Wirtschaftsinformatik

und Maschinelles Lernen

Stiftung Universitat Hildesheim

Marienburger Platz 22

31141 Hildesheim

Prof. Dr. Dr. Lars Schmidt-Thieme

Lukas Brinkmeyer

Master – Seminar Data Analytics II

Prototypical Networks for Few-shot Learning

Summer Semester 2020

Ajith Gumudavelly

305643, gumudavelly@uni-hildesheim.de

Abstract

The main objective of the authors of this paper is to propose a new prototpical networks for the problems which are faced by few-shot classification. So the new classifier should be able to generalize to the new classes which are not seen in the training set for every few number of samples. Basically prototpical networks has ability to learn metric space where the classifiscation can be done by calculating the distances to the prototype representations. Plenty of researches are done on this idea with excellent results, though authors wanted to present the idea can be done better by adding few clear analysis. The data which is consider here to show the results is: CU-Birds dataset.

Table of Contents

Introduction

The authors of the paper which I am going to be reviewing are Jake Snell: Ph.d student at University of Toronto, Kevin Swersky: Ph.d student at University of Toronto and research assisstant at Google Brain under the supervision of Richard S. Zemel: Professor at University of Toronto, Department of Computer Sciencee and also the Director of research at the Vector Institute for Artificial Intelligence.

Before this paper, there is abundant of research done on few-shot learning while two of such produced good results. This paper [1] proposed a new method called matching networks,this used the method attention mechanism on set of examples to predict classes for the points which are not labelled. The main technique here is to create a similar few shot task by mimicing which are called as episodes while training. These episodes creates an environment which are good for the training problem and provides a clear route for imporving generalization.

The second one [2] is an added result to the before paper. Here we use episodes but by meta-learning approach. This used LSTM such that it goes with training the custom model for every episode instead of going with one model on various episodes. These approcahes provided good results but authors mainly focused on a problem called overfitting[3]. This is caused because of repetation of training model on the new data which becomes hard for few shot classification to give space to the new class which was not present in the training data. The authors not only performed few shot but also zero shot learning. They drew connections to the matching networks in the one-shot setting and examine the distance function which is used in the model.

The authors related the prototypical networks to the clustering[4] because to prove that using of class means as a prototype when distances are calculated with a Bregman divergence, such as squared Euclidean distance. Here the choosing of distance calculator matters and prototypical networks are more approachable for few-shot and zero-shot learning because of its efficiency and simplicity than the meta learning algorithms.

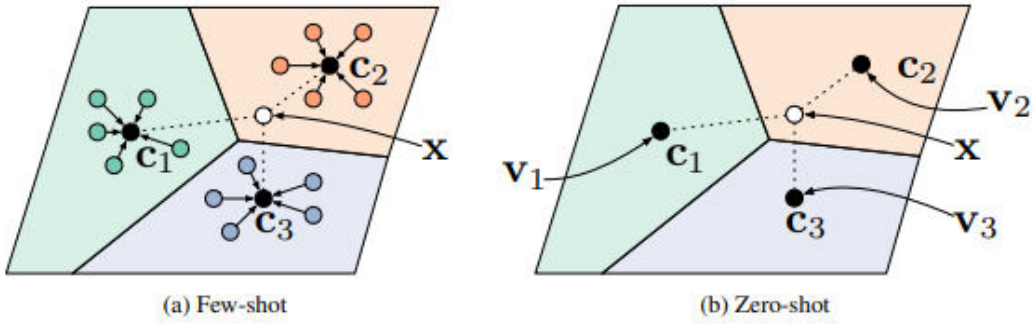

Figure 1: *Prototypical networks in the few-shot and zero-shot scenarios. Left: Few-shot prototypes ck are computed as the mean of embedded support examples for each class. Right: Zero-shot prototypes ck are produced by embedding class meta-data vk. In either case, embedded query points are classified via a softmax over distances to class prototypes:* $p\varphi(y = k|x) \propto exp(-d(f\varphi(x), ck))$*.[14]*

Related Work

The are many works done by many authors on this work. Few of them which makes sense to our approach are

One is Non - Linear extension of NCA[5] because the authors used Neural Networks and uses Euclidean distance for calculating the distance. A minute differnce is that authors

formed softmax directly over classes unlike the individual points which are calculated from each and every class to the distances for the representation and will be easy for predictions.

The second one is to nearest class mean approach[6]. Here they used use mean of examples. This is developed to combine the new classes to classifier without reinforcement and designed to deal with large set of examples. But in this approach the authord used Neural Networks to non - linearly embed points and combine this with episodic training to handle the few shot scenario. This[6] appraoch also go for non-linear classification but by accepting classes to have many more prototypes. These prototypes are available while pre-processing when using k-means on the input space and goes with multi-modal variant of the linear embedding whereas prototypical networks in our approach uses a non-linear embedding without any pre-processing and authors mainly used Bregman divergences unlike distance used in this paper[6].

Third one is meta-learning approach from this paper[2] and LSTM is important dynamics here. From any given episode here the LSTM is trained itself to train a model to perform good on query points. We can consider both prototypical Networks and Matching Networks under meta-learning which produces simple classifiers from the new form of training episodes. But in the few shot the data is so less that simple inductive bias seems to work very well and there is no need of learning a custom embedding for every episode.

There is another approach which is related to this one like neural statistician[7] which is from the generative modeling literature.

Summary

Before this approach many researchers from different universities tried solving the problem of overfitting with different approcahes and using different distance metric but in this paper authors mainly concentrated on using the relevant and effecient distance metric called squared Euclidean distance and this works effeciently because of cosine distance which is used previous method is not being an Bregman Divergence.

Zero-shot Learning

The other part the authors implemented is zero-shot learning which is different from few-short which in case they considered a class of meta-data vectors for every class. These are learned from the raw text:

$$-\| f_{\phi}(x) - c_k \|^2 \quad = \quad - f_{\phi}(x)^T f_{\phi}(x) + 2c^T{}_k f_{\phi}(x) - c_k{}^T c_k$$

Model	Dist.	Fine Tune	5-way Acc. 1-shot	5-way Acc. 5-shot	20-way Acc. 1-shot	20-way Acc. 5-shot
MATCHING NETWORKS [29]	Cosine	N	98.1%	98.9%	93.8%	98.5%
MATCHING NETWORKS [29]	Cosine	Y	97.9%	98.7%	93.5%	98.7%
NEURAL STATISTICIAN [6]	-	N	98.1%	99.5%	93.2%	98.1%
PROTOTYPICAL NETWORKS (OURS)	Euclid.	N	**98.8%**	**99.7%**	**96.0%**	**98.9%**

Table1: Few - shot classification accuracies on Omniglot [14]

To deal with zero-shot prototypical networks should be modified and that will be a direct approach so they define it is $c_k = g_{\phi}(v_k)$, which is a separate embedding for the meta-data vector.

Overall, the authors performed experiements on Omniglot[8] with the splits proposed for few-shot learning and on Caltech UCSD bird dataset(CUB-200 2011)[9] for zero-shot.

Experiments

For Few shot classification authors used Omniglot dataset which has 1623 handwritten characters from 50 alphabets and these are recycled to grayscale 28 * 28. They also used 1200 characters plus 4800 classes in total for the test. The embedding architecture is of four blocks and each is 64-filter 3 * 3 convolution. The initial learning rate is 10^{-3} and it is cut half for every 2000 episdoes. These prototypical networks are trained using Euclidean distance with 60 classes and 5 query points per class. It is necessary that the matching the training-shot with test-shot is useful. At last the authors calculated classification accuracy for model over 1000 randomly generated episodes from test set and resutls are in Table 1.

The another dataset authors used for Few shot classification is miniImageNet dataset which was proposed from this paper[1] and derived from ILSVRC-12 dataset[10]. This dataset consists of 60 thousand color images of 84 * 84 size which is divided into 100 classes with 600 examples. The splits here used are extracted from [2].

The table2 shows the few shot classification accuracies on miniImageNet data and all the results are averaged considering 600 test episdoes and reported with 95% confidence intervals.

Model	Dist.	Fine Tune	5-way Acc. 1-shot	5-way Acc. 5-shot
BASELINE NEAREST NEIGHBORS*	Cosine	N	28.86 ± 0.54%	49.79 ± 0.79%
MATCHING NETWORKS [29]*	Cosine	N	43.40 ± 0.78%	51.09 ± 0.71%
MATCHING NETWORKS FCE [29]*	Cosine	N	43.56 ± 0.84%	55.31 ± 0.73%
META-LEARNER LSTM [22]*	-	N	43.44 ± 0.77%	60.60 ± 0.71%
PROTOTYPICAL NETWORKS (OURS)	Euclid.	N	**49.42 ± 0.78%**	**68.20 ± 0.66%**

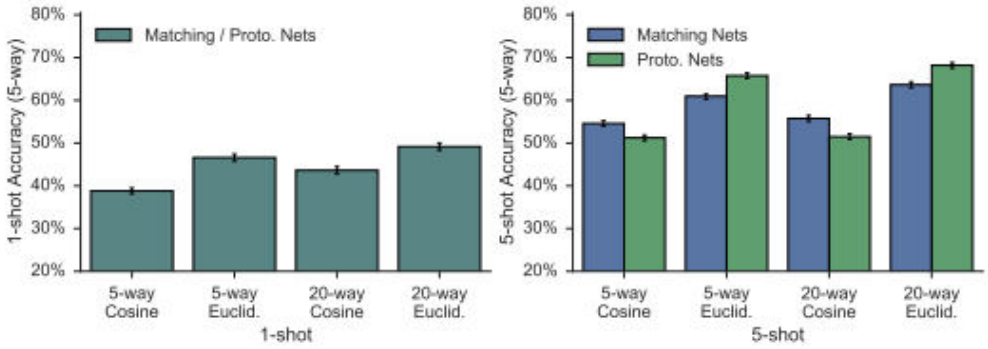

Table2 *Comparison showing the effect of distance metric and number of classes per training episode on 5-way classification accuracy for both matching and prototypical networks on miniImageNet. [14]*

For Zero-shot classification, the authors used Caltech-CUB-200 Birds (CUB) dataset 200-2011[9] which has 11,788 images of different bird species of 200. Here the split is into 100 training, 50 validation and 50 test. The Table3 shows the results are achieved with good margin comparatively to other methods like [11][12][13].

Model	Image Features	50-way Acc. 0-shot
ALE [1]	Fisher	26.9%
SJE [2]	AlexNet	40.3%
SAMPLE CLUSTERING [17]	AlexNet	44.3%
SJE [2]	GoogLeNet	50.1%
DS-SJE [23]	GoogLeNet	50.4%
DA-SJE [23]	GoogLeNet	50.9%
PROTO. NETS (OURS)	GoogLeNet	**54.6%**

Table3: Zero-shot classification accuracies on CUB-200. [14]

Discussion

Discussion (i)

The motivation of this paper is solving overfitting problem which is caused because of repeatation of using of same model over new data. To solve that authors, tried with a new and effecient approach of episodic approach. The research question which raised while introducing the paper are well formulated.

The related work which are mentioned in the related section are mostly with similar approaches but there are minute differences like usage of distances, perfoming operations on individual spaces and also using different algorthms. Overall the pipeline is similar but the usecase are changing in each and every appraoch and I didn't feel like authors left any space for other approaches to add.

When we look at the method, authors were stick to their motivation which is solving the overfitting problem and considered relevant distance metrics to solve the problem. The research questions are well defined.

At the end if you look at the results for few shot classification accuracies on ImagenNet the results are averaged over 600 test episdoes and overall they are reported with 95% confidence intervals.

For CUB-200 zero shot classification accuracies on CUB-200 dataset the model achieved 54.6% which is good than remaining models whereas few shot classification accuracies on Omniglot achieved 98.9% .

To conclude, as a reader I would like to point out that authors did a development for the model by achieving results with little increase in accuracy.

Discussion (ii)

The paper is well crafted and easily readable for an average english speaker as well but some words and sentences are quite tricky to undertstand while describing about the approaches like while describing about the other similar approaches.

Formation is good but after references there are few additional results which I think will good if the authors added before it with a separate heading.

Coming to citations there are few mentioned because those approaches are similar to what authord performed here.

Titles, description and details for figures are correct and explained accordingly whenever there is a need for describing about it.

All the references, citations and link to other related information are also mentioned correctly. But as I can't read all the information from all the papers mentioned it is quite difficult for reader to check the source is legitimate or not considering time. For few which I checked seems to be relative. There are 31 references from different authors.

Conclusion

The main idea which authors proposed is that they represent each and every class by the mean of the examples which is a representation learned by the neural network and the

approach is called prototypical networks for few-shot learning. Those networks are trained using episodic training. When you look at the approaches above in the related work section this is simpler comparatively and produces state-of-the-art.

The important thing the authors considered here is the distance while other approaches ignored it's importance and also by modifying the episodic learning procedure. They also went with the approach of prototypical neworks for zero-shot setting as well and also achieved results on the CUB-200 dataset. Coming to the future work as said before the authors gave importance to the distance metric and in future it is Bregman distance is used. They also conducted few explorations like learning a variance per dimension for every class. At the end the authors achieved the efficiency of prototypical networks which makes a one of the best approach for few shot learning.

References

[1] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In Advances in Neural Information Processing Systems, pages 3630–3638, 2016.

[2] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. International Conference on Learning Representations, 2017.

[3] Overfitting [Online] Available From: https://en.wikipedia.org/wiki/Overfitting [Accessed on August 14 2020]

[4] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. Journal of machine learning research, 6(Oct):1705–1749, 2005.

[5] Ruslan Salakhutdinov and Geoffrey E. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In AISTATS, pages 412–419, 2007

[6] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(11):2624–2637, 2013.

[7] Harrison Edwards and Amos Storkey. Towards a neural statistician. International Conference on Learning Representations, 2017.

[8] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. In CogSci, 2011.

[9] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015.

[11] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attributebased classification. In Computer Vision and Pattern Recognition, pages 819–826, 2013.

[12] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In Computer Vision and Pattern Recognition, pages 2927–2936, 2015.

[13] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning deep representations of fine-grained visual descriptions. arXiv preprint arXiv:1605.05395, 2016.

[14] Jake Snell,Kevin Swersky,Richard S. Zemel: Prototypical Networks for Few-shot Learning