Differential neural ensemble search with diversity control

Petr Babkin

Moscow Institute of Physics and Technology

Course: My first scientific paper (Strijov's practice)/Group B05-003

Expert: O. Bakhteev

Consultants: K. Yakovlev, K. Petrushina

2023

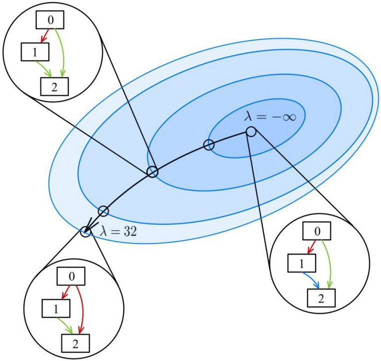
Goals of research

- Propose a novel method of sampling deep learning models with diversity control
- Investigate sampled models in terms of diversity and performance
- Test different ensemles of the sampled models
- Compare performance with other state-of-art methods

Problem of creating NN ensembles

Method:

1. Find optimal architecture
2. Sample architectures with diversity control
3. Give answer as equal voting



Problem of sampling new models for ensemble:

$$\min_{\alpha} \mathbb{E}_{\lambda \sim U(0,\Lambda)}[\mathcal{L}_{val}(w^*, \alpha(\lambda)) - \lambda JS(\alpha^*, \alpha(\lambda))]$$

$$s.t.\ w^* = \arg\min_{w} \mathbb{E}_{\lambda \sim U(0,\Lambda)}[\mathcal{L}_{train}(w, \alpha(\lambda))]$$

Literature

- Yao Shu1, Yizhou Chen, Zhongxiang Dai, Bryan Kian, Hsiang Low: Neural Ensemble Search via Bayesian Sampling
- Hanxiao Liu, Karen Simonyan, Yiming Yang: DARTS: Differentiable Architecture Search
- Konstantin Yakovlev, Olga Grebenkova, Oleg Bakhteev, Vadim Strijov: Neural Architecture Search with Structure Complexity Control
- Ashwin Raaghav Narayanan, Arber Zela, Tonmoy Saikia, Thomas Brox, Frank Hutter: Multi-headed Neural Ensemble Search

Problem statement

Classic problem of searching for NN ensembles

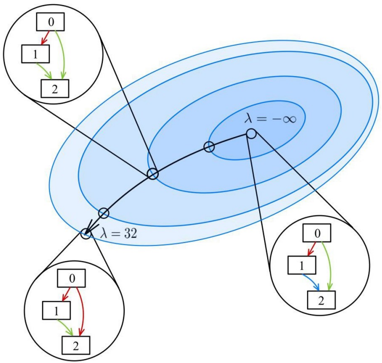
$$\min_{S} \mathcal{L}_{val} \left(\frac{1}{|S|} \sum_{\boldsymbol{\alpha} \in S} f(\boldsymbol{w}^*_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) \right)$$
$$s.t.\ \forall \boldsymbol{\alpha} \in S\ \ \boldsymbol{w}^*_{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{w}} \mathcal{L}_{train}(f(\boldsymbol{w}^*_{\boldsymbol{\alpha}}, \boldsymbol{\alpha}))$$

Rearranged problem of searching for NN ensembles

$$\min_{\alpha} \mathbb{E}_{\lambda \sim U(0,\Lambda)}[\mathcal{L}_{val}(w^*, \alpha(\lambda)) - \lambda JS(\alpha^*, \alpha(\lambda))]$$
$$s.t.\ w^* = \arg\min_{w} \mathbb{E}_{\lambda \sim U(0,\Lambda)}[\mathcal{L}_{train}(w, \alpha(\lambda))]$$

where λ is a diversity parameter

Hypotheses and model



Architectural space

- Architectural space is continuous
- Architectures differ in terms of JSd
- The further architecture locates the worse accuracy it performs
- Diversity and performance are both important for ensembling

Solution

We sample architectures using hypernetwork, a parametric mapping

$$h : [0, \Lambda] \times \mathbb{R}^u \to \mathbb{R}^s,$$

where $\mathbb{R}^u$ is hypernetwork parametric space and $\mathbb{R}^s$ is architectural space

Algorithm

Initialize: $N \in \mathbb{N}$, $\mathcal{S} = \varnothing$, hypernetwork
S $\leftarrow \{\ \alpha^*\}$ ▷ Result os NAS
for $i = 1, \dots, N$ **do**
    Sample $\lambda \sim U(0, \Lambda)$
    S $\leftarrow$S $\cup \{\alpha(\lambda, \alpha^*)\}$ ▷ α gaind from hypernetwork
end for
Return: S as a resulting ensemble

Goals of computational experiment

To conduct experiments on fashionMNIST dataset following two main problems

Comparison of architectures

- Investigate performance of architectures
- Investigate diversity of architectures

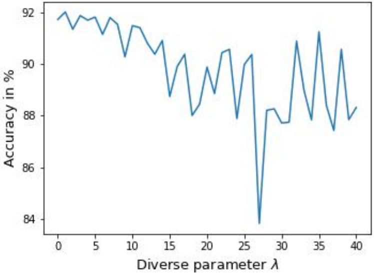
Ensembling effectiveness investigation

- Try different diversity parameter limits Λ for architectures to be diverse enough and show good performance
- Compare results with state-of-art NES algorithms

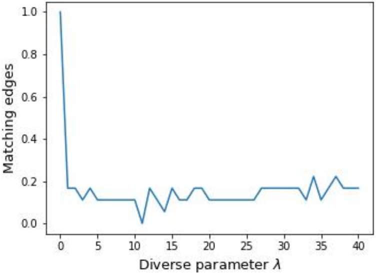
Comparison of architectures

As it was expected the more diverse architectures are the worse they perform

However, due to non-convexity of problem, even for minor λ architectures almost do not intersect



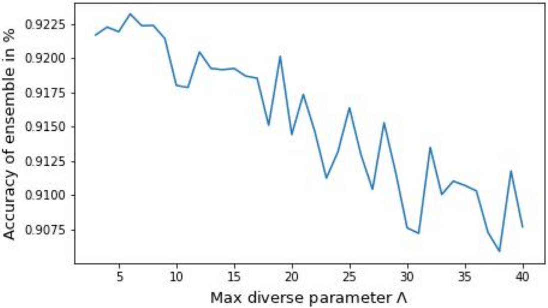
Accuracy on λ



Matching edges on λ

Resulting ensemble

Least diverse ensemble showed best performance on this problem, although all ensembles perform competible accuracy



Ensembles performance on diversity limit

Conclusion

Achieved results

- A novel method of sampling deep learning models ensembles was proposed
- Sampled architectures were investigated in terms of diversity and performance
- Diversity of sampled models was controlled
- Resulting ensemble showed compatible performance

Further investigations

- Try varying diverse parameter λ for architectures to be closer to optimum
- Try another regularizer based on intersections function
- Try starting algorithm from an optimal architecture