# Экспериментальное исследование современных методов оптимизации, учитывающих кривизну функции потерь, для обучения нейронных сетей

Кузнецова Мария Павловна

Научный руководитель д.ф-м.н., профессор А. Ф. Измаилов

МГУ им М.В.Ломоносова

2 июня 2023 г.

# Цели работы

## Основная цель

Численное сравнение и тестирование современных методов оптимизации для обучения нейронных сетей на задачах классификации.

**Поставленные задачи**

- Программно реализовать обучение нейронных сетей с разными архитектурами с помощью разных методов.
- Визуализировать результаты работы разных методов и сделать выводы.

# Постановка задачи классификации

Пусть $X \subset \mathbb{R}^n$ — конечное множество векторов (описаний объектов), $Y = \{1, \ldots, c\}$ — конечное множество допустимых ответов. Совокупности пар $(x_i, y_i)$ объект-ответ $D^k = \{(x_i, y_i) \in X \times Y | i = \overline{1, k}\}$ и $D^h = \{(x_i, y_i) \in X \times Y | i = \overline{k+1, m}\}$ называются обучающей выборкой и контрольной выборкой соответственно, где $h = m - k$.

Пусть $P^c = \{p \in \mathbb{R}^c : p_1 + \ldots + p_c = 1, p_i \geq 0, i = \overline{1, c}\}$, $f(\theta, \cdot) : \mathbb{R}^n \to P^c$ — функция нейронной сети, где $\theta$ представляет собой вектор, состоящий из всех параметров (весовых коэффициентов) нейронной сети, объединенных вместе.

# Постановка задачи классификации

Функция потерь — это неотрицательная функция $L(\theta, D^k)$, характеризующая величину ошибки функции нейронной сети при данном наборе параметров $\theta$ на обучающей выборке.

## Задача обучения нейронной сети

$$L(\theta, D^k) \to \min_{\theta}. \tag{1}$$

Для анализа результатов обучения исследовались значения функции потерь на обучающей и контрольной выборках, а также точность (доля правильных ответов) на тестовой выборке.

# Методы оптимизации

В процессе обучения на шаге $t$ случайным образом выбирается часть обучающей выборки (батч), и с использованием её делается шаг. Эпоха — последовательность шагов, в ходе которой используются все батчи.

## Стохастический градиентный спуск (SGD)

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta L(\theta_t, D_t^k), \tag{2}$$

$\alpha$ — шаг обучения.

## EvoLved Sign Momentum (Lion)

$$\theta_{t+1} = \theta_t - \alpha(\text{sign}(c_t) + \lambda\theta_t), \tag{3}$$

$c_t$ — это оценки момента первого порядка градиента, $\lambda$ — коэффициент регуляризации.

## Метод приближённой кривизны с учетом факторизации Кронекера (K-FAC)

K-FAC является аппроксимацией метода натурального градиента, шаг которого определяется как

$$\theta_{t+1} = \theta_t - \alpha F^{-1} \nabla_\theta L(\theta, D_t^k), \tag{4}$$

где через $F$ обозначается так называемая информационная матрица Фишера (которая используется для вычисления ковариационных матриц, связанных с оценками максимального правдоподобия).

## Метод минимизации с учётом остроты (SAM)

В основе данного метода лежит идея поиска параметров, в некоторой окрестности которых функция потерь имеет значения близкие к минимальным. То есть ищутся неострые минимумы.

# Методы оптимизации

## Предобусловленная стохастическая оптимизация (Shampoo)

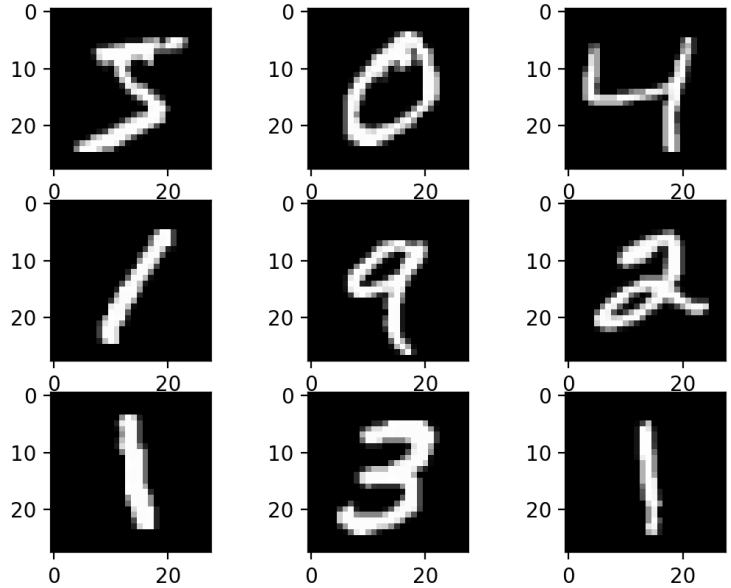
Данный метод описывает алгоритм, в котором вместо одного предобуславливателя предлагается использовать набор предобуславливателей, каждый из которых работает с одним измерением.

## Планировщик скорости обучения

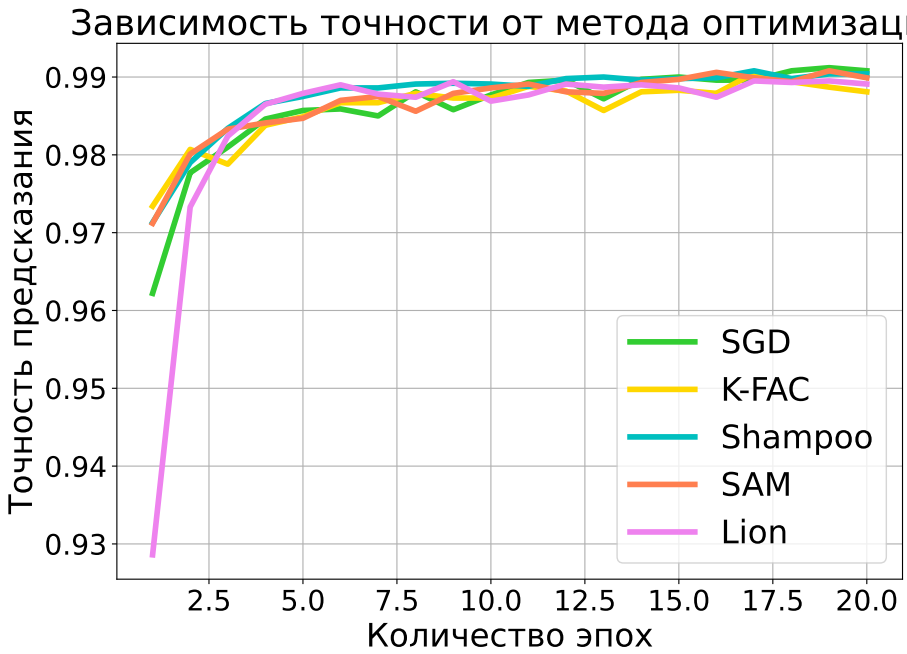
Планировщик скорости обучения регулирует шаг $\alpha$ в процессе обучения. Уменьшение шага обучения с каждой эпохой позволяет улучшить сходимость.

# Задача классификации цифр

**Данные**: чёрно-белые изображения цифр от 0 до 9 (MNIST).
**Нейронная сеть**: свёрточная, 6-ти слойная (LeNet5).



# Эксперименты для задачи классификации цифр



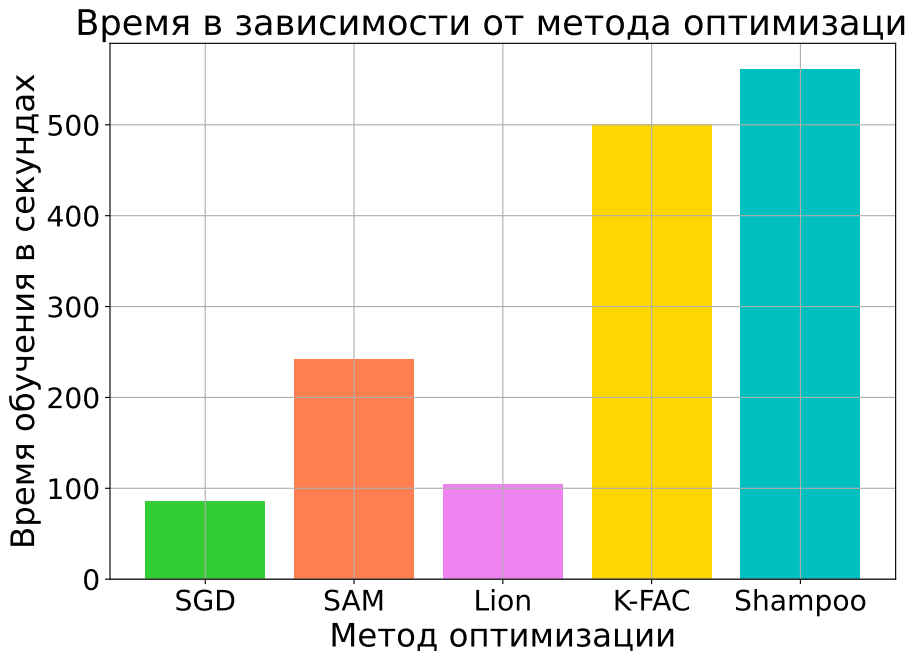
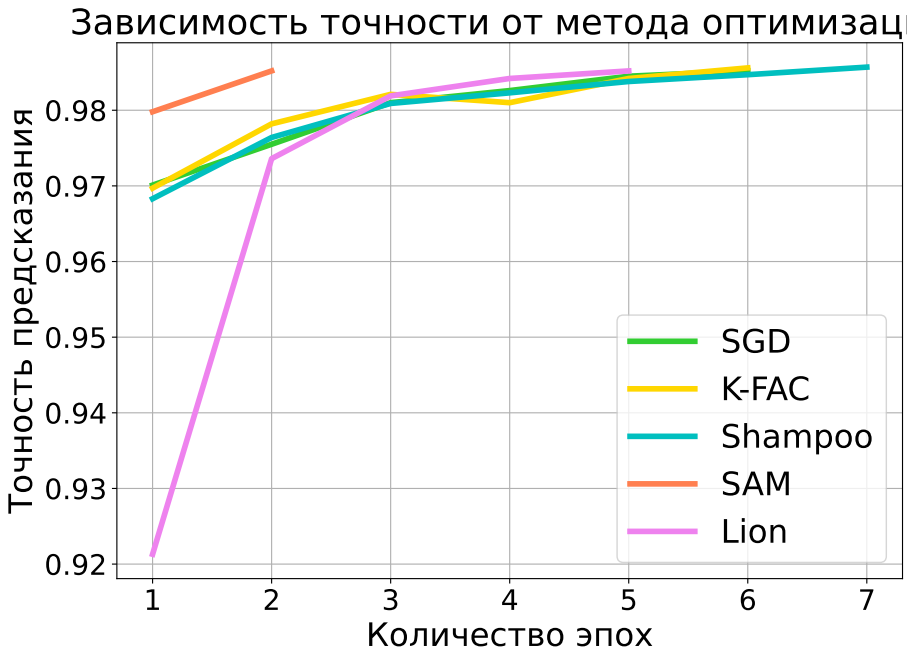




Рис.: Эксперименты с обучением нейронной сети разными методами в течение 20 эпох.

# Эксперименты для задачи классификации цифр



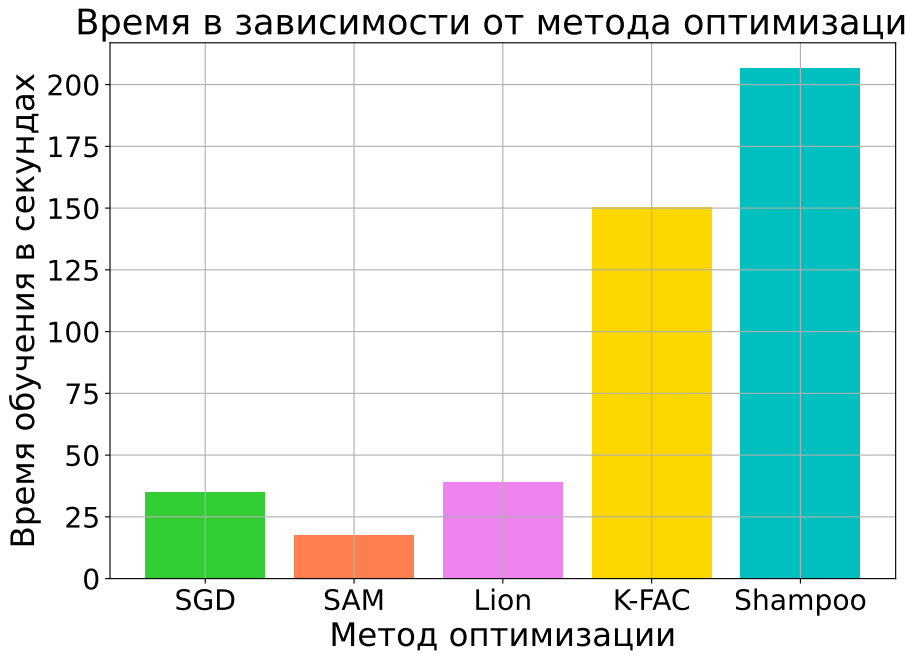
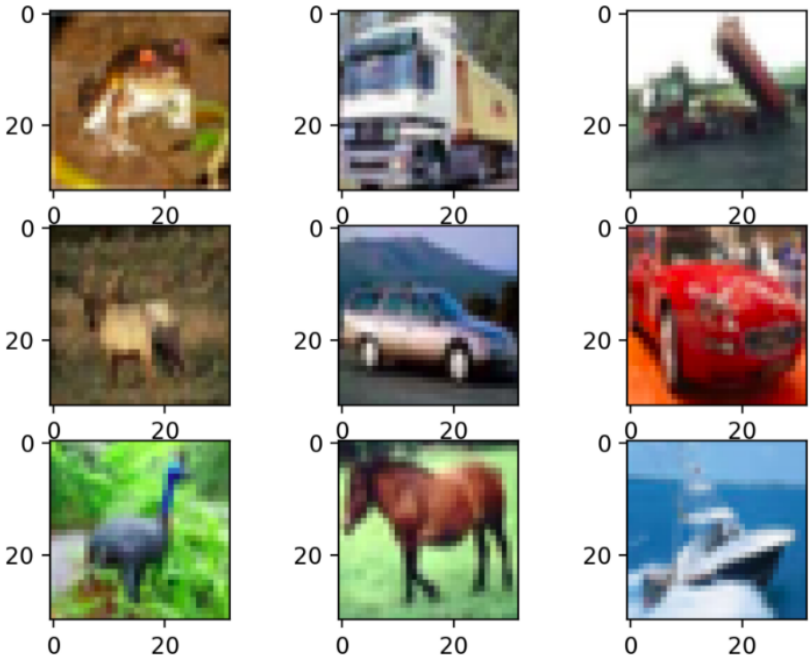




Рис.: Эксперименты с обучением нейронной сети до точности 0.985 разными методами.

# Задача классификации цветных изображений

**Данные**: цветные изображения из 10 классов (CIFAR-10).
**Нейронная сеть**: свёрточная, 18-ти слойная (ResNet-18).



# Эксперименты для задачи классификации цветных изображений

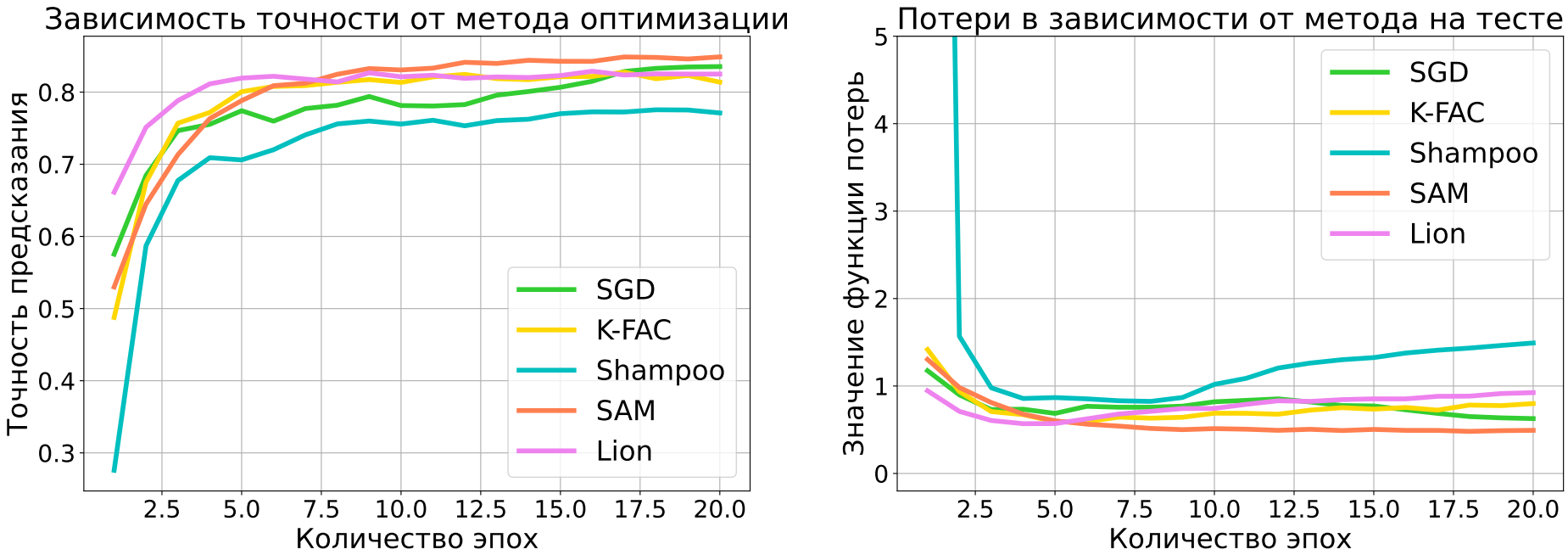



Рис.: Эксперименты с обучением нейронной сети разными методами в течение 20 эпох.

# Эксперименты для задачи классификации цветных изображений

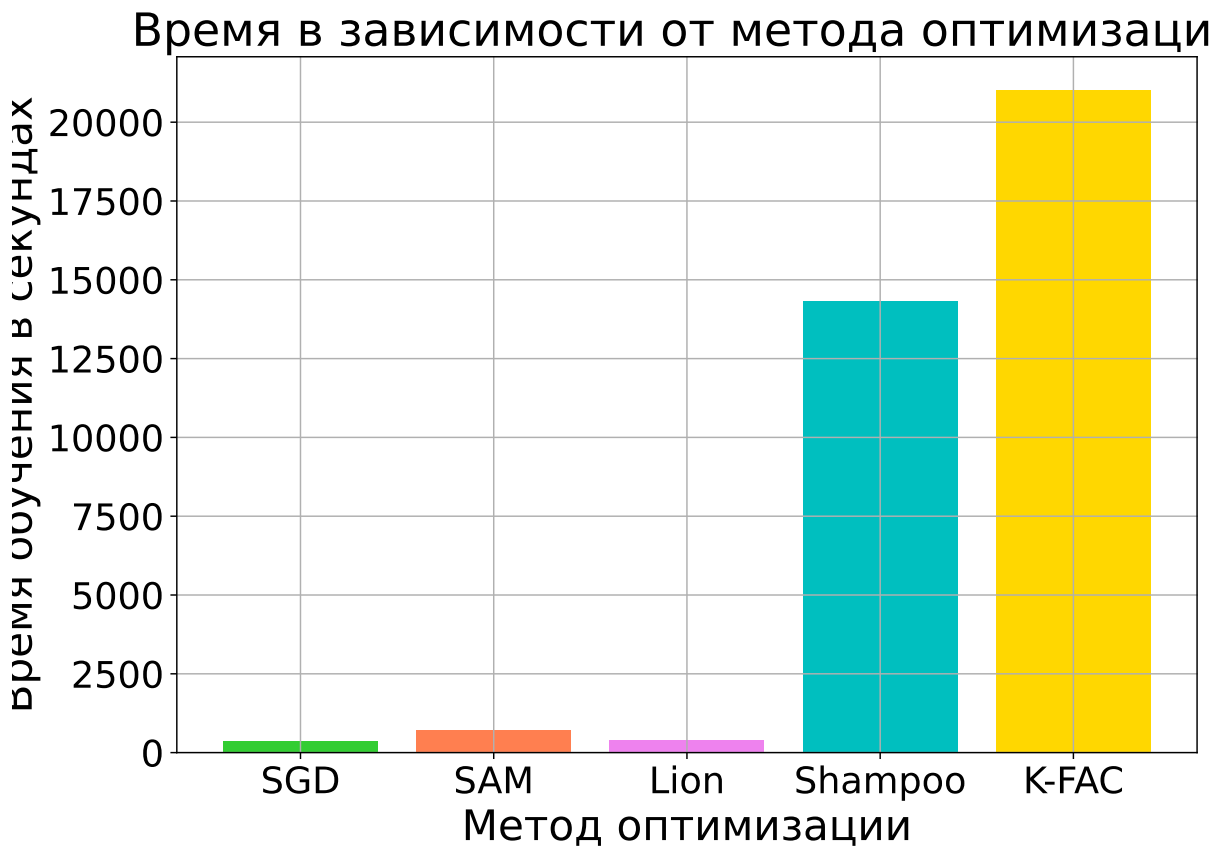



Рис.: Диаграмма времени, потраченного на обучение нейронной сети в течение 20 эпох с помощью разных методов.

# Эксперименты для задачи классификации цветных изображений

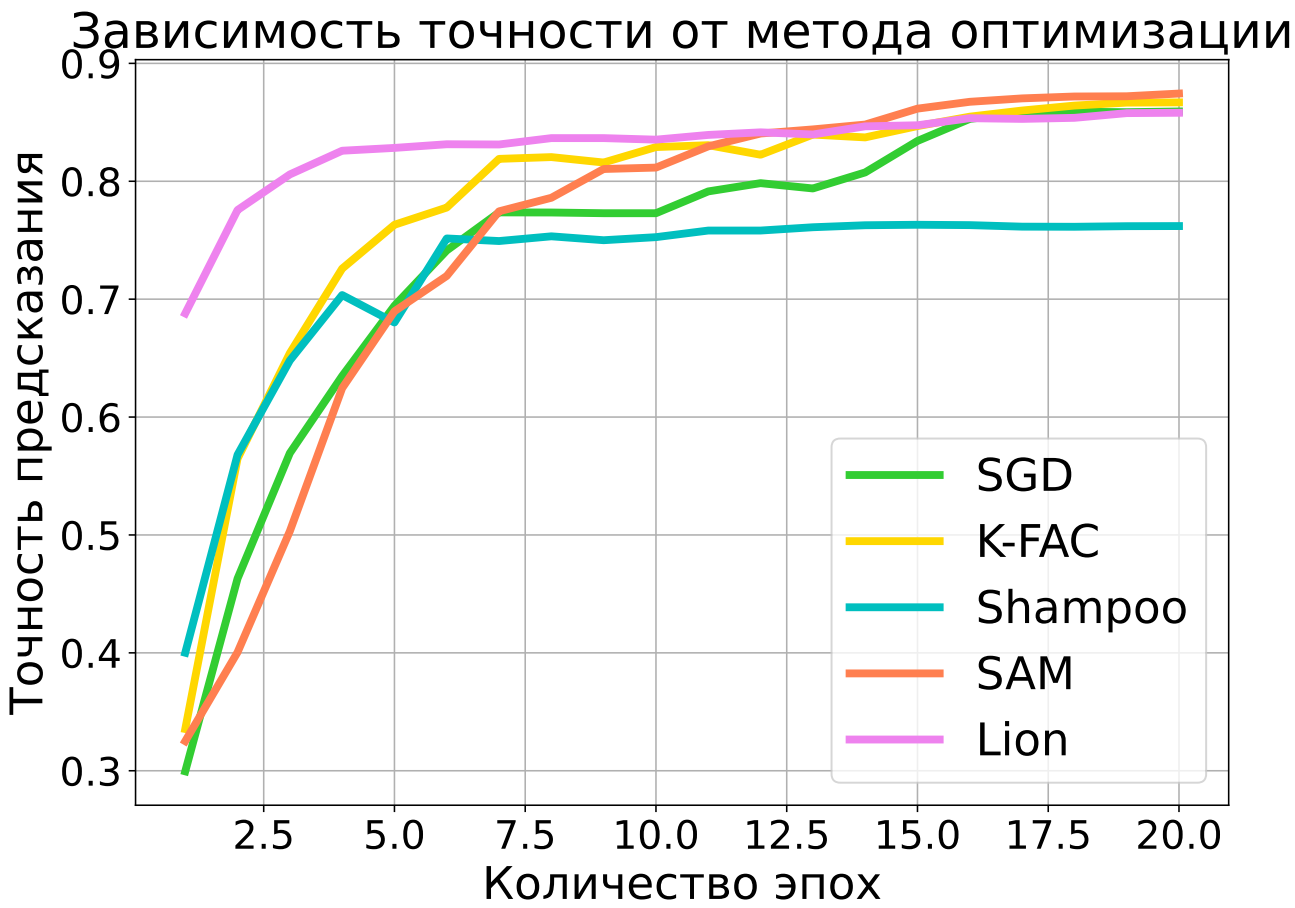



Рис.: Эксперименты с планировщиком скорости обучения.

# Задача классификации именованных сущностей

**Данные**: предложения на английском языке (CoNLL).
**Нейронная сеть**: рекуррентная, 2-х слойная.

# Эксперименты для задачи классификации именованных сущностей

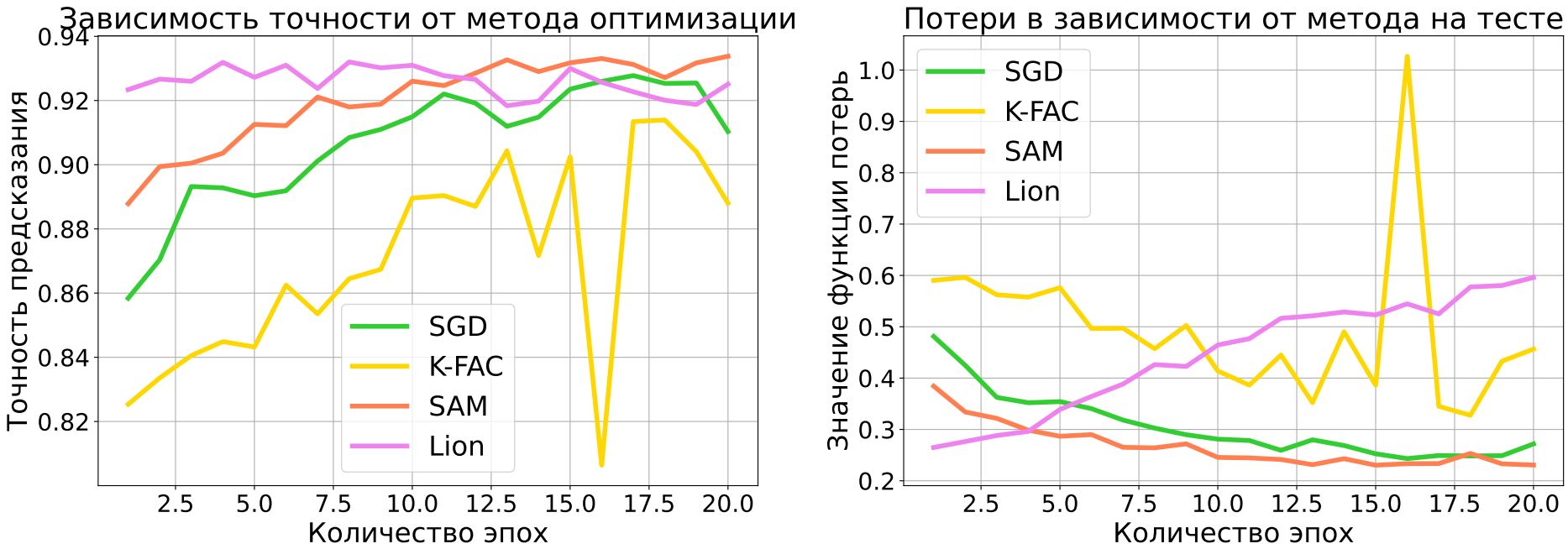
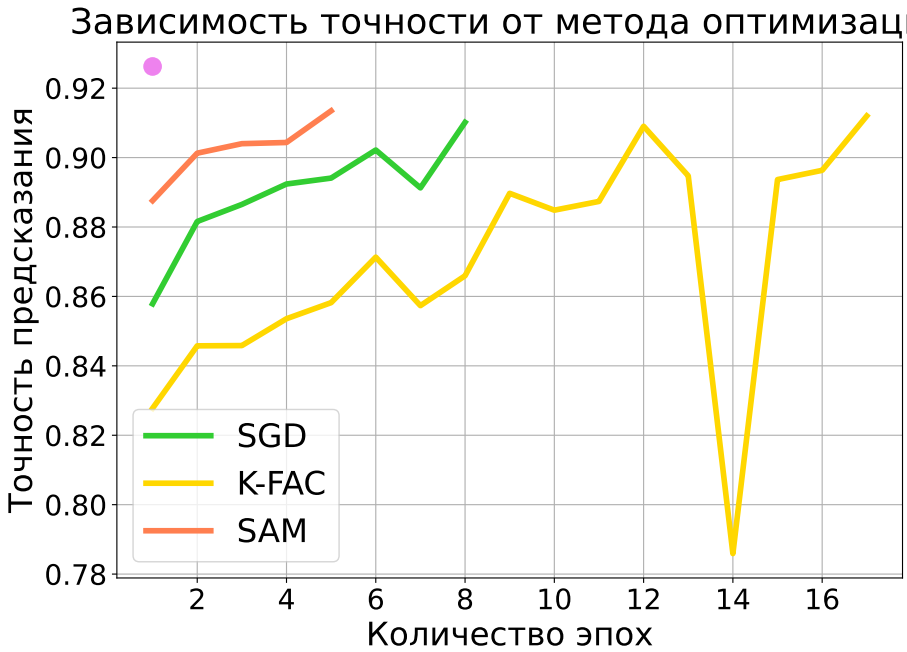



Рис.: Эксперименты с обучением нейронной сети разными методами в течение 20 эпох.

# Эксперименты для задачи классификации именованных сущностей



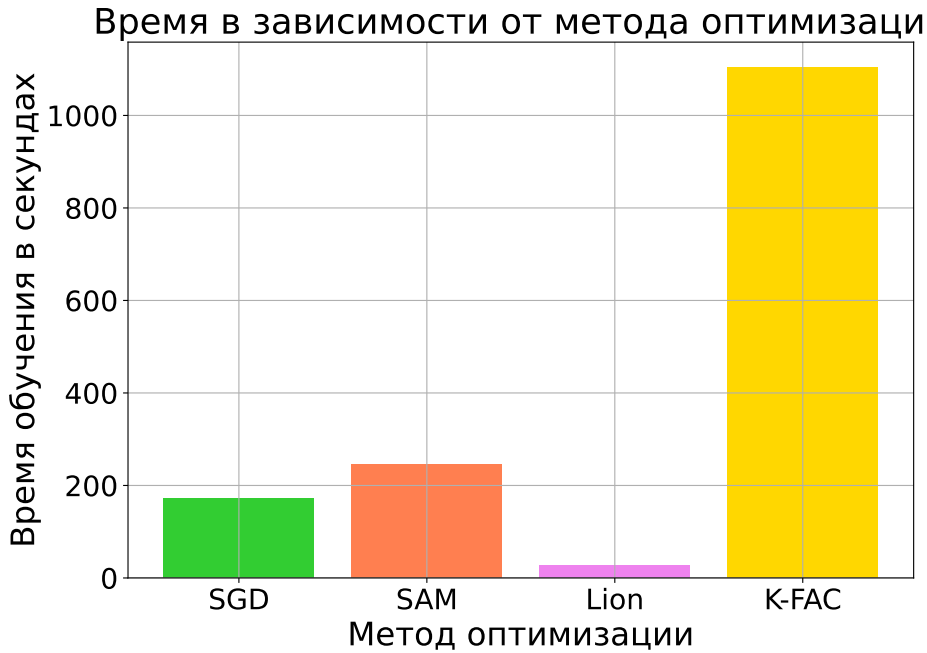




Рис.: Эксперименты с обучением нейронной сети до точности 0.91 разными методами.

# Результаты работы

1. **Shampoo**: низкие результаты, большие временные затраты.
2. **K-FAC**: хорошие результаты с свёрточными нейронными сетями, плохие с реккурентными, большие временные затраты.
3. **Lion**: высокие результаты с первых эпох, время сравнимо с временем работы SGD, переобучение.
4. **SAM**: лучшие результаты с нейронными сетями разных типов, нет переобучения, затраты времени в 2 раза больше чем при использовании SGD.

# Литература


- **Shampoo**: Gupta V., Koren T., Singer Y. 2018.
- **K-FAC**: Martens J., Grosse R. B. 2015.
- **Lion**: Chen X., Liang C., Huang D., Real E., Wang K., Liu Y., Pham H., Dong X., Luong T., Hsieh C., Lu Y., Le Q. 2023.
- **SAM**: Foret P., Kleiner A., Mobahi H., Neyshabur B. 2021.