the number of nodes and layers, learning rates, regularization parameters, number of epochs, number of episodes, etc.). The main focus was set on testing outcomes of four different few-shot learning problems: 3-way 1-shot, 3-way 5-shot, 5-way 1-shot, and 5-way 5-shot, all of them with 15 samples in the query set. Both datasets are split into the training, validation, and testing subsets, with each of them containing a different number of classes. The RelationNet model was obtained by performing a grid search over the before-mentioned hyperparameter space, to identify the ones that out-turned the best mean and standard deviation accuracy metrics on the validation subset after averaging over 1000 episodes. The optimizer used for all the models was Adam. Once the models were trained, mean testing accuracies over 5 different seeds were averaged to give the final performance of the algorithms, reported in the tables below.

3.1. Swiss-Prot

The Swiss-Prot dataset is split into a training subset with 59 classes, a validation subset with 15 classes, and a testing subset with 9 classes. Tables 1 and 2 depict mean accuracy

and standard deviation, as the simplest metrics for evaluation of the models in the benchmark.

Table 1. 3-way classification mean accuracy and standard deviation comparison for 1-shot and 5-shot setting over different models for Swiss-Prot dataset

MODELS	3-WAY 1-SHOT	3-WAY 5-SHOT
PROTONET	62.0 ± 1.3	71.6 ± 0.9
MATCHINGNET	65.3 ± 0.8	72.8 ± 0.7
RELATIONNET	63.0 ± 0.7	58.5 ± 1.3
MAML	64.7 ± 1.4	65.4 ± 1.1
BASELINE	64.3 ± 1.0	75.5 ± 1.0
BASELINE++	63.5 ± 1.1	69.8 ± 1.0

Table 2. 5-way classification mean accuracy and standard deviation comparison for 1-shot and 5-shot setting over different models for Swiss-Prot dataset

MODELS	5-WAY 1-SHOT	5-WAY 5-SHOT
PROTONET	54.6 ± 1.3	63.4 ± 1.1
MATCHINGNET	53.2 ± 1.2	61.7 ± 1.0
RELATIONNET	46.1 ± 1.3	44.0 ± 0.9
MAML	49.6 ± 1.1	45.1 ± 1.1
BASELINE	53.6 ± 1.0	63.0 ± 1.0
BASELINE++	52.1 ± 1.2	59.2 ± 1.0

In the context of **3-way 1-shot** classification, it can be deduced that all Meta-learning-based algorithms as well as Transfer-learning based algorithms exhibit comparable mean accuracy, with MatchingNet emerging as the front runner, followed by MAML with negligibly smaller mean accuracy, but higher standard deviation. In the domain of **3-way 5-shot** classification, an overall elevation in mean accuracies is noted, aligning with expectations due to the increased number of samples per class during training episodes. It is noteworthy that both RelationNet and MAML exhibit significant declines in accuracy, phenomena that will be further investigated in subsequent analysis. On the other hand, the Baseline model demonstrates superior performance.
Regarding the **5-way 1-shot** classification, it is apparent that mean accuracy values are lower in comparison to the 3-way 1-shot classification scenario. This is anticipated as the model deals with an increased number of classes while having only one sample per class. Notably, among Meta-learning-based algorithms, there exists a significant performance disparity. MatchingNet and ProtoNet exhibit superior performance in this specific scenario, while MAML and RelationNet show comparatively lower efficacy. This suggests that the latter models may be more susceptible to overfitting across a larger and potentially more diverse feature space. In the case of **5-way 5-shot**, following the previously observed pattern, the mean accuracy of all models except RelationNet and MAML is higher than their 1-shot counterparts.

3.2. Tabula Muris

The Tabula Muris dataset is split into a training subset with 59 classes, a validation subset with 47 classes, and a testing subset with 37 classes.

Table 3. 3-way classification comparison for 1-shot and 5-shot setting over different models for Tabula Muris dataset

MODELS	3-WAY 1-SHOT	3-WAY 5-SHOT
PROTONET	86.8 ± 0.7	93.3 ± 0.5
MATCHINGNET	81.3 ± 1.1	89.7 ± 0.9
RELATIONNET	85.8 ± 1.0	88.4 ± 1.0
MAML	81.1 ± 1.1	89.3 ± 0.8
BASELINE	87.1 ± 0.8	93.8 ± 1.1
BASELINE++	78.1 ± 1.2	89.8 ± 1.0

Table 4. 5-way classification comparison for 1-shot and 5-shot setting over different models for Tabula Muris dataset

MODELS	5-WAY 1-SHOT	5-WAY 5-SHOT
PROTONET	81.8 ± 1.0	89.4 ± 0.8
MATCHINGNET	73.6 ± 1.1	83.2 ± 0.7
RELATIONNET	68.9 ± 1.0	78.6 ± 0.9
MAML	67.9 ± 1.2	86.2 ± 1.1
BASELINE	78.6 ± 0.9	88.6 ± 1.0
BASELINE++	66.9 ± 1.0	81.2 ± 0.7

The results of the classification on the Tabula Muris dataset show a significant improvement in performance for all of the models compared to the results observed for Swiss-Prot. In both the case of **3-way 1-shot** and **3-way 5-shot** classification (table 3), the Baseline model outperforms its counterparts, followed closely by ProtoNet with slightly lower mean accuracy but more favorable standard deviation. Compared to the 3-way 1-shot, all models show a notable increase in mean accuracy for the 3-way 5-shot classification. Even though RelationNet had the third-best mean accuracy among the models during 3-way 1-shot learning, in the 5-shot case it performed worse than the other models.
Observing the case of **5-way 1-shot** classification (table 4), we notice a great contrast in the performance of different models. Baseline and ProtoNet yield competitive accuracies, compared to MAML, MatchingNet, and Baseline++ which drastically underperform. In **5-way 5-shot** learning, ProtoNet, and Baseline again demonstrate superior performance, with the highest accuracies while RelationNet significantly tails behind.

3.3. Ablation study

Besides presenting the models with the best accuracy, we performed the ablation study on each of the meta-learning-based models to evaluate their sensitivity to the changes in the hyperparameters. Conducting such

evaluations is imperative, as the resilience of the models to variations in parameters significantly influences their overall performance.

For any number of classes in the classification problem, **ProtoNet** models with multiple linear blocks displayed a propensity for overfitting on the training set. This issue can be mitigated by introducing a small regularization parameter to the optimizer or simplifying the network architecture. Both approaches yielded comparable results for a smaller number of classes, while discrepancies between validation and training accuracy persisted for larger class sets. Regarding training epochs, ProtoNet achieved competitive results with 30% fewer iterations than the optimal case, and similar outcomes were observed for a smaller number of training episodes. In 3-way classification, ProtoNet demonstrated insensitivity to a tenfold increase in the learning rate, whereas for 5-way classification, a larger learning rate led to significant overfitting. Despite addressing overfitting and showcasing stability in response to parameter changes, ProtoNet emerged as a robust model in our analysis. In assessing the **MatchingNet** model, we adopted a methodology consistent with that employed for ProtoNet. Manipulating hyperparameters in the context of 3-way and 5-way classification produced analogous findings to those observed with the ProtoNet model. The model avoids overfitting and yields better performance for the smaller number of layers, which remains stable or slightly worsened when decreasing the number of epochs or the number of episodes. The **MAML** model exhibited stability with smaller class numbers but demonstrated severe overfitting for larger N-way classifications, particularly with the proposed hyperparameters, such as inner learning rate and number of tasks. This led to superior training dataset performance but underwhelming results on the testing set.

Before delving into a more comprehensive examination of the evaluation conducted on the **RelationNet** model, it is imperative to acknowledge its atypical performance. As discerned from the aforementioned tables 1 and 2, the augmentation of shot numbers in the context of 3-way and 5-way classification did not yield a notable increase in the mean accuracy for this model and Swiss-Prot dataset. Even more surprising is the observation that its accuracy diminishes with an increasing number of available samples. We employed a range of methodologies to probe into the determinants influencing this behavior, including modifications to the model architecture concerning variations in the number of nodes and layers within linear blocks, adjustments to dropout rates, manipulation of learning rates, regularization parameters, and number of epochs and episodes, among others. These modifications did not lead to a significant accuracy enhancement, but the investigation did show an additional finding: this architecture of the RelationNet model exhibits pronounced inconsistency even with minor variations in its parameters. The obtained mean accuracy and standard deviation for RelationNet are competitive to its rival models in the case of a 1-shot learning problem, but tail behind them significantly in a 5-shot learning one. This implies that this architecture of RelationNet might be failing to adjust to sequence data and aggregate features efficiently in the context of our benchmark since it is reported that architecture based on convolutional blocks performs well on image-based data. Therefore, the next step in the research would be to assess the performance of such an architecture, similar is addressed in the (Cao et al., 2020), on the Swiss-Prot dataset. We speculated that the reason for this could be the way the author originally implemented the aggregation of the sample features in K-shot learning when $K > 1$. Changing this method from summing the sample features to taking their mean, even though it includes only additional scaling, did result in overall higher training accuracy, but failed to produce a substantial increase on the testing dataset. What is also interesting to address is the performance of this RelationNet architecture on the Tabula Muris dataset, which acknowledges an increase in the mean accuracy across 1-shot and 5-shot problems and also produces comparable results to the ones provided in the (Cao et al., 2020). This may suggest that the Tabula Muris dataset, with twice as many features compared to the Swiss-Prot dataset, provides a more discernible feature space for this particular architecture, implying that the data in the former dataset is more conducive to learning within this framework. Nevertheless, we contend that a systematic examination of other existing RelationNet architecture adaptations for this kind of dataset should be conducted initially to discern the fundamental cause of the observed suboptimal accuracy.

4. Conclusion

The project aimed to make a comprehensive and complete evaluation of the performance of different few-shot learning models, by expanding and analyzing the previously implemented benchmark. Through the addition of RelationNet, we were able to get a better understanding of its performance compared to other state-of-the-art methods. While in some cases RelationNet yielded competitive results to other models, it failed to do so in others, most notably in the 5-shot setting, on the Swiss-Prot dataset. A potential reason for this inconsistency could be the replacement of its convolutional structure with fully-connected layers, a decision we made to account for the difference in data from the original implementation on image-based data. Therefore, as we assessed this RelationNet architecture, we also opened the door for future research on how different approaches to this model behave.

5. Others

Table 5. Best parameters for RelationNet model architecture obtained by grid search hyperparameter tuning on Swiss-Prot dataset and 3-way problem

ARCHITECTURE	3-WAY 1-SHOT	3-WAY 5-SHOT
EMBEDDINGS MODULE	[512, 512]	[512]
RELATION MODULE	[256, 256]	[256]
LINEAR LAYER SIZE	128	128
LEARNING RATE	0.001	0.001
EPOCHS	60	60

Table 6. Best parameters for RelationNet model architecture obtained by grid search hyperparameter tuning on Swiss-Prot dataset and 5-way problem

ARCHITECTURE	5-WAY 1-SHOT	5-WAY 5-SHOT
EMBEDDINGS MODULE	[128]	[512, 512]
RELATION MODULE	[256, 64]	[256, 256]
LINEAR LAYER SIZE	128	128
LEARNING RATE	0.001	0.001
EPOCHS	60	40

Table 7. Best parameters for RelationNet model architecture obtained by grid search hyperparameter tuning on Tabula Muris dataset and 3-way problem

ARCHITECTURE	3-WAY 1-SHOT	3-WAY 5-SHOT
EMBEDDINGS MODULE	[64, 64]	[64]
RELATION MODULE	256	[256, 256]
LINEAR LAYER SIZE	128	128
LEARNING RATE	0.001	0.01
EPOCHS	60	60

Table 8. Best parameters for RelationNet model architecture obtained by grid search hyperparameter tuning on Tabula Muris dataset and 5-way problem

ARCHITECTURE	5-WAY 1-SHOT	5-WAY 5-SHOT
EMBEDDINGS MODULE	[64, 64]	[64, 64]
RELATION MODULE	256	256
LINEAR LAYER SIZE	128	64
LEARNING RATE	0.01	0.001
EPOCHS	60	40

References

Brbiclab. https://brbiclab.epfl.ch/.

Cao, K., Brbic, M., and Leskovec, J. Concept learners for generalizable few-shot learning. *CoRR*, abs/2007.07375, 2020. URL https://arxiv.org/abs/2007.07375.

Chen, W., Liu, Y., Kira, Z., Wang, Y. F., and Huang, J. A closer look at few-shot classification. *CoRR*, abs/1904.04232, 2019. URL http://arxiv.org/abs/1904.04232.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017. URL http://arxiv.org/abs/1703.03400.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017. URL http://arxiv.org/abs/1703.05175.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning, 2018.

Vinyals, O., Deepmind, G., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. *Matching Networks for One Shot Learning*. URL https://arxiv.org/pdf/1606.04080.pdf.