Metaparameters and metaoptimization

MIPT

2023

Hyperparameters

Definition

Prior for parameters w and structure $\boldsymbol{\Gamma}$ of the model f is a distrubution $p(\mathsf{W}, \boldsymbol{\Gamma}|\mathsf{h}) : \mathbb{W} \times \Gamma \times \mathbb{H} \to \mathbb{R}^+$, where $\mathbb{W}$ is a parameter space, Γ is a structure space.

Definition

Hyperparameters $\mathsf{h} \in \mathbb{H}$ of the models are the parameters of $p(\mathsf{w}, \boldsymbol{\Gamma}|\mathsf{h})$ (parameters of prior f).

Metaparameters

Wikipedia

A parameter that controls the value of one or more others.

Definition

Metaparameters $\boldsymbol{\lambda}$ are the parameters of optimization problem.

Is dropout rate a metaparameter or hyperparameter?

Consider using Gumbel-Softmax as a variational distribution for model deep neural network structure. Is temperature a metaparameter or hyperparameter?

Is SGD learning rate a metaparameter or hyperparameter?

Gradient descent for evidence estimation

Statement

Let L be a Lipschitz function, and optimization operator be a bijection. Then entropy difference for two steps is:

$$\mathrm{S}(q'(\mathrm{w})) - \mathrm{S}(q(\mathrm{w})) \simeq \frac{1}{r}\sum_{g=1}^{r}\left(-\beta \mathrm{Tr}[\mathrm{H}(\mathrm{w}'^{g})] - \beta^2 \mathrm{Tr}[\mathrm{H}(\mathrm{w}'^{g})\mathrm{H}(\mathrm{w}'^{g})]\right).$$

Final estimation for the τ optimization step:

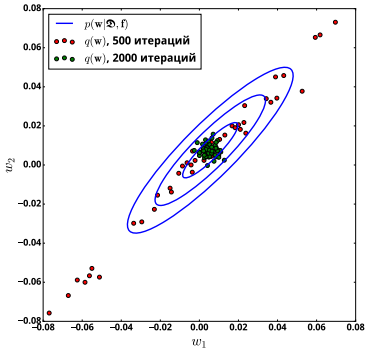
$$\log \hat{p}(\mathrm{Y}|\mathfrak{D},\mathrm{h}) \sim \frac{1}{r}\sum_{g=1}^{r} L(\mathrm{w}_{\tau}^{g},\mathfrak{D},\mathrm{Y}) + \mathrm{S}(q^0(\mathrm{w})) +$$

$$+\frac{1}{r}\sum_{b=1}^{\tau}\sum_{g=1}^{r}\left(-\beta \mathrm{Tr}[\mathrm{H}(\mathrm{w}_b^g)] - \beta^2 \mathrm{Tr}[\mathrm{H}(\mathrm{w}_b^g)\mathrm{H}(\mathrm{w}_b^g)]\right),$$

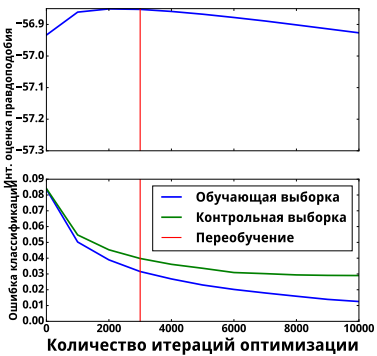
w_b^g is a parameter vector for optimization g on the step b, $\mathrm{S}(q^0(\mathrm{w}))$ is an initial entropy.

Overfitting, Maclaurin et. al, 2015

Gradient descent does not optimize KL-divergence $\mathrm{KL}(q(\mathrm{w})||p(\mathrm{w}|\mathfrak{D},\mathrm{h}))$. Evidence estimation gets worse while optimization tends to the optimal parameter values. This can be considered as a overfitting start.



Convergence

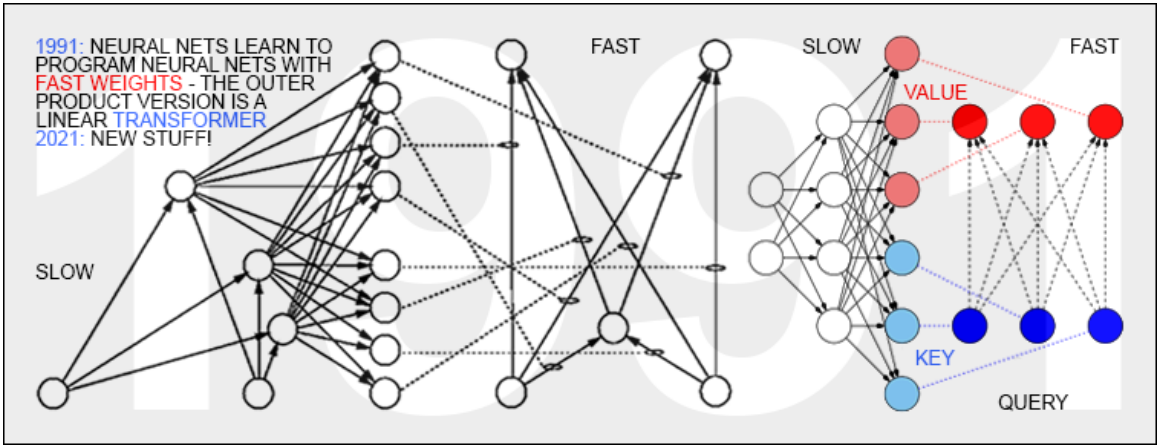


Overfitting start

A neural network that embeds its own meta-levels

A model is decomposed into submodels with different subtasks:

- “Normal” model: optimization and inference.
- Evaluation model: validation quality estimation.
- Analyzing model: aanalysis of model parameters.
- Modifiyng model: parameter modification.



https://people.idsia.ch/~juergen/fast-weight-programmer-1991-transformer.html

Learning to learn by gradient descent by gradient descent

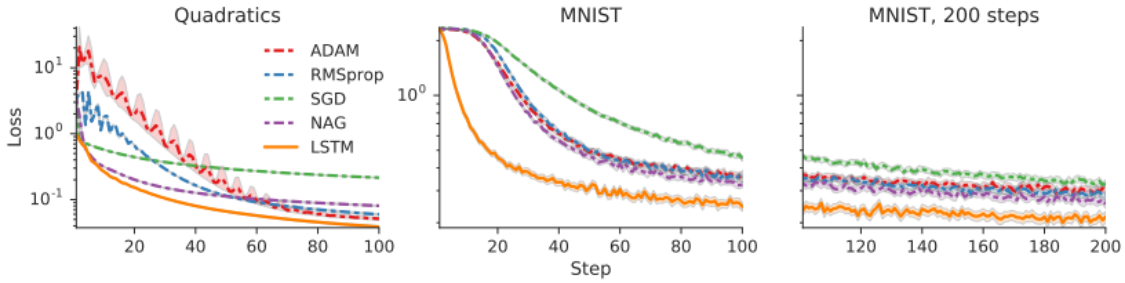
Idea: consider optimization T as a differential function:

$$T(\theta) = \text{LSTM}(\theta).$$

Optimization problem:

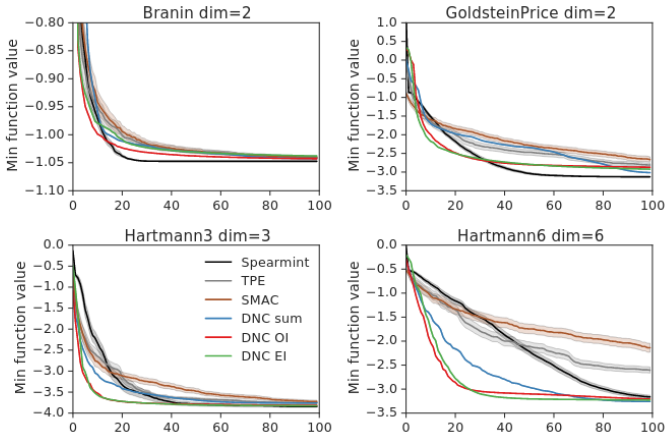
$$\sum_{t=t_0}^{t_\eta} L\left(T^t(\theta_{t_0})\right) \to \max.$$

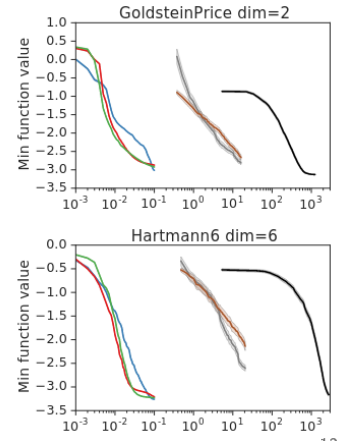
LSTM has a small number of parameters, it shares parameters for each metaparameters.



Learning to Learn without Gradient Descent by Gradient Descent

- Using the same idea, but for hyperparameters optimization (alternative to GP)
- Can be used with a combination of different optimization functions (e.g. EI)
- Does not require gradient at the «test» time

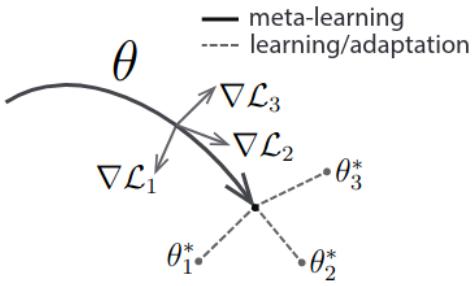




Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

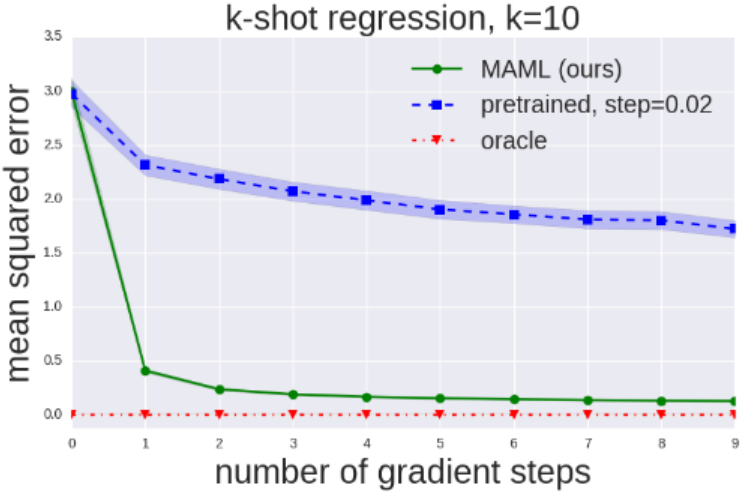
Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

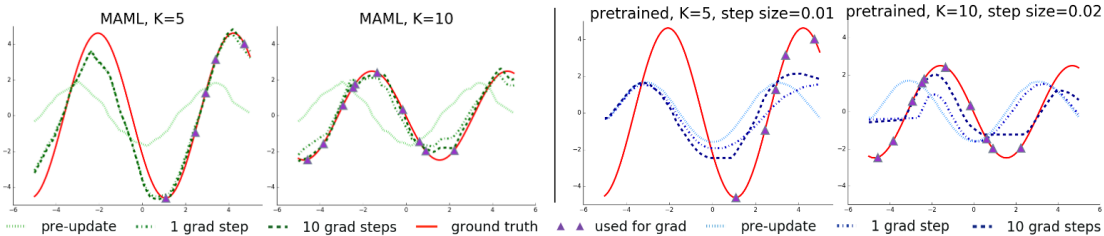


2

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks



Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks



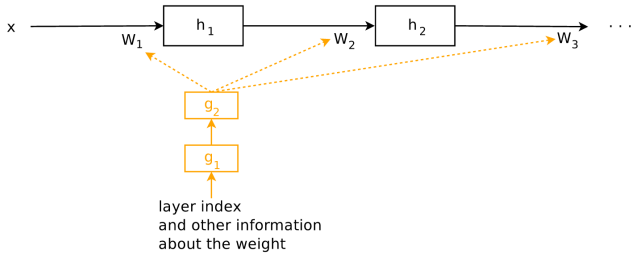
Гиперсети

Definition

Given a set Λ.
Hypernetwork is a parametric mapping from Λ to set $\mathbb{R}^n$ of the model f parameters:

$$G : \Lambda \times \mathbb{R}^u \to \mathbb{R}^n,$$

where $\mathbb{R}^u$ is a set of hypernetwork parameters.



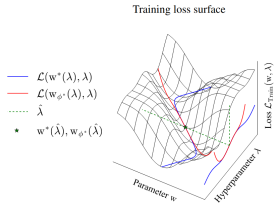
Ha et al., 2016

Stochastic Hyperparameter Optimization through Hypernetworks

$$\mathrm{E}_{\lambda}\left(-\log\ p(\mathfrak{D}|\mathrm{w}(\lambda)) + \lambda||\mathrm{w}(\lambda)||_2^2\right) \rightarrow \min$$

Theorem

Sufficiently powerful hypernetworks can learn continuous best-response functions, which minimizes the expected loss for all hyperparameter distributions with convex support.



Lorraine et al., 2016

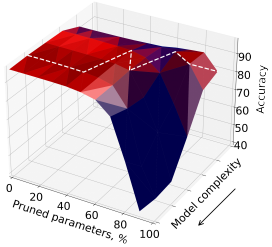
Deep learning model selection with parametric complexity control

Theorem (Grebenkova, 2023)

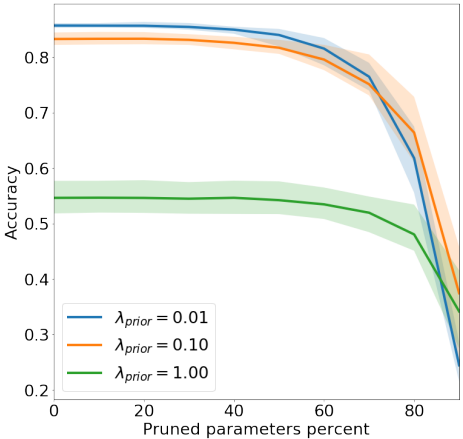
Hypernetowkr allows to learn not only the best-response functions, but also the statistical properties of the models.

Corollary

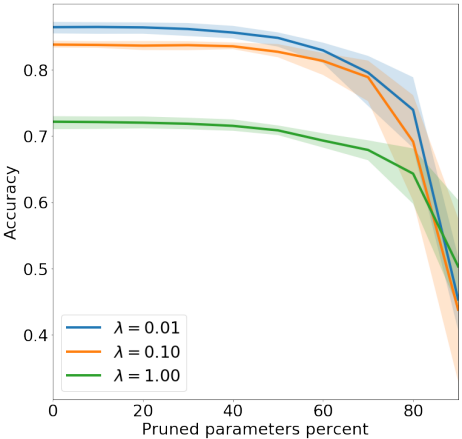
We can control the metaparameters, such as pruning rate or model complexity.



Example: CIFAR-10



CNN



CNN with hypernetwork

References

- Schmidhuber, Jürgen. "A neural network that embeds its own meta-levels."IEEE International Conference on Neural Networks. IEEE, 1993.
- Dougal Maclaurin et. al, Gradient-based Hyperparameter Optimization through Reversible Learning, 2015
- Andrychowicz M. et al. Learning to learn by gradient descent by gradient descent //Advances in neural information processing systems. – 2016. – С. 3981-3989.
- Chen Y. et al. Learning to learn without gradient descent by gradient descent //International Conference on Machine Learning. – PMLR, 2017. – С. 748-756.
- Ha D., Dai A., Le Q. V. Hypernetworks //arXiv preprint arXiv:1609.09106. – 2016.
- Lorraine J., Duvenaud D. Stochastic hyperparameter optimization through hypernetworks //arXiv preprint arXiv:1802.09419. – 2018.
- Гребенькова О. С., Бахтеев О. Ю., Стрижов В. В. Вариационная оптимизация модели глубокого обучения с контролем сложности //Информатика и её применения. – 2021. – Т. 15. – №. 1. – С. 42-49.
- Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks."International conference on machine learning. PMLR, 2017.
- Olga Grebenkova, Oleg Bakhteev, Vadim Strijov , Deep Learning Model Selection With Parametric Complexity Control , ICAART 2023.