UMass Amherst | College of Information & Computer Sciences

Optimizing Neural Networks with Gradient Lexicase Selection

Li Ding[1] & Lee Spector[2,1]

[1] College of Information & Computer Sciences, University of Massachusetts Amherst
[2] Department of Computer Science, Amherst College
liding@umass.edu, lspector@amherst.edu

ICLR
International Conference On Learning Representations

Aggregated Performance Measure

- Modern data-driven learning algorithms are usually optimized by computing the aggregate performance on the training data.

Fitness function

Genetic Algorithms

Loss function

Neural Networks

Seeking "Compromises"

- One potential drawback for aggregated performance measurement is that the model may learn to seek "compromises" and getting stuck at local optima.

			what we prefer
✅ 0.9		✅ 0.99	✅ 0.6
❌ 0.1		❌ 0.1	✅ 0.6
✅ 0.9	some training steps →	✅ 0.99	✅ 0.6
✅ 0.9		✅ 0.99	✅ 0.6
✅ 0.9		✅ 0.99	✅ 0.6
loss: 2.724		loss: 2.343	loss: 2.554

Lexicase Selection [1]

- Uncompromising problems have been recently explored in evolutionary computation for tasks such as program synthesis.
- Instead of using an aggregated fitness function, lexicase selection gradually eliminates candidates by evaluating on each individual training case.
- It has also been used in rule-based learning, symbolic regression, constraint satisfaction problems, machine learning, and evolutionary robotics to improve model generalization.

[1] Thomas Helmuth, Lee Spector, and James Matheson. Solving uncompromising problems with lexicase selection. IEEE Transactions on Evolutionary Computation, 19(5):630–643, 2014.

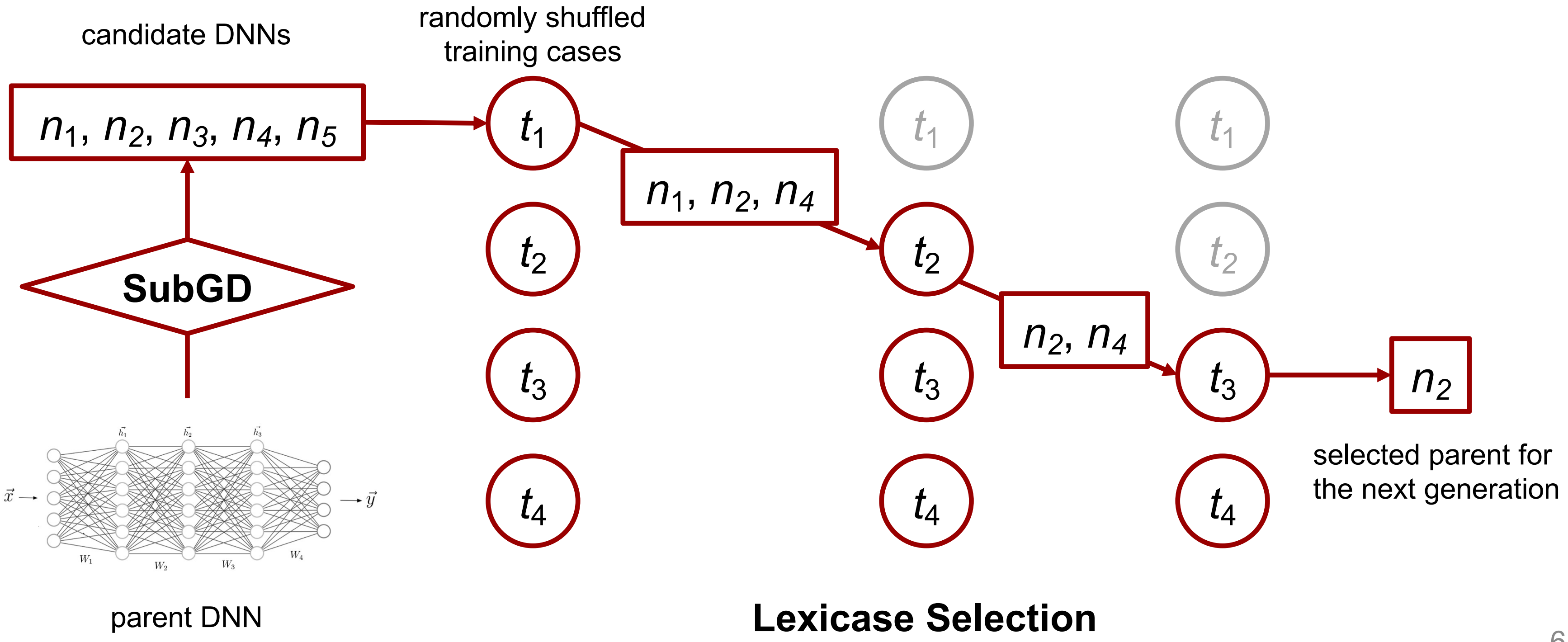
This Work: Gradient Lexicase Selection

- Our goal is to integrate the idea of lexicase selection to improve the generalization of DNNs, while at the same time keep the efficiency of the popular gradient-based learning.
- Our method has two main components: subset gradient descent (SubGD) and lexicase selection.

Mutation by Subset Gradient Descent

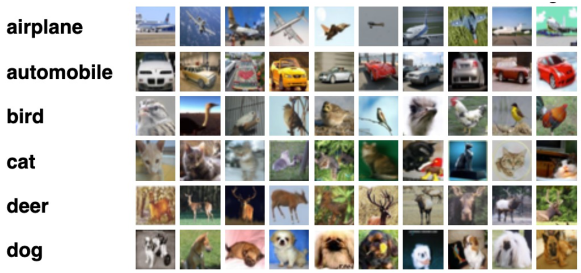
- We propose a gradient-based mutation method: the training set is randomly divided into subsets. Each model candidate is then trained on one of the subsets using stochastic gradient descent.
- There are several advantages:
  - All the candidates are trained with different non-overlapping training samples, so they are more likely to evolve diversely, especially when data augmentation is also included.
  - Each candidate is trained using gradient descent for efficiency.
  - Candidates can be trained in parallel to further reduce computation time.
- The goal is to find a balance between exploration and exploitation towards the whole evolution process.

Gradient Lexicase Selection



Experiments

- Three image classification benchmarks (CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and SVHN (Netzer et al., 2011)) are used for evaluation.
- We implement the proposed algorithm on six popular DNN architectures (VGG (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), MobileNetV2 (Sandler et al., 2018), SENet (Hu et al., 2018), EfficientNet (Tan & Le, 2019)).
- We also implement the original momentum-SGD training as baselines.



Results

Table 1: Image classification results. We report the mean percentage accuracy (*acc.*) with standard deviation (*std.*) obtained by running the same experiment with three different random seeds. The last column (*acc.* ↑) calculates the difference of accuracy by using our method compared to baseline, where positive numbers indicate improvement.

Dataset	Architecture	Baseline		Lexicase		
		acc.	*std.*	*acc.*	*std.*	*acc.* ↑
CIFAR-10	VGG16	92.85	0.10	93.40	0.13	**0.55**
	ResNet18	94.82	0.10	95.35	0.06	**0.53**
	ResNet50	94.63	0.46	94.98	0.18	**0.34**
	DenseNet121	95.06	0.31	95.38	0.04	**0.32**
	MobileNetV2	94.37	0.19	93.97	0.12	-0.39
	SENet18	94.69	0.14	95.37	0.23	**0.68**
	EfficientNetB0	92.60	0.18	93.00	0.22	**0.40**
CIFAR-100	VGG16	72.09	0.52	72.53	0.20	**0.44**
	ResNet18	76.33	0.29	76.68	0.40	**0.35**
	ResNet50	76.82	0.96	77.44	0.25	**0.63**
	DenseNet121	78.72	0.82	79.08	0.26	**0.36**
	MobileNetV2	75.87	0.28	75.57	0.30	-0.30
	SENet18	76.97	0.06	77.22	0.29	**0.25**
	EfficientNetB0	71.03	0.86	71.36	0.87	**0.33**
SVHN	VGG16	96.27	0.06	96.29	0.08	**0.02**
	ResNet18	96.43	0.14	96.62	0.08	**0.19**
	ResNet50	96.69	0.21	96.74	0.07	**0.04**
	DenseNet121	96.82	0.16	96.87	0.03	**0.05**
	MobileNetV2	96.23	0.13	96.26	0.07	**0.03**
	SENet18	96.62	0.19	96.59	0.11	-0.03
	EfficientNetB0	96.14	0.12	95.94	0.10	-0.20

Results

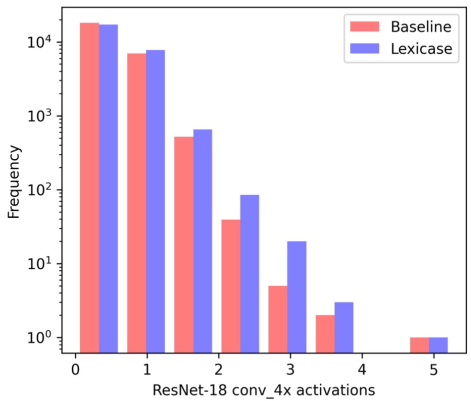
- Comparing other selection methods.

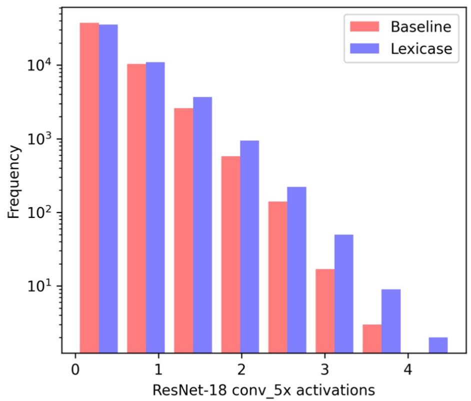
Table 2: Comparing gradient lexicase selection to other selection methods on CIFAR-10. We report the mean percentage accuracy (*acc.*) with standard deviation (*std.*) obtained by running the same experiment with three different random seeds.

Architecture	SGD		Random		Tournament		Lexicase	
	acc.	*std.*	*acc.*	*std.*	*acc.*	*std.*	*acc.*	*std.*
VGG16	92.85	0.10	92.97	0.15	93.12	0.12	**93.40**	0.13
ResNet18	94.82	0.10	94.99	0.12	94.90	0.14	**95.35**	0.06
ResNet50	94.63	0.46	94.75	0.13	94.77	0.04	**94.98**	0.18
DenseNet121	95.06	0.31	95.13	0.04	95.12	0.02	**95.38**	0.04
MobileNetV2	**94.37**	0.19	94.02	0.14	93.91	0.09	93.97	0.12
SENet18	94.69	0.14	95.04	0.15	95.01	0.23	**95.37**	0.23
EfficientNetB0	92.60	0.18	92.77	0.11	92.83	0.12	**93.00**	0.22

Ablation Studies

- Representation Diversity
  - Lexicase selection has been shown to improve population diversity in GP, it may as well help DNNs learn more diverse representations, which improves model generalization.
  - We visualize the feature activations in ResNet-18 trained using normal SGD and gradient lexicase selection, which shows that our method produces more diverse representations.





Conclusion

- We propose Gradient Lexicase Selection, an evolutionary algorithm that incorporates lexicase selection with gradient descent to help optimizing DNNs for better generalization.
- Experimental results show that the proposed method can improve the generalization of popular DNN architectures on the image classification benchmarks.
- Several ablation studies further validates our method. Qualitative analysis also shows that our method can produce better representation diversity.

UMass Amherst | College of Information & Computer Sciences

COMPUTING FOR THE COMMON GOOD

Li Ding (liding@umass.edu) &

Lee Spector (lspector@amherst.edu)