В статье предлагается метод адаптации больших языковых моделей путем добавления обучаемых малоранговых матриц к предобученным весам. При количестве обучаемых параметров в 10000 раз меньшем, чем в дообучаемой модели, на задачах NLG и NLU удается добиться качества, превосходящего полный файн-тюнинг.

Сильные стороны:

- Представлены актуальные и практически применимые результаты.
- Представлен обширный обзор работы по смежным направлениям, предложенный метод показан как логичное продолжение предшествующих исследований.
- Представлены не только результаты сравнения предложенного метода с существующими, но также и результаты совместного применения LoRA и prompt tuning.
- Приведены все гиперпараметры, необходимые для воспроизведения результатов, исходный код также выложен в открытый доступ.
- Статья написана чисто, подробно и последовательно, её приятно читать.

Слабые стороны:

- Нет экспериментов на SuperGLUE, хотя их следовало бы провести для подтверждения эффективности метода на NLU задачах. В частности, было бы очень интересно увидеть результаты генеративных моделей, дообученных с помощью LoRA на SuperGLUE в условиях, близких к [1], поскольку там приводится сильный результат на T5 и возникает естественный вопрос, нельзя ли улучшить его с помощью LoRA. В статье на этот вопрос ответа нет, так как приводятся только эксперименты с RoBERTa и DeBERTa на GLUE, при этом в первой версии не было даже их.
- Нет честного сравнения с SOTA prompt tuning методами. В частности, упоминается их нестабильность, хотя в [3] было показано, что можно добиться более стабильного обучения путем репараметризации обучаемых эмбеддингов с помощью LSTM. В [1] также показано, что можно обойтись даже без репараметризации, проинициализировав обучаемые эмбеддинги эмбеддингами из словаря. Оба подхода игнорируются и сравнение ведется с оригинальным prefix tuning из [2].

Комментарии: Поскольку LoRA проигрывает своему главному конкуренту – prompt tuning – в оверхеде на переключение между задачами, то также было бы интересно увидеть исследование по снижению этого оверхеда, например, путем снижения количества измененных весов после обучения.

Кроме этого не хватает экспериментов по совмещению LoRA с prompt tuning для одновременного повышения качества и сохранения возможности обслуживать огромное количество задач одной моделью.

Оценка: В статье отсутствуют важные эксперименты, которые бы позволили точнее позиционировать её среди конкурирующих методов. Тем не менее, значительность вклада несомненна, и я считаю, что статья заслуживает оценки **8**. Уверенность оцениваю как **4**.

Список литературы

[1] Brian Lester, Rami Al-Rfou и Noah Constant. “The Power of Scale for Parameter-Efficient Prompt Tuning”. В: *arXiv:2104.08691 [cs]* (апр. 2021). URL: http://arxiv.org/abs/2104.08691.

[2] Xiang Lisa Li и Percy Liang. “Prefix-Tuning: Optimizing Continuous Prompts for Generation”. В: *arXiv:2101.00190 [cs]* (янв. 2021). URL: http://arxiv.org/abs/2101.00190.

[3] Xiao Liu и др. “GPT Understands, Too”. В: *arXiv:2103.10385 [cs]* (март 2021). URL: http://arxiv.org/abs/2103.10385.