A Study on the Predictability of Sample Learning Consistency

Alain Raymond-Sáez, Julio Hurtado, Álvaro Soto

Learning Consistency

- Learning Consistency is a measure of how consistently a certain sample is learned by a set of models.
- A metric called C-Score [1] has been proposed to estimate this value, which is the ratio of models that learn a particular sample correctly.

Motivation

- Curriculum Learning [2] is a strategy to show easier examples first and progressively add harder examples during a model's training, which has shown good results for learning faster and more robust models [3]. However, having access to difficulty scores is not trivial.
- Curriculum Learning currently uses C-Score as a proxy for sample difficulty.
- Due to the multiple trained models needed to obtain the final score, C-Score is a computationally intensive solution.

What did we do?

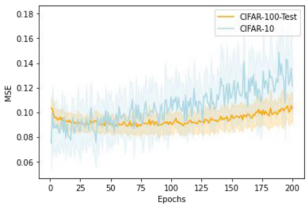
- Since C-Score is such a computationally intensive method we attempt to train a model ***to predict a sample's C-Score from its features alone.***
- With this model, we could alleviate its use and hopefully use it on datasets where we have no such metric.

Methodology

- We predict C-Scores using 3 different methods:
  - Regression
  - Bayesian Personalized Ranking (BPR)
  - Binning
- Each method presents a progressive relaxation of the original regression problem.
- To test each method we use Spearman Rank Correlation (SRC) between the model's prediction and the ordering induced by the ground truth C-Scores.

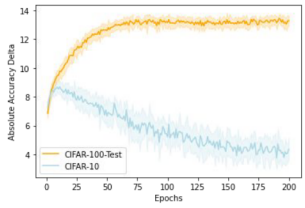
Training Dataset: CIFAR-100
Epochs: 200
N° Seeds: 10
Evaluation Datasets: CIFAR-100 (in-distribution) CIFAR-10 (out of distribution).

Regression



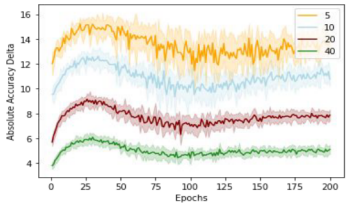
BPR

BPR defines a loss function that encourages the model to learn how to rank a pair of samples appropriately.

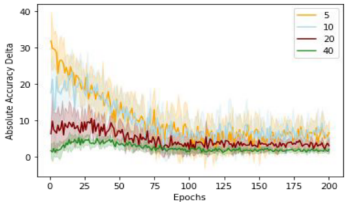


Binning

- We divide the C-Score range of [0,1] in equal width bins.
- Then, we train a model to learn to classify which bin a certain image belongs to given a traditional Cross Entropy Loss.
- We run experiments with 5, 10, 20, 40 bins.



In distribution results



Out of distribution results

References

[1] Jiang, Z., Zhang, C., Talwar, K., and Mozer, M. C. Characterizing structural regularities of labeled data in over-parameterized models.

[2] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning.

[3] Wu, X., Dyer, E., and Neyshabur, B. When do curricula work?

Link to Paper!

Results

- As can be seen in Table 1, for all methods generalization capabilities within the same dataset are quite low, while extrapolation to a new task is barely above random chance. This suggests that **the task may require additional information to be solved**.
- The best method is BPR, which achieves a 0.44 correlation with the ground truth ordering on CIFAR-100 and 0.23 correlation on CIFAR-10. Binning performs below all other methods in all cases, even when using 5 bins, which is a much easier task than BPR or regression.

Dataset	Method	SRC	Std. Dev
CIFAR-100	**BPR**	**0.443082**	**0.006912**
CIFAR-100	REGRESSION	0.368591	0.011231
CIFAR-100	BINS-5	0.300274	0.009779
CIFAR-100	BINS-10	0.326956	0.008409
CIFAR-100	BINS-20	0.337483	0.009155
CIFAR-100	BINS-40	0.324599	0.008783
CIFAR-10	**BPR**	**0.227766**	**0.022339**
CIFAR-10	REGRESSION	0.202101	0.014833
CIFAR-10	BINS-5	0.102341	0.009429
CIFAR-10	BINS-10	0.105720	0.006573
CIFAR-10	BINS-20	0.132004	0.011930
CIFAR-10	BINS-40	0.137987	0.010076

Table 1 - Spearman Rank Correlation for different methods

Conclusions

- We find that these models have limited predictive power within the distributions they were trained on. Out of distribution, these models regress to barely above random chance.
- This suggests that a sample's difficulty is not entirely explained by its features but rather other factors.
- We conjecture that the relation between a sample and its neighbours in an embedding space can help explain the missing factors. Future work will explore this.