Few-shot Learning with Novel Metrics

Patrik Gergely, Ana Sofia Uzsoy, Stuart Burrell

UNIVERSITY OF CAMBRIDGE

Summary

Motivation

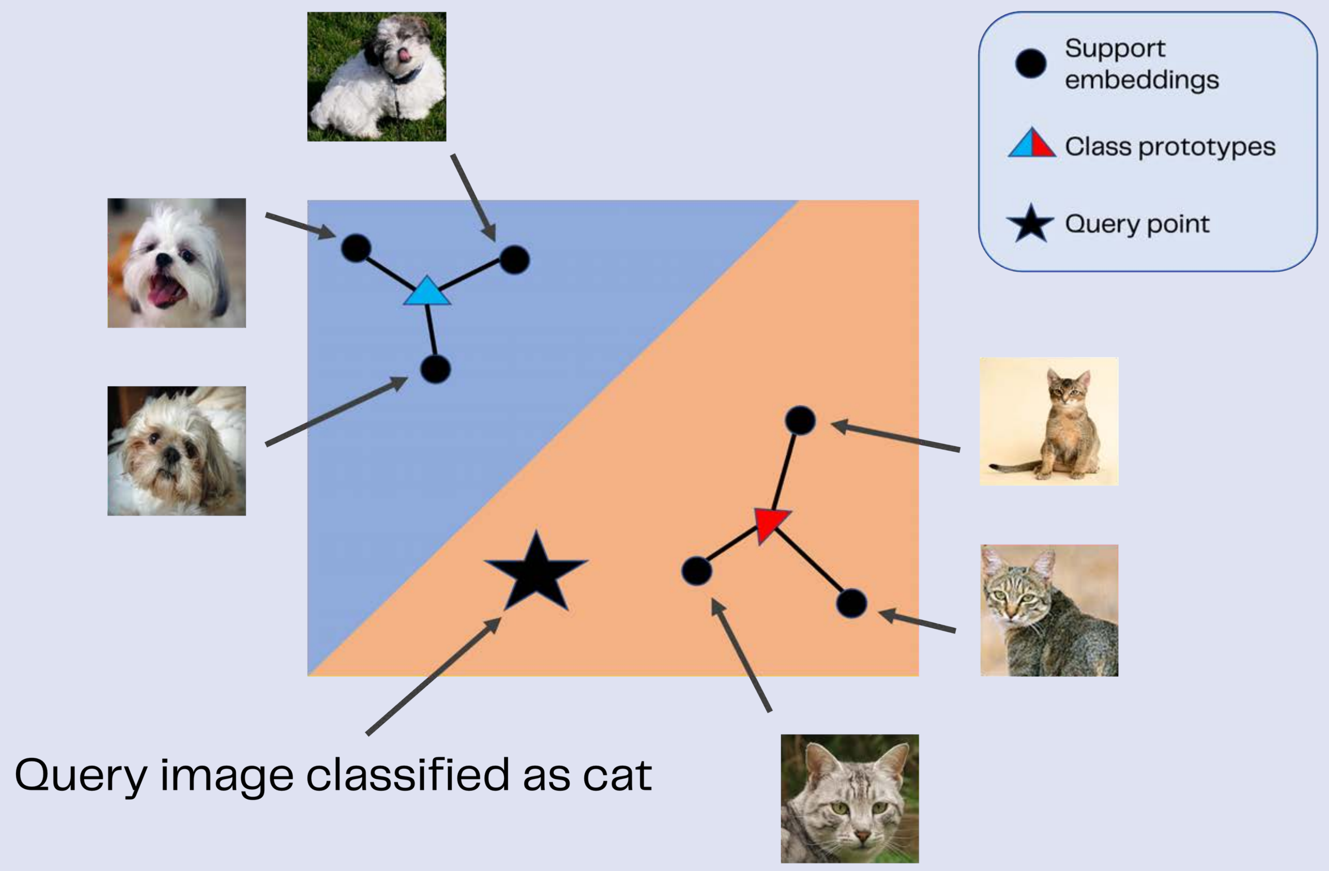
Traditional methods must be re-trained to learn new classes, which is computationally expensive.

The Task

Classify unseen classes with a pre-trained model from a handful of examples.

A Solution

Prototypical Networks, Snell et al. (2017)



Future Directions

Learnable Metrics: model learns the best distance metric for the task

Doc2Vec Pre-training: use pretrained models to improve NLP performance

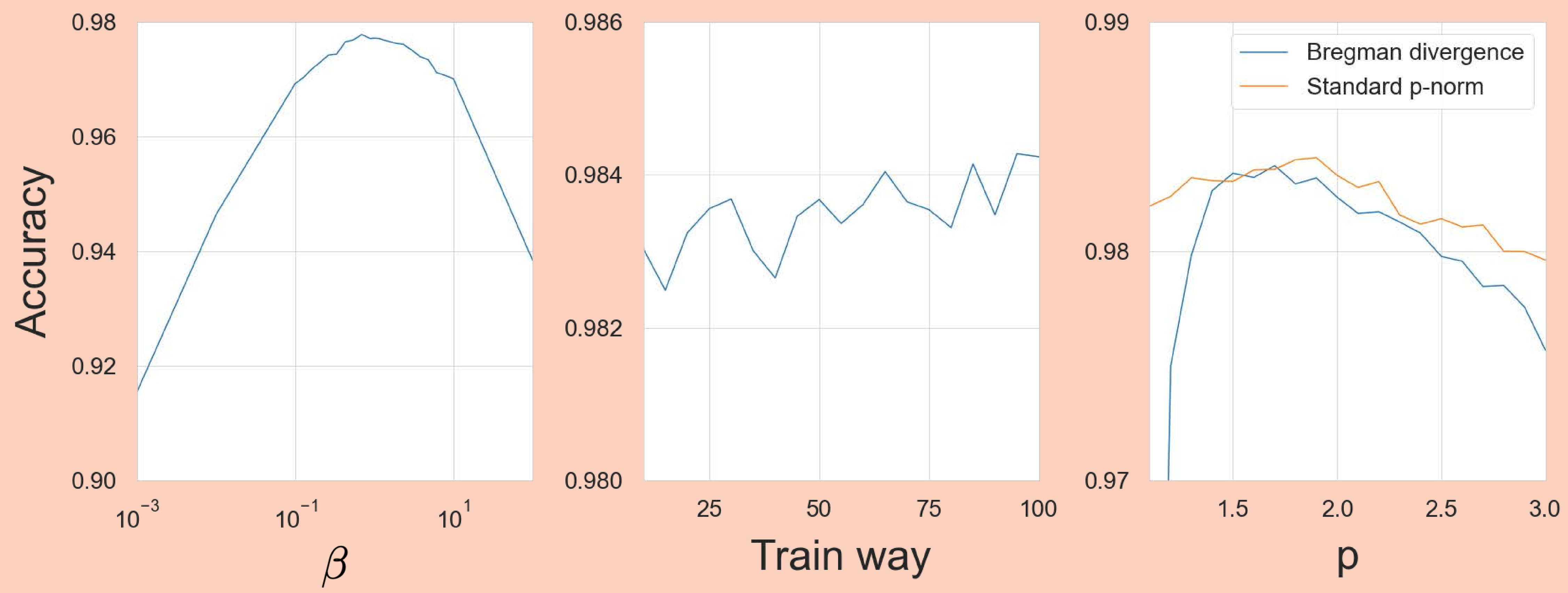
NLP-Focused architectures: use recurrent or LSTM layers to improve NLP performance

Results

Table 1: Percentage accuracy of different metrics trained on three different datasets. Testing was done 5-way with either 1-shot or 5-shot.

	Omniglot		Mini-imagenet		Reuters	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Snell's Euclidean	98.8	99.7	49.42	68.20	–	–
Squared Euclidean	98.32	99.51	48.35	65.82	24.01	27.99
KL Divergence	67.46	78.13	38.86	44.26	22.18	34.47
Generalized I-div.	74.08	87.45	28.83	43.95	22.31	34.98
Cosine Similarity	72.69	83.30	38.07	46.40	24.33	25.42
Cosine with Softmax	82.24	88.45	39.88	54.51	21.57	26.31

Hypothesis: Euclidean distance outperforms cosine similarity as it is a Bregman divergence.
Finding: KL divergence and Generalized I-divergence are two Bregman divergences which are outperformed by cosine similarity.



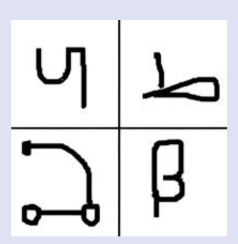
(a) Model accuracy vs. β for Mahalonobis distance.

(b) Model accuracy vs. train way using standard distance.

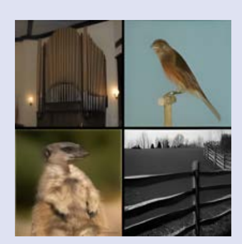
(c) Model accuracy vs. p for Bregman divergence & Standard p-norm.

Default values fom Snell et al. (2017) are 5-shot 60-way train, 20-way test with 15 query points for each.

Data



Omniglot



Mini-imagenet

Reuters

Glossary

Support: Example points for prototype
Shot: Number of support points
Way: Number of classes
Prototype: Embedded class representative
Query: Point to be classified

Mathematical Background

Pseudo-metrics

Squared Euclidean $\sum_{j=1}^{d}(x_j - y_j)^2$

Generalised I-divergence $\sum_{j=1}^{d} x_j \log(\frac{x_j}{y_j}) - \sum_{j=1}^{d}(x_j - y_j)$

KL Divergence $\sum_{j=1}^{d} x_j \log_2(\frac{x_j}{y_j})$

Cosine Similarity $\frac{x \cdot y}{|x||y|}$

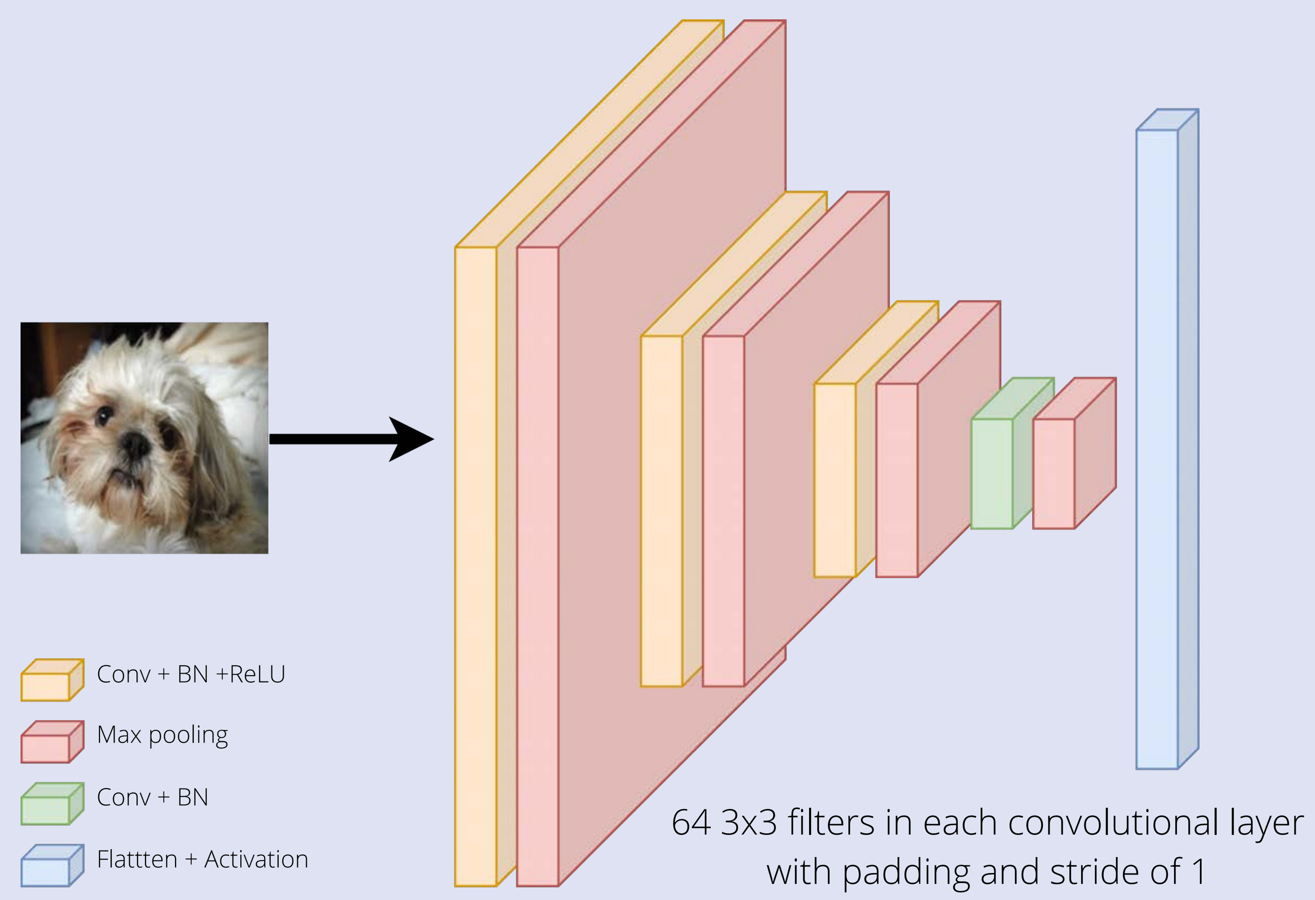
Squared Mahalanobis $\frac{1}{2}(\mathbf{x}-\mathbf{y})^T \underbrace{\mathbf{Q}_k^{-1}}_{\lambda_k \boldsymbol{\Sigma}_k + (1-\lambda_k)\boldsymbol{\Sigma}_k + \beta \mathbf{I}} (\mathbf{x}-\mathbf{y})$

Bregman Divergences

The Bregman divergence for generating function ϕ is given by

$$d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle.$$

Architecture



References

J. Snell, K. Swersky, and R. Zemel.
Prototypical networks for few-shot learning. NIPS (2017).

Q. Le and T. Mikolov.
Distributed representations of sentences and documents. ICML (2014)

A. Banerjee, S. Merugu, I. S. Dhillon and J. Ghosh.
Clustering with Bregman Divergences. JMLR (2005)

P. Bateni, R. Goyal, V. Masrani, F. Wood, L. Sigal.
Improved few-shot visual classification. CVPR (2020)

Code