Overcoming Multi-Model Forgetting

In one of the first NAS papers using Reinforcement Learning, Zoph et al. (Google) used more than 800 GPUs in parallel for two weeks.

Weight Sharing was introduced in NAS to speed up the process.
Assumptions

Our hypothesis:

1. Weight-sharing can negatively affect architectures.

2. If justified, this can lead to a wrong evaluation of candidates in NAS, making the evaluation phase closer to random
Multi-Model Forgetting

Top 1 Accuracy vs Training iterations

- Model A decrease
- Model B increase

A after (1)
Baseline
A during (2) B during (2)
1 shared 1 shared
2 shared 2 shared
3 shared 3 shared
Study of Weight-Sharing

Simple scenario of two models sharing parameters:

\[ f_1(\mathcal{D}; \theta_1, \theta_s) \text{ and } f_2(\mathcal{D}; \theta_2, \theta_s) \]

Assume that we have access to the optimal parameters \((\hat{\theta}_1, \hat{\theta}_s)\) of the first model \(f_1(\mathcal{D}; \theta_1, \theta_s)\).

**Maximizing** the posterior distribution \(p(\theta \mid \mathcal{D}), \theta = (\theta_1, \theta_2, \theta_s)\)

\[
\mathcal{L}_{WPL}(\theta_2, \theta_s) = \mathcal{L}_2(\theta_2, \theta_s) + \frac{\lambda}{2} (\|\theta_s\|^2 + \|\theta_2\|^2) + \frac{\alpha}{2} \sum_{\theta_{s_i} \in \theta_s} F_{\theta_{s_i}} (\theta_{s_i} - \hat{\theta}_{s_i})^2
\]

Cross-entropy loss  L2 regularization

**Weight importance**
Experiments on Two Models

- **WPL** reduces multi-model forgetting

- **WPL** have a **minimal** effect on the learning of the second model
ENAS on PTB

(a) Mean diff.
(b) Best 5 mean diff.
(c) Max diff.

Error Difference (diff.)

Cross Entropy
WPL

Epochs
Summing up

To recap, our main contributions are:

1. **Weight Sharing** negatively impacts NAS
2. **Weight Sharing** can cause the **search phase** in NAS to become closer to random
3. **WPL** reduces Multi-Model Forgetting

Pacific Ballroom #19
(6:30pm - 9pm)