Название статьи (авторы статьи): **Beyond Fully-Connected Layers with Quaternions: Parameterization of Hypercomplex Multiplications with 1/n Parameters (Aston Zhang, Yi Tay, SHUAI Zhang, Alvin Chan, Anh Tuan Luu, Siu Hui, Jie Fu)**
Автор рецензии: Иван Сафонов

1. **Краткое описание статьи:** в статье предлагается способ уменьшения количества параметров линейного слоя в n раз (без значительного уменьшения качества и увеличения времени работы). Он обобщает метод сжатия в 4 раза с помощью кватернионного умножения. Авторы показали применимость метода для сжатия рекуррентных сетей и трансформера.
2. **Сильные стороны:**
    - Придумано интересное обобщение метода, использующего кватернионы. При n=4 представленный метод с теоретической точки зрения не хуже (то есть является обобщением).
    - Показано, что метод сжатия дает небольшое уменьшение качества при n кратном уменьшении количества параметров (на NLP задачах).
    - С помощью метода можно настраивать во сколько раз уменьшать количество параметров.
    - С помощью метода можно увеличить матрицы внутри трансформера, а затем уменьшить количество параметров методом в n раз (тем самым получить модель с тем же количеством параметров), и это может дать улучшение качества относительно изначального транформера.
    - Статью легко читать, все математические выкладки очень хорошо поданы. Идеи были теоретически обоснованы (например переход от метода с кватернионами к именно такому методу).
    - С научной точки зрения представлен иной (не похожий на предыдущие) метод уменьшения количества параметров нейронной сети.
    - Параметры экспериментов достаточно подробно описаны и код выложен в github.
3. **Слабые стороны:**
    - В экспериментах нет сравнения с очевидным и стандартным методом сжатия с помощью SVD (матрица k x d → k x r и r x d).
    - Возможно можно ускорить умножение на PHM матрицу в n раз, потому что она имеет специальный вид. В целом возможно можно записать PHM матрицу в тензорном виде через параметры S, A и это может дать какое-то иное представление метода.
    - В методе параметры матрицы получаются перемножениями обучаемых параметров, что на практике может быть нестабильно. Не хватило комментариев по этому поводу.
    - Эксперименты вызывают вопросы, в целом было много экспериментов с разными NLP задачами, но многие из них кажутся довольно нестандартными; также большой разброс в результатах разных экспериментов вызывает тревогу относительно возможной случайности результатов и их незначимости.
4. **Насколько хорошо написана статья:** статья написана доходчиво, сложностей в понимании не возникло.
5. **Воспроизводимость:** статья написана достаточно подробн, также авторы выложили код экспериментов. При самореализации проблем возникнуть не должно, но есть опасения насчет того, что метод успешно обучится.
6. **Дополнительные комментарии, предложения по улучшению:** хотелось бы добавить другие методы сжатия в эксперименты, а также доверительные интервалы для всех результатов.
7. **Оценка:** 7
8. **Уверенность:** 4