# Optimizer Combination Analysis for Differentiable Neural Architecture Search

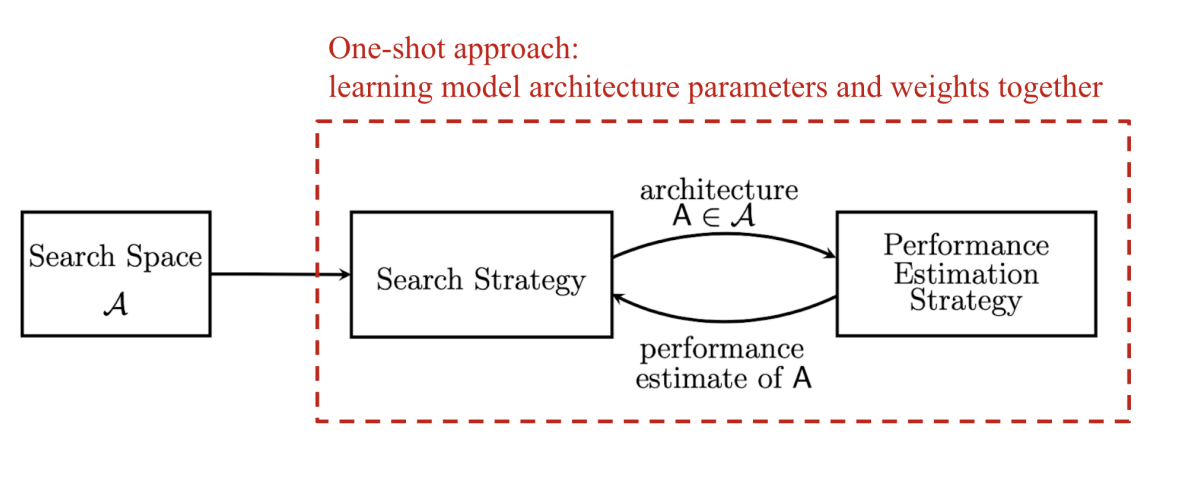
## CSED 490Y: Optimization for Machine Learning

Jiwoong Shin

2022.05.30

Group 9

# Neural Architecture Search (NAS)




**Neural Architecture Search** is a concept to automatically search neural architecture in specific search space.

- **Search Space**
  - All candidate architectures
- **Search Strategy**
  - How to search?
- **Performance Estimation Strategy**
  - How to evaluate performance of an architecture?

# DARTS

**DARTS** changes neural architecture search problem to continuous optimization problem based on bi-level optimization.

$$\min_{\alpha} \ \mathcal{L}_{val}(w^*(\alpha), \alpha)$$
$$\text{s.t. } w^*(\alpha) = \text{argmin}_w \mathcal{L}_{train}(w, \alpha)$$

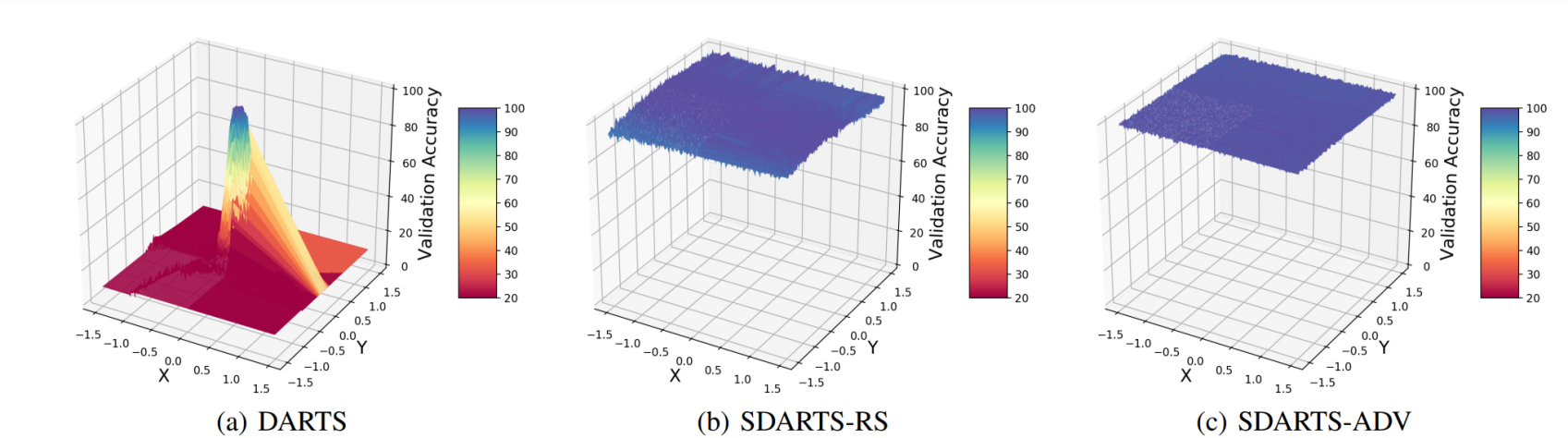
- $w$: weight parameters
- $\alpha$: architecture parameters
- $w^*(\alpha)$: optimal weight parameters for architecture $\alpha$

## Architecture Gradient Approximation

- First: $\nabla_\alpha \mathcal{L}_{val}(w^*(\alpha), \alpha) \approx \nabla_\alpha \mathcal{L}_{val}(w, \alpha)$
- Second: $\nabla_\alpha \mathcal{L}_{val}(w^*(\alpha), \alpha) \approx \nabla_\alpha \mathcal{L}_{val}(w - \xi \nabla_w \mathcal{L}_{train}(w, \alpha), \alpha)$

We alternatively update architecutre parameter $\alpha$ and weight parameter $w$.

# Smooth DARTS (SDARTS)




(a) DARTS (b) SDARTS-RS (c) SDARTS-ADV

- After searching with DARTS, landscape of validation accuracy regarding the architecture weight is uneven.
  - If we discretize DARTS's continuous encoding to derive a architecture, the architecture get lower accuracy than we expected.
- How about force **the landscape of $\mathcal{L}_{val}(\bar{w}(A), A + \Delta)$ to be more smooth with respect to the perturbation**?

# Smooth DARTS (SDARTS)

- Smooth DARTS focus on reducing the **Hessian matrix of architecture parameters**.
  - A smaller Hessian norm results in a flatter loss landscape.
- Injection noise in architecture parameters implicit regularize the Hessian norm

**Smooth DARTS's objective**

$$\min_{A} \mathcal{L}_{val}(\bar{w}(A), A)$$

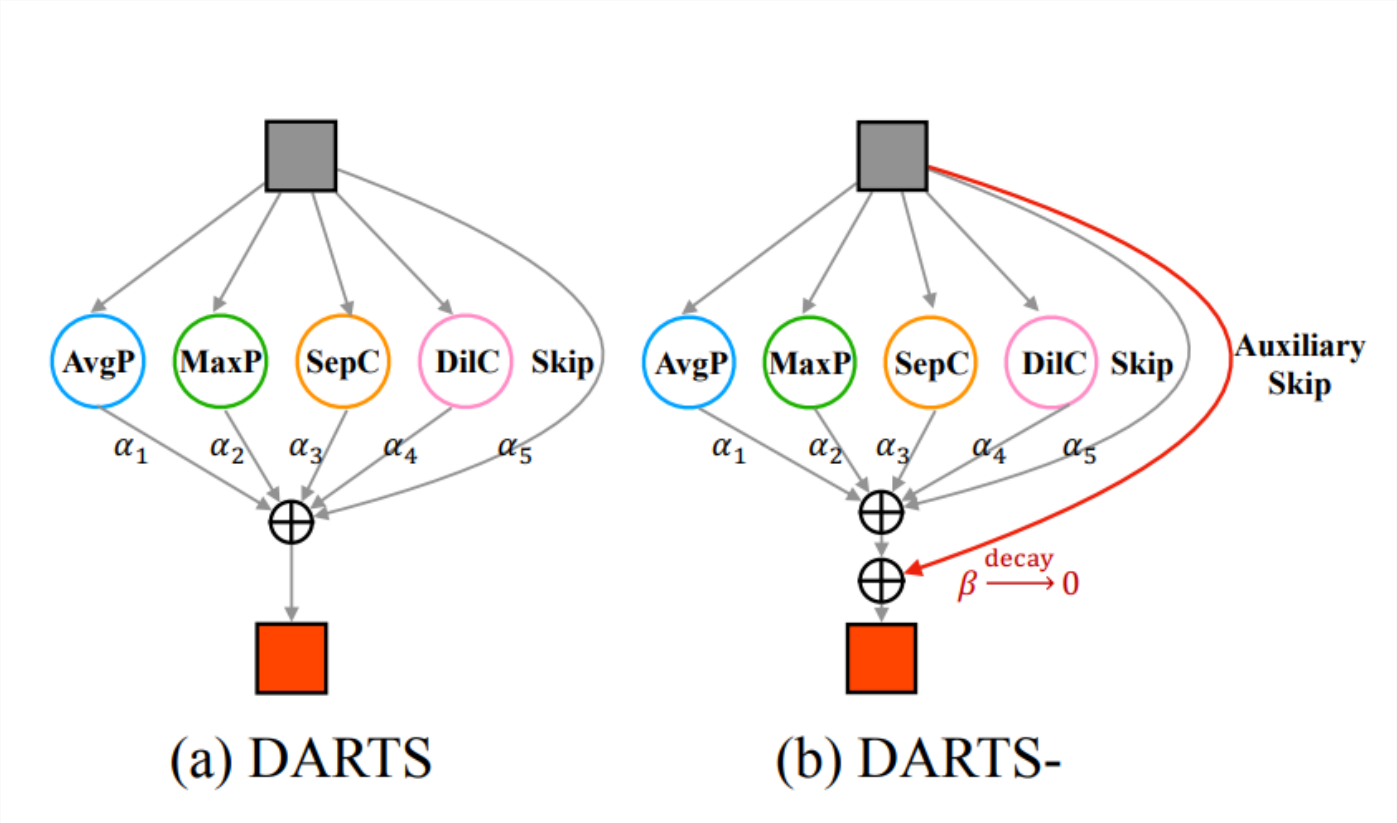
SDARTS-RS: $\bar{w}(A) = \arg\min_{w} \mathbb{E}_{\delta \sim U[-\epsilon,\epsilon]} \mathcal{L}_{train}(w, A + \delta)$

# DARTS-

- In differentiable architecture search, the skip connections play **two roles**.
  1. Alleviate vanishing gradient
  2. Be considered as one of candidate operations
- In searching phase, importance of skip connection not only influenced from role 2 but also **role 1**.
  - This makes domination of skip connections!
- We need to separate two functions of skip connection to eliminate unfair advantage!

# DARTS-

- DARTS- separate the functions of skip connections by adopting **auxiliary skip connection**.
  - In search phase, auxiliary skip connection is gradually decreased to eliminate the impact of itself.



# NAS Bench 201

**NAS Bench 201** is a benchmark dataset for various neural architecture structures.

- Small but all architectures of search space are evaluated.
  - # of architectures: 15625
  - Benchmark dataset has various information for each architecture.
    - e.g. Accuracy for CIFAR10, CIFAR100, and ImageNet dataset, Loss for CIFAR10, CIFAR100, and ImageNet dataset etc.
- It makes us to track the performance of current architecture with current architecture encoding immediately.

# Fixed Optimizer Setting

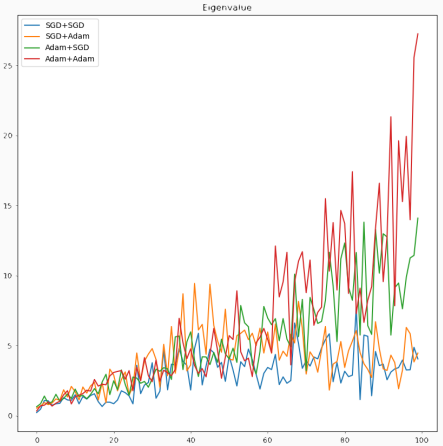
DARTS[3] and its variants[1, 2] use **fixed optimizer setting** to search architecture in supernet.

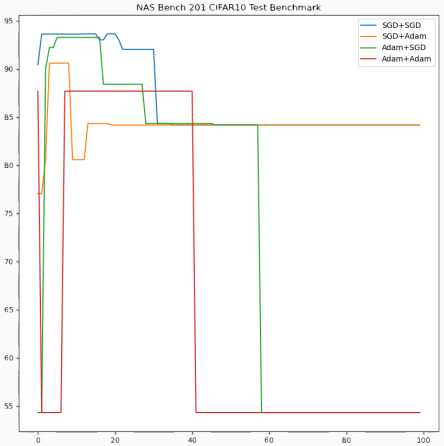
- weight parameters: SGD
- architecture parameters: Adam

However, is it the optimal choice? Is there any room for improvement?

- No specific investigation on this

# Optimizer Comparison for Architecture Search


- Supernet Accuracy
  - SGD + SGD, SGD + Adam > Adam + SGD, Adam + Adam
- Eigenvalue of Hessian matrix for architecture parameters
  - SGD + SGD, SGD + Adam < Adam + SGD, Adam + Adam
- NAS Bench 201
  - SGD + SGD, SGD + Adam > Adam + SGD, Adam + Adam

For weight parameters: SGD > Adam

# Why SGD is better than Adam?

## Property of supernet

In recent work[5], the convergence of network weights $w$ can heavily rely on $\beta_{skip}$ in the supernet.

- Settings
  - Three operations: convolution, skip connection, none
  - Loss: MSE Loss

By single update step, the training loss can be reduced by ratio $(1 - \eta_w \varphi/4)$ with probability of at least $1 - \sigma$.

- $\eta_w$: Learning rate
- $\varphi$: To be introduced in next slide

# Why SGD is better than Adam?

$\varphi$ obeys

- $h$: number of supernet layers

DARTS

$$\varphi \propto \sum_{i=0}^{h-2} \left[ \left( \beta_{conv}^{(i,h-1)} \right)^2 \prod_{t=0}^{i-1} \left( \beta_{skip}^{(t,i)} \right)^2 \right]$$

- $\varphi$ depends more on $\beta_{skip}$ than $\beta_{conv}$
- That is, it makes wrong gradient (guidance) for supernet update.

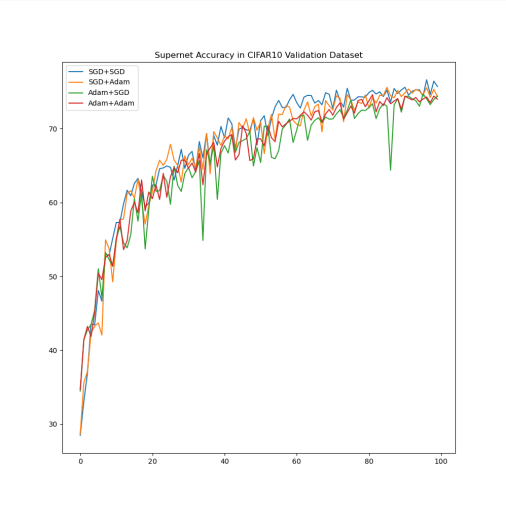
If the Adam algorithm with a faster convergence speed is used, this bias can be more affected.

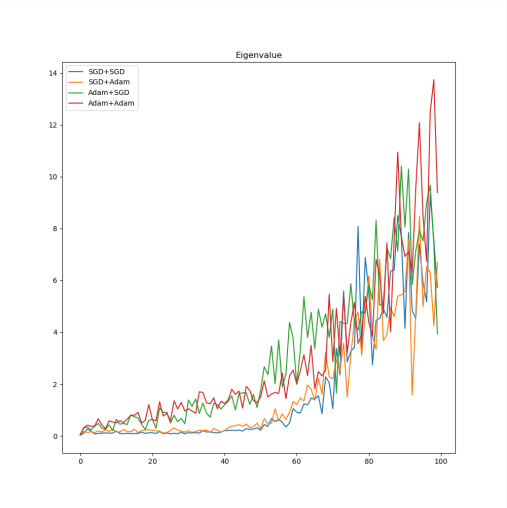
# Why SGD is better than Adam?

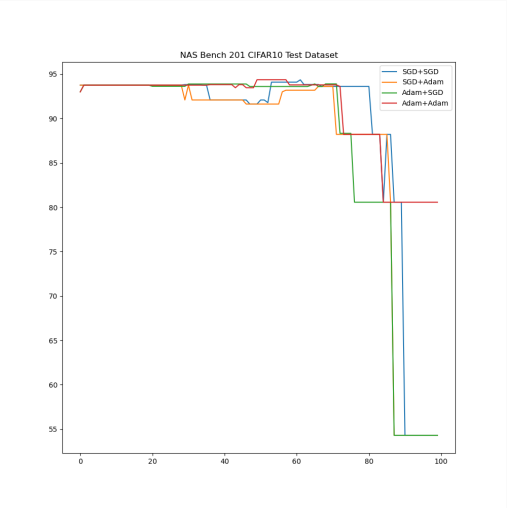
## Generalization ability of SGD

- SGD can show better generalization ability than Adam [4]
- This property makes the searched architecture more robust in discretization process.
  - We can update architecture parameters on more reliable shared weight parameters.

# Different Search Algorithms: DARTS-




DARTS- shows similar pattern with DARTS

- In early and middle time, It shows powerful search performance.

# Different Search Algorithms: DARTS-

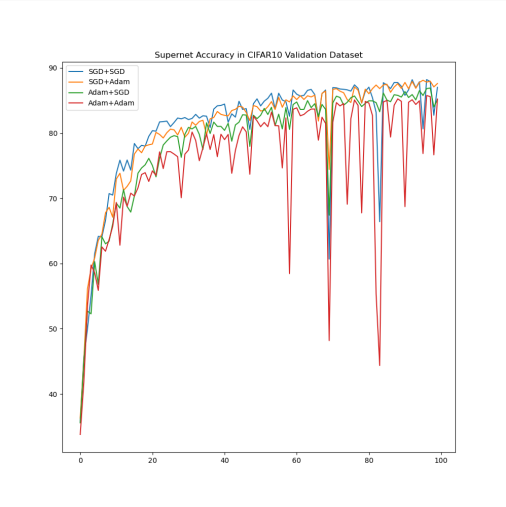
When we use DARTS-, supernet convergence can be formulated as follows.

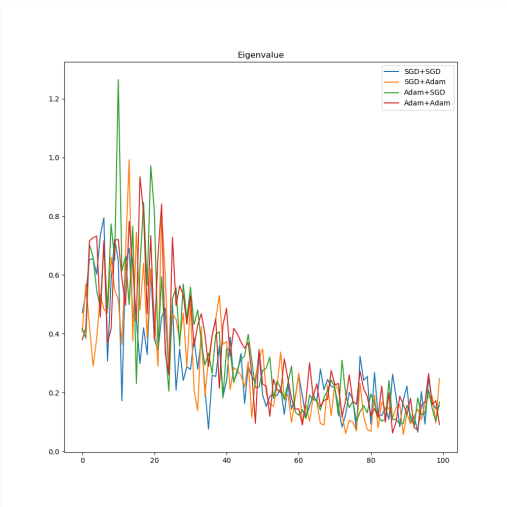
$$\varphi \propto \sum_{i=0}^{h-2} \left[ \left( \beta_{conv}^{(i,h-1)} \right)^2 \prod_{t=0}^{i-1} \left( \beta_{skip}^{(t,i)} + \beta \right)^2 \right]$$

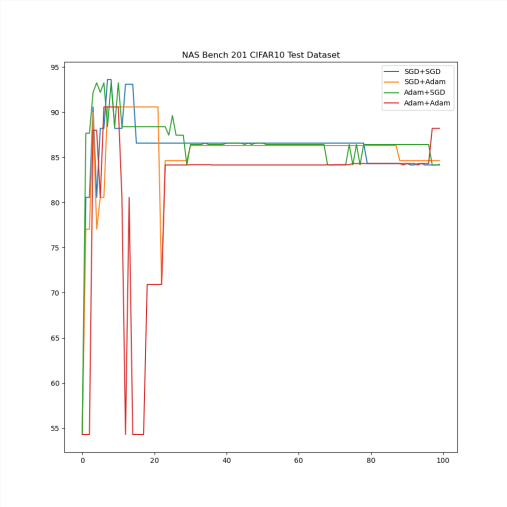
$\beta$ means weight for auxiliary skip connection.If $\beta >> \beta_{skip}$, $\beta_{skip}$ do not determine supernet update.

- However, in later part of optimization, because we decay the auxiliary skip connection, DARTS's problem occurs on DARTS- too.

# Different Search Algorithms: SDARTS



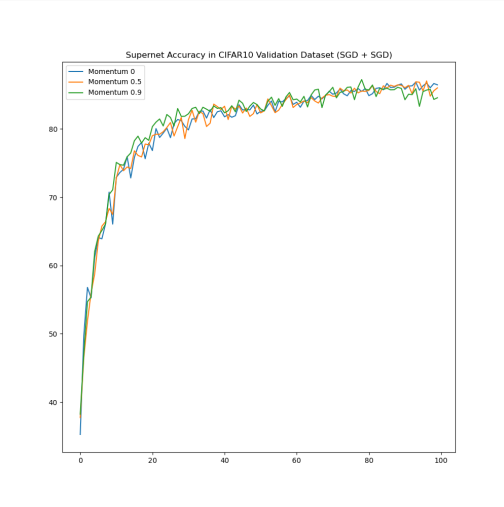



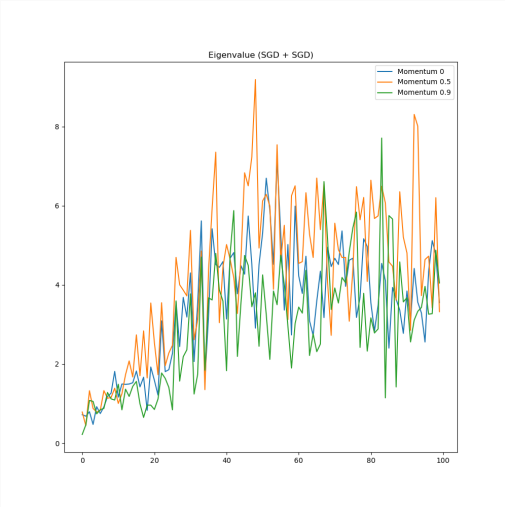

SDARTS shows different pattern with DARTS and DARTS-

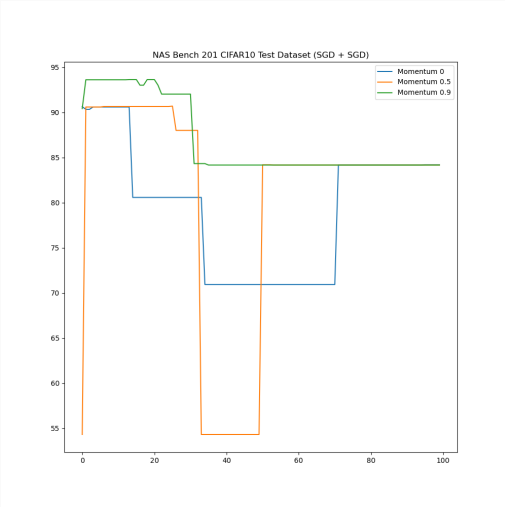
- In SDARTS, **Adam + Adam** shows the best performance in searching.
- This is because SDARTS uses additional regularization technique.
    - Add noise!
- The regularization prevents search process taking unfair advantage.

# Momentum comparison for Architecture Search

## DARTS

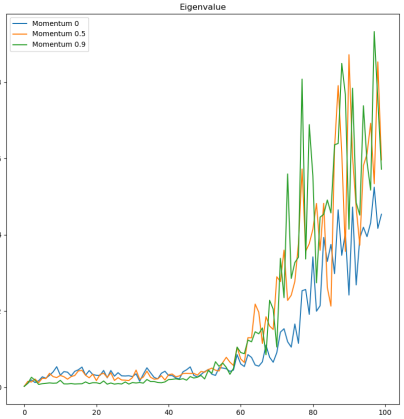


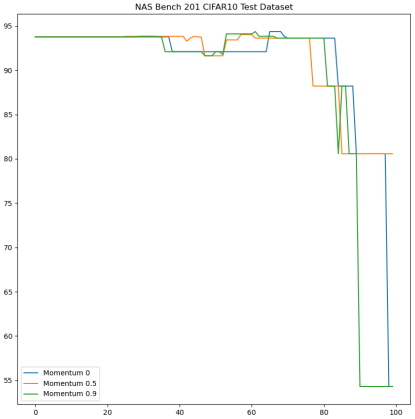



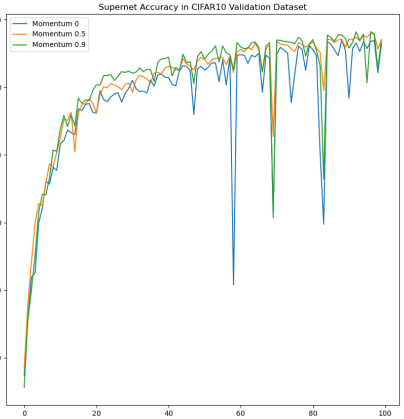
# Momentum comparison for Architecture Search

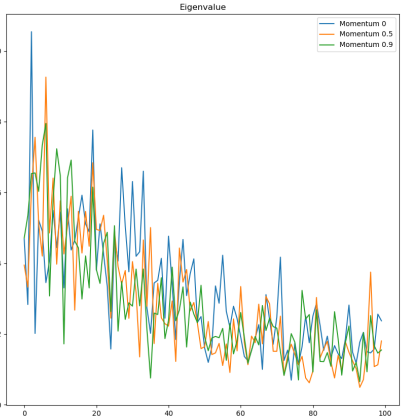
## DARTS-

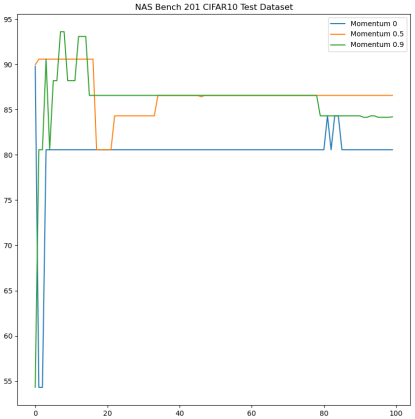



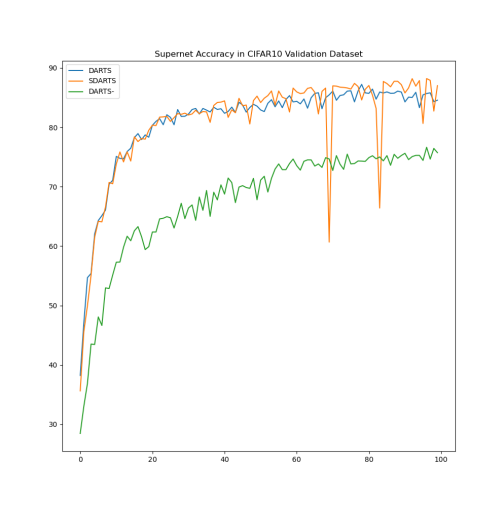

## SDARTS

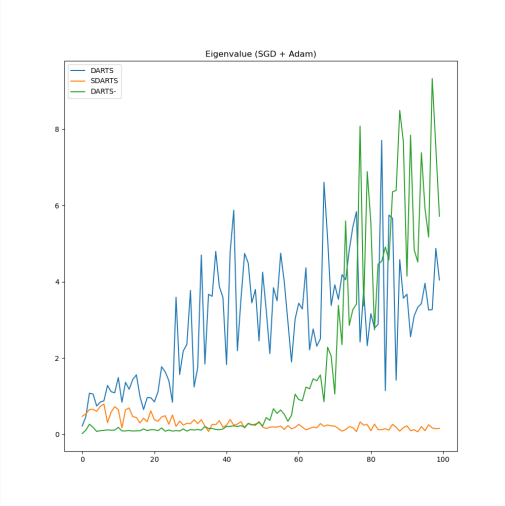


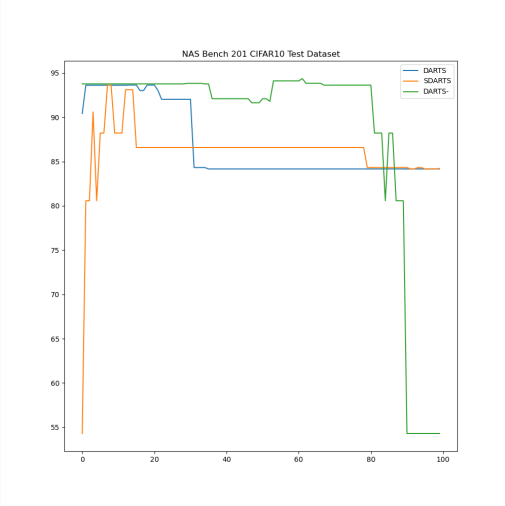



# Algorithm Comparison: Fundamental Problem




Not converged network shows powerful performance

- In early stage, all methods show powerful performance, but lose their performance in later.

# Conclusion

1. In differentible architecture search, SGD optimizer is better than Adam optimizer for weight parameters if there is no regularization method.
2. Moderate level of momentum can help to reduce performance degradation.
3. There are fundamental problem in differentiable architecture search that fully trained supernet does not give powerful architecture.
    - We need to find suitable regularization method to make supernet's performance be indicator of the performance of searched architecture.

# References i


X. Chen and C. Hsieh.
**Stabilizing differentiable architecture search via perturbation-based regularization.**
In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1554–1565. PMLR, 2020.

X. Chu, X. Wang, B. Zhang, S. Lu, X. Wei, and J. Yan.
**DARTS-: robustly stepping out of performance collapse without indicators.**
In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

# References ii


H. Liu, K. Simonyan, and Y. Yang.
**DARTS: differentiable architecture search.**
In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

P. Zhou, J. Feng, C. Ma, C. Xiong, S. C. Hoi, and W. E.
**Towards theoretically understanding why sgd generalizes better than adam in deep learning.**
In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

# References iii

P. Zhou, C. Xiong, R. Socher, and S. C. Hoi.
**Theory-inspired path-regularized differential network architecture search.**
In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.