# Рецензия на рукопись

*“Adaptation of the architecture of the deep learning model with performance control”*

Автор  Бородин С.С.

В статье описывается метод прунинга нейросетей, учитывающий технические возможности целевого аппаратного обеспечения. Исследуется зависимость качества модели от регуляризации относительно времени работы. Теоретически доказано существование нетривиального решения и проведены эмпирические исследования по его поиску. В качестве оптимизируемой модели использована ResNet18, эксперименты проводятся с помощью датасета CIFAR-10.

## Название

Предложил бы более “острое” название, содержащее больше существительных, например “*Adaptive Deep Learning Model Architecture With Performance Control*"

## Аннотация

- “with respect to **the** target hardware…”
- “computational resources" вместо «computation budget”

## Введение

- “Since **the** introduction of…”
- “without **suffering the problem of** vanishing/exploding…”
- “We measure latency of **the** model”
- “**Finally**, we ensure that original and relaxed problems **are equivalent**.”

## Постановка задачи

- V не определена. В целом, можно догадаться, какой смысл имеет обозначение, но хотелось бы увидеть расшифровку.
- Связь между \tilde{\gamma} и \tilde{\gamma}_e тоже понятна, но лучше показать, что \tilde{\gamma}_e - это компонента \tilde{\gamma}.
- “We will call a computational subgraph not broken if there **is a** path in the…” (общее замечание: лучше “there exists” заменить на “there is a”)
- “Hence, sufficient condition for **a** graph to be not broken is non-constant output.”
- В теореме не определена T. Лучше обозначить связь со временем операции T(e).

## Вычислительный эксперимент

- Хотелось бы немного больше услышать про RelaxedBernoulli (хотя бы о том, что это гладкая аппроксимация распределения Бернулли)
- График №2 слегка нерепрезентативен: на нем изображены accuracy и latency, посоветовал бы разделить графики между собой (и указать параметр по оси Oy).

Помимо этого хотелось бы увидеть больше значений на оси Ox, так как в тексте обращаются к параметру \lambda = 5, но его на графике нет.
- Предлагаю в Таблице №2 acceleration также указывать в процентах.
- Хотелось бы увидеть, как метод ведет себя для разных latency (на примере различных устройств). То есть продемонстрировать, что модель, адаптированная для одного оборудования, показывает на нем качество лучше, чем на другой платформе.

### Заключение

- “**We managed to achieve** better accuracy for the same latency and better…”
- Можно добавить информацию о том, в каком направлении может развиваться исследование в дальнейшем

### Общие замечания к тексту

- Посоветовал бы поправить пунктуацию, в некоторых местах не хватает запятых.
- Артикли (особенно неопределенные) тоже необходимо добавить

### Замечания к коду

Код имеет удобную структуру и хорошо оформлен. Есть только несколько предложений:

- В ноутбуках с экспериментами не хватает markdown-ячеек с заголовками и описанием процесса, было бы неплохо их добавить
- Чтобы ячейки с установкой библиотек не занимали много места, можно воспользоваться *IPython.utils.io.capture_output()*
- Чтобы графики не накладывались друг на друга, можно использовать *plt.tight_layout()*
- Предлагаю вынести файлы из */code/src* в */src*

### Вывод

Никаких серьезных замечаний к статье нет, текст хорошо структурирован, полученные результаты являются новыми и актуальными, они в полной мере объяснены и подкреплены соответствующими аналитическими соображениями. Оценка строго положительная, очень рекомендую статью к публикации.

*Рецензент:*
Овчаренко К.А.