Learning to learn by gradient descent by gradient descent

Гущенко-Чеверда Иван, 141

Классические методы оптимизации

Градиентный спуск

$$\theta_{t+1} = \theta_t - \alpha_t \nabla f(\theta_t) .$$

Методы использующие информацию о кривизне(метод Ньютона, BFGS, L-BFGS).

Методы оптимизации нейросетей

- Стохастический градиентный спуск
- RMSprop

$$\theta_{t+1,i} \leftarrow \theta_{t,i} - \frac{\gamma}{\sqrt{v_{t,i}^2 + \varepsilon}} \nabla_{\theta_i} \ell_t(\theta_t),$$
$$v_{t,i}^2 \leftarrow \beta v_{t-1,i}^2 + (1-\beta)(\nabla_{\theta_i} \ell_t(\theta_t))^2$$

- Adam

$$\theta_{t+1,i} \leftarrow \theta_{t,i} - \frac{\gamma}{\sqrt{\hat{v}_{t,i}^2 + \varepsilon}} \hat{m}_{t,i},$$
$$v_{t,i}^2 \leftarrow \beta_2 v_{t-1,i}^2 + (1-\beta_2)(\nabla_{\theta_i} \ell_t(\theta_t))^2, \qquad \hat{v}_{t,i}^2 \leftarrow \frac{v_{t,i}^2}{1-\beta_2^t},$$
$$m_{t,i} \leftarrow \beta_1 m_{t-1,i} + (1-\beta_1) \nabla_{\theta_i} \ell_t(\theta_t) \qquad \hat{m}_{t,i} \leftarrow \frac{m_{t,i}}{1-\beta_1^t}$$

Формализация обучения

Обозначения:

- Оптимизатор(optimizer) m – параметризован параметрами ϕ
- Оптимизируемый функционал(optimizee) f – зависит от параметров θ

Expected loss:

$$\mathcal{L}(\phi) = \mathbb{E}_f \left[f\left(\theta^*(f, \phi)\right) \right]$$

Параметризация оптимизатора

$$\mathcal{L}(\phi) = \mathbb{E}_f \left[\sum_{t=1}^{T} w_t f(\theta_t) \right] \qquad \text{where} \qquad \begin{aligned} \theta_{t+1} &= \theta_t + g_t \,, \\ \begin{bmatrix} g_t \\ h_{t+1} \end{bmatrix} &= m(\nabla_t, h_t, \phi) \end{aligned}$$

Если взять w_t равным 1 в точке T и в остальных 0, то мы перейдем к формуле с предыдущего слайда.

Параметризация для Adam

$$\mathcal{L}(\phi) = \mathbb{E}_f \left[\sum_{t=1}^{T} w_t f(\theta_t) \right] \qquad \text{where} \qquad \theta_{t+1} = \theta_t + g_t \,,$$

$$\begin{bmatrix} g_t \\ h_{t+1} \end{bmatrix} = m(\nabla_t, h_t, \phi)$$

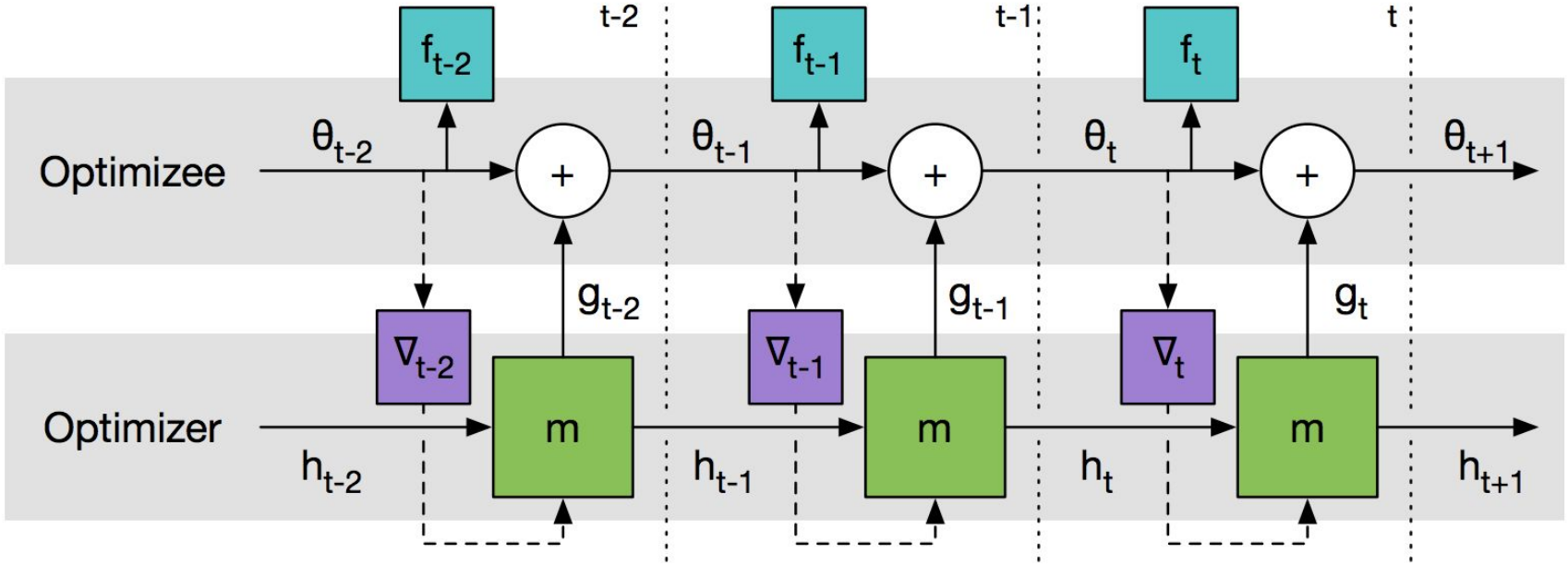
Adam:

$$\theta_{t+1,i} \leftarrow \theta_{t,i} - \frac{\gamma}{\sqrt{\hat{v}_{t,i}^2} + \varepsilon} \hat{m}_{t,i},$$

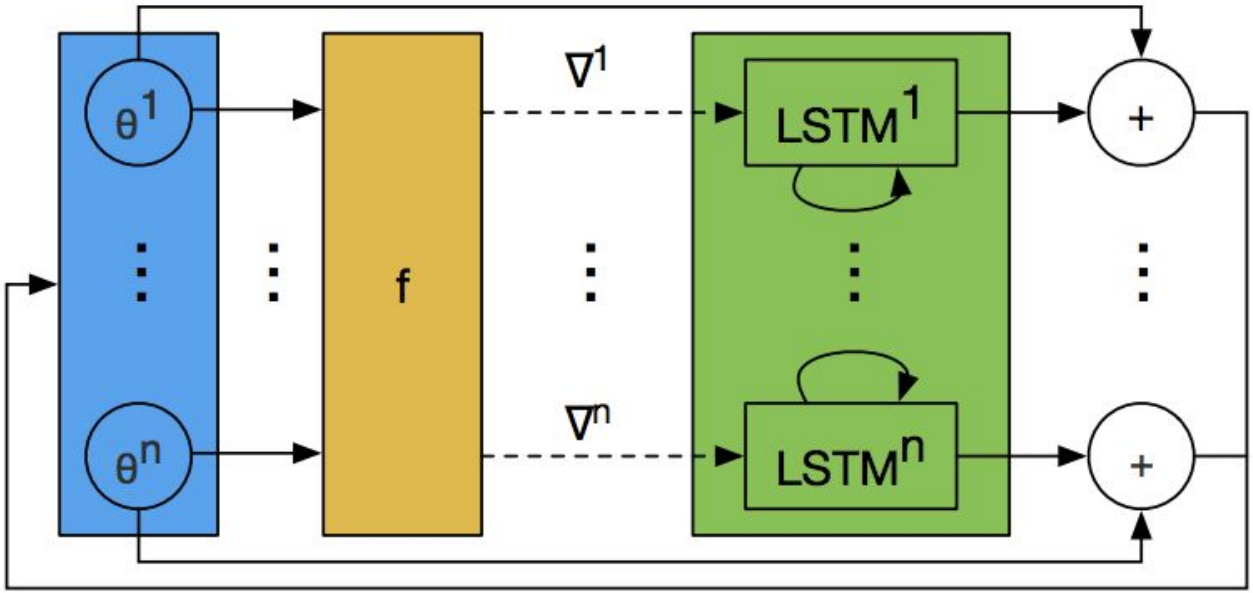
$$v_{t,i}^2 \leftarrow \beta_2 v_{t-1,i}^2 + (1-\beta_2)(\nabla_{\theta_i} \ell_t(\theta_t))^2, \qquad \hat{v}_{t,i}^2 \leftarrow \frac{v_{t,i}^2}{1-\beta_2^t},$$

$$m_{t,i} \leftarrow \beta_1 m_{t-1,i} + (1-\beta_1) \nabla_{\theta_i} \ell_t(\theta_t) \qquad \hat{m}_{t,i} \leftarrow \frac{m_{t,i}}{1-\beta_1^t}$$

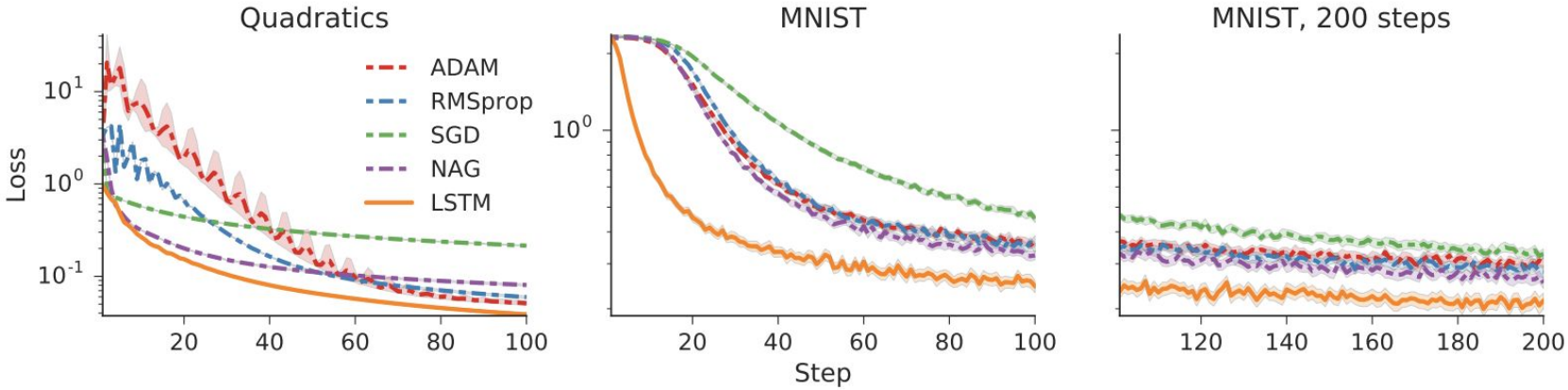
Параметризация при помощи нейросети



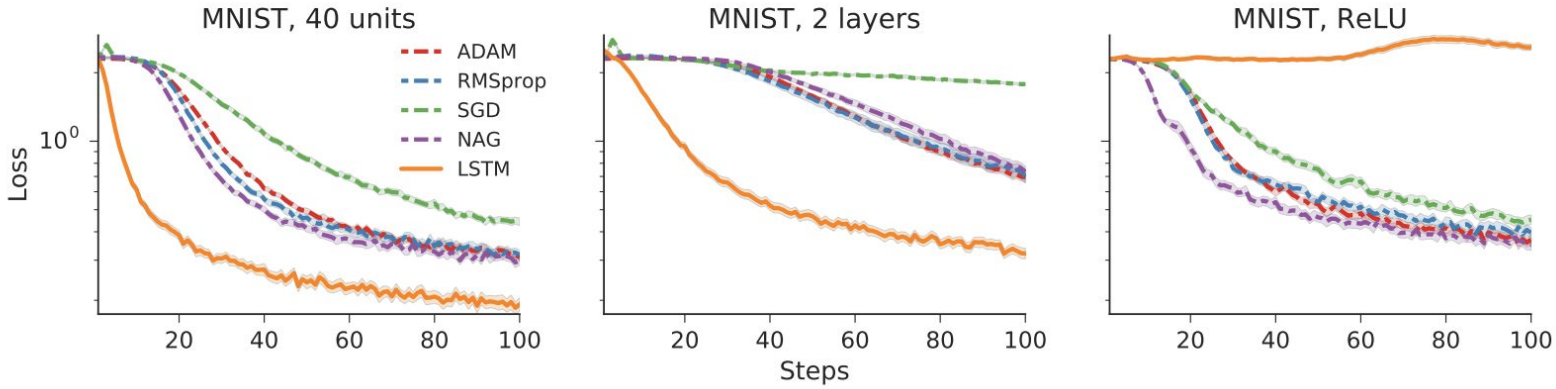
Один шаг оптимизации



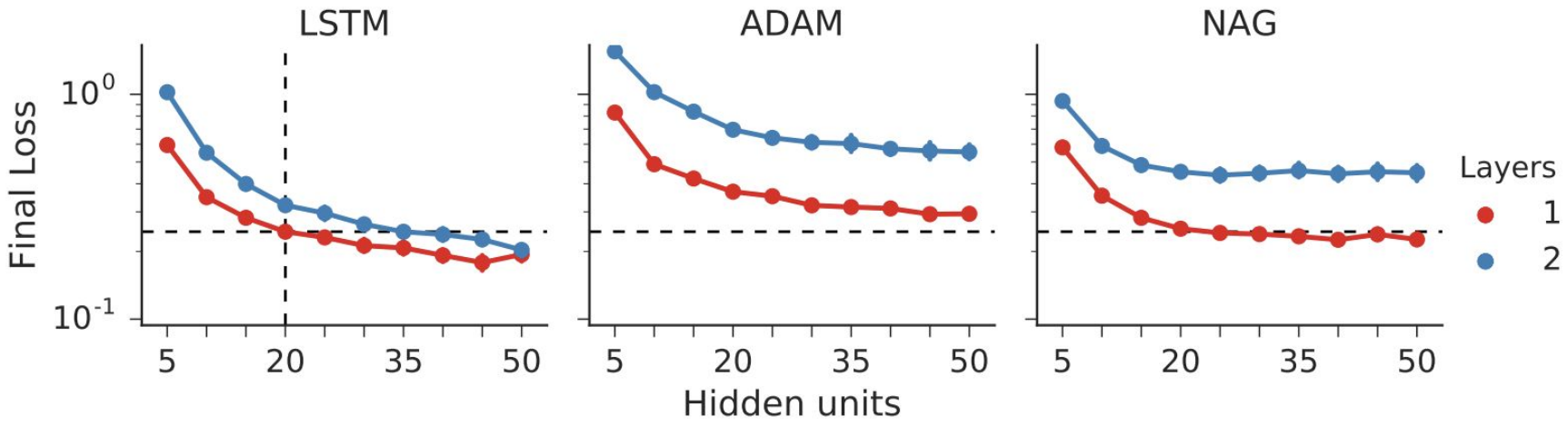
Эксперименты. Задачи



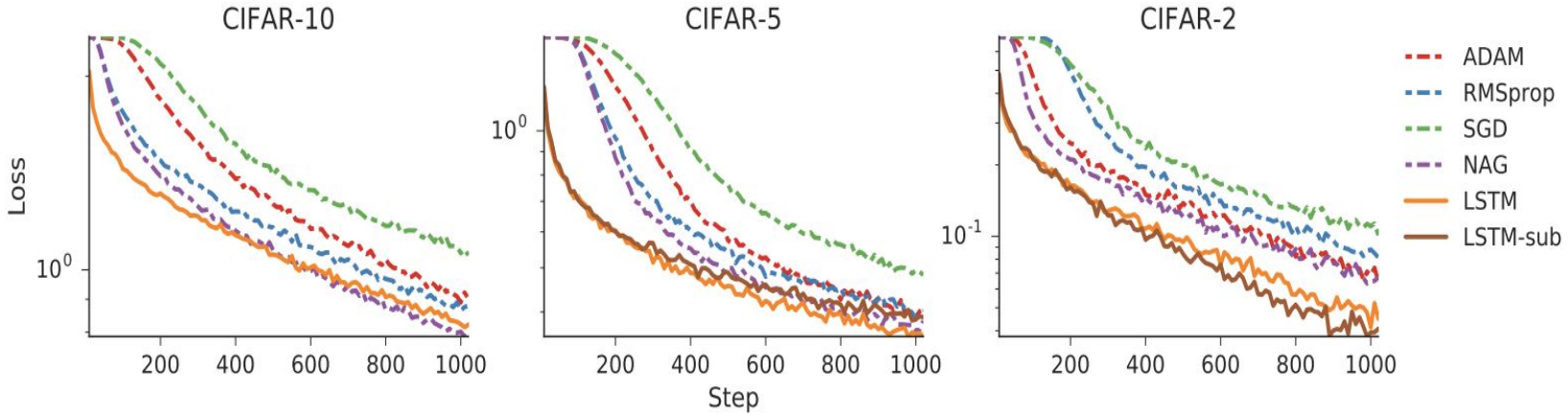
Эксперименты. Обобщающая способность



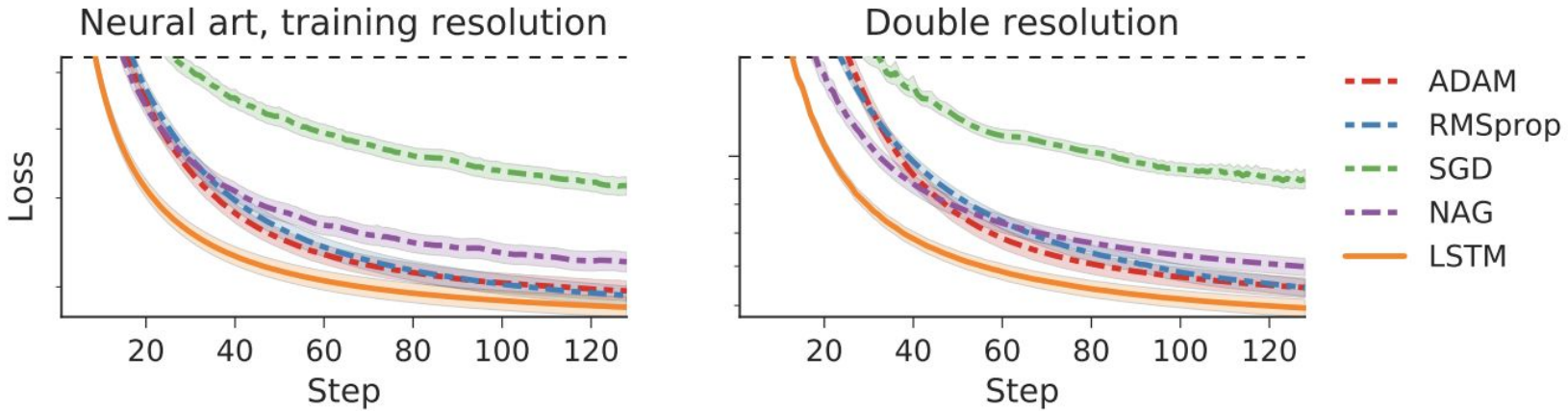
Эксперименты. Обобщающая способность



Эксперименты. Обобщающая способность



Эксперименты. Обобщающая способность



Оригинальная статья

Learning to learn by gradient descent by gradient descent(2016)
https://arxiv.org/pdf/1606.04474.pdf.