VeLO: обзор-рецензия

Асланов Алишер, БПМИ191

Авторы: **Google Research, Brain Tram**

Luke Metz*, James Harrison†, C. Daniel Freeman, Amil Merchant,
Lucas Beyer, James Bradbury, Naman Agarwal, Ben Poole,
Igor Mordatch, Adam Roberts, Jascha Sohl-Dickstein‡

Luke Metz
Meta-learning, GANs

James Harrison
Robotics, RL, meta-learning

C. Daniel Freeman
Meta-learning, RL, Bayesian Inference

Amil Merchant
Meta-learning

Небольшая аналогия

Deep learning

engineering features → learning features

SIFT (Lowe et. al. 1999)
HOG (Dalal et. al. 2005)

LeNet (LeCun et. al. 1998)
AlexNet (Krizhevsky et. al. 2012)

Meta learning

engineering to learn → learning to learn

SGD (Robbins et. al. 1951, Bottou 2010)
Autoencoders (Hinton et. al. 2006)

Learning To Learn (Hochreiter et. al. 2001)
Learned Optimizers (Andrychowicz et. al. 2016, Li et. al. 2016, Wichrowska et. al. 2017, Metz et. al. 2018, 2019)

* Luke Metz, Towards General Purpose Learned Optimizers

Предшествующие статьи

Bengio et al. On the Optimization of a Synaptic Learning Rule (1992) — обучение простых правил обновления параметров для маленьких моделей.

$$\Delta w(i,j) = \theta_0 + \theta_1\, y(i) + \theta_2\, x(j) + \theta_3\, y(mod(j)) + \theta_4\, y(i)\, y(mod(j)) + \theta_5\, y(i)\, x(j) + \theta_6\, y(i)\, w(i,j)$$

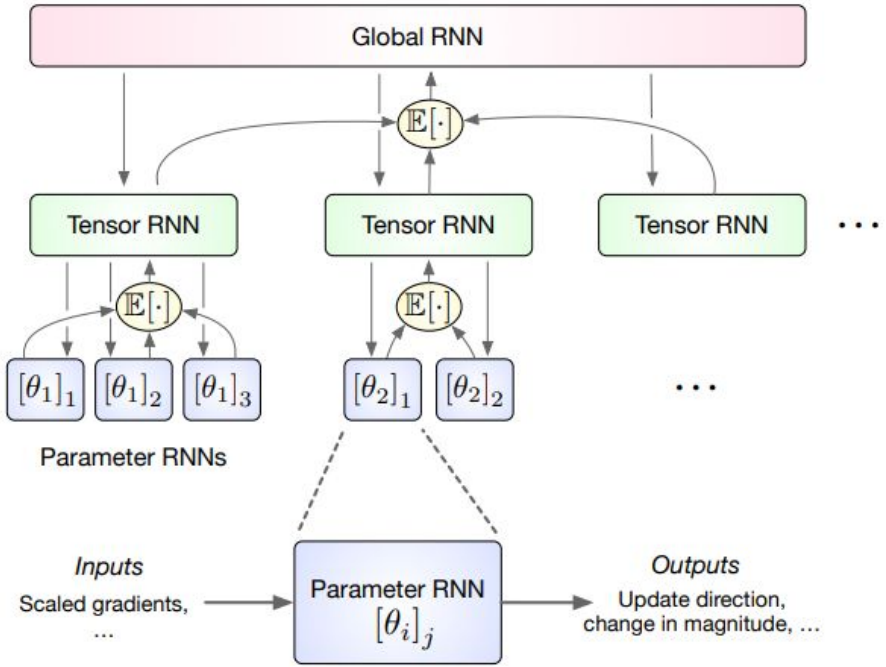
Andrychowicz et al. Learning to learn by gradient descent by gradient descent (2016) — обучение рекуррентных нейросетей для обновления параметров.

- LSTM для предсказания “направления спуска”;
- Эксперименты на выпуклых функциях, MNIST, CIFAR-10;
- Модель получилась лучше стандартных оптимизаторов + есть обобщающая способность для похожих задач и чуть бОльших моделей (масштаб ~ 50 тыс. параметров).

Предшествующие статьи

Wichrowska et al. Learned Optimizers that Scale and Generalize (2017) — первая работа с практически аналогичной архитектурой оптимизатора.

- Простые (синтетические) оптимизационные задачи в мета-обучающей выборке;
- На вход тензорной части подается только среднее, в отличие от VeLO, где также подаются другие статистики;
- Обобщающая способность для MLP, CNN (которых не было в мета-обучающей выборке!).



Предшествующие статьи

Metz et al. Tasks, stability, architecture, and compute: Training more effective learned optimizers, and using them to train themselves (2020) — более ранняя работа от авторов VeLO.

- Здесь meta-loss = average loss. В статье про VeLO авторы пишут, что это работает хуже, по сравнению с VeLO (где meta-loss = значение лосса на последней итерации).

Metz et al. Practical tradeoffs between memory, compute, and performance in learned optimizers (2022) — еще одна недавняя работа от авторов VeLO.

- Здесь MLP-часть обучаема и также принимает на вход скрытые представления от LSTM-части, отчего становится больше => ее дольше обучать.

Достоинства и недостатки работы

Достоинства:

- Практическая применимость для задач “среднего” размера с качеством, не уступающим стандартным методам оптимизации (а иногда даже превосходящим их!);
- Открытый исходный код;
- Множество различных экспериментов с задачами, не входящими в мета-обучающую выборку.

Недостатки:

- Мало исследуется обобщающая способность моделей, обученных с помощью VeLO;
- Неприменимость при большом числе итераций оптимизации и/или большом количестве параметров.

Идеи для улучшения?

- А получится ли адаптировать идею обучаемых оптимизаторов для методов нулевого порядка? Звучит не очень реалистично — как это, обучать нейросети без backpropagation!?

- А получится ли адаптировать идею обучаемых оптимизаторов для методов *второго* порядка? Возможно, но считать гессианы может быть дорого.

- А что если положить meta-loss = validation loss? Кажется это не вполне корректно, т.к. все же мы стремимся оптимизировать ошибку на обучении. Но с другой стороны мы хотим, чтобы наши модели хорошо обобщали (что далеко не всегда следует из низкой ошибки на обучении).

Еще немного источников

- Простое объяснение статьи про VeLO
- Еще одна работа по теме от тех же авторов