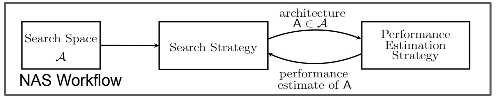
Investigating Differentiable Neural Architecture Search

Data Science Capstone (Fall 2019)

Michael S. Emanuel
Julien Laasri
Dylan Randle
Jiawei Zhuang

Neural Architecture Search (NAS)

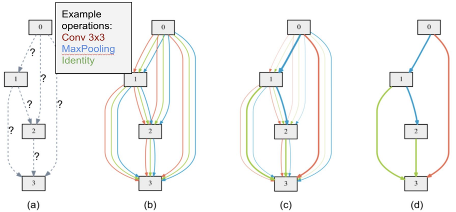
Deep learning frees us from feature engineering, but creates a new problem: "architecture engineering". We use NAS to automate neural network design, with applications to novel scientific datasets.



Differentiable NAS (DARTS)

Liu et. al, 2019: https://arxiv.org/abs/1806.09055

- Continuous relaxation allows *efficient* optimization of "architecture parameters" via gradient descent



Goal: Find the optimal cell, by placing proper operations (e.g. conv, pooling) at edges

Superpose: each edge is the sum over the outputs of multiple operations, weighted by continuous "architecture parameters" α

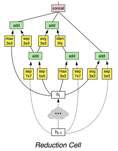
Search: Optimize the architecture weights α, using gradient descent on validation loss

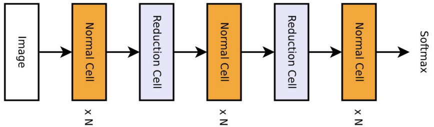
Discretize: select the operation with the highest architecture weight, to be the final architecture

- Different operations are weighted by relative importances

$$\bar{o}^{(i,j)}(x) = \sum_{o \in \mathcal{O}} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in \mathcal{O}} \exp(\alpha_{o'}^{(i,j)})} o(x)$$

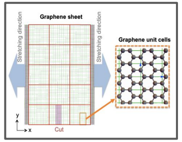
- Train regular weights on train set, architecture on validation
- Discover the best cell and stack them sequentially as final model





Datasets

- **MNIST:** classifying images of handwritten digits
- **Graphene Kirigami:** cutting simulated graphene to optimize stress/strain properties
- **Galaxy Zoo:** classifying galaxy morphology from telescope images
- **Chest X-Ray:** predicting 15 diseases from chest x-ray images



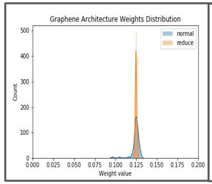
Results

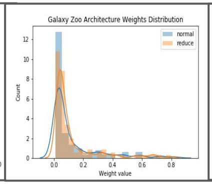
Model	MNIST	Graphene	Galaxy Zoo	Chest X-Ray
DARTS (Continuous)	99.07	0.89	**0.094**	**0.157**
DARTS (Discrete)	99.27	**0.92**	0.114	0.163
Random Search	99.31	0.90	0.098	0.169
ResNet	**99.40**	**0.92**	0.095	0.163
Metric	Acc.	R^2	RMSE	BCE

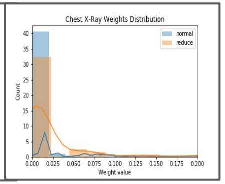
- DARTS best on complex datasets (e.g. Galaxy Zoo, Chest X-Ray)
- *Discretization can fail:* yields worse model on Galaxy Zoo and Chest X-Ray
- ResNet and random search can be competitive
- DARTS search takes ~10x longer than training single model (e.g. ResNet); network 10x bigger, batch size 1/10th
- Random search was run for same GPU time as DARTS

Discussion

Architecture Weights



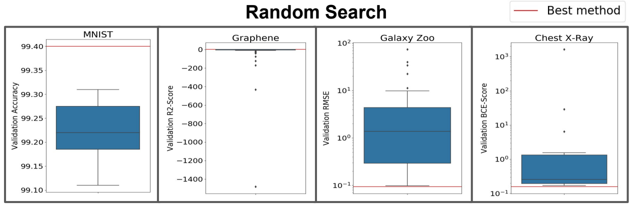




$\sigma_\alpha = 0.003$ $\sigma_\alpha = 0.168$ $\sigma_\alpha = 0.290$

- Architecture weights initialized to ~0.125
- Architecture considered sparse if many weights near 0
- Observation: degree of *architecture sparsity varies considerably* across datasets

Random Search



$\sigma_{ACC} = 0.068$ $\sigma_{R^2} = 220$ $\sigma_{RMSE} = 12.4$ $\sigma_{BCE} = 405$

- High variance: performance *sensitive* to architecture
- Low variance: performance *insensitive* to architecture

Conclusions & Future Work

- DARTS is a useful tool, but overkill on simple tasks
- ResNet and random search could be good enough
- DARTS introduces many additional hyperparameters
- DARTS discretization step is heuristic
- *Future work:* encourage sparsity in DARTS architectures (e.g. sparsemax vs. softmax) to prevent discretization failure