Mask-Predict: Parallel Decoding of Conditional Masked Language Models

1. В чем заключается проблема мультимодальности (multimodality problem), возникающая при неавторегрессионном декодировании последовательностей? Приведите пример.
2. Какие токены маскируются на i-ой итерации алгоритма декодирования в модели Conditional Masked Language Model?
3. Опишите алгоритм декодирования последовательности для модели Conditional Masked Language Models. В чем заключаются его преимущества перед алгоритмами декодирования для авторегрессионных моделей?

On the Discrepancy between Density Estimation and Sequence Generation

1. Есть ли корреляция между логарифмом правдоподобия и BLEU? Для каких моделей?
2. Выпишите формулу логарифма правдоподобия для авторегрессионной модели, формулу априорного гауссовского распределения для модели со скрытыми переменными.
3. Какие модели дают лучшее качество перевода: авторегрессионные или неавторегрессионные (модели со скрытыми переменными)? А какие дают лучшее моделирование распределения?
4. Определим преобразование пары векторов $x, y \in \mathbb R^n$ по следующей формуле $f(x, y) = (x, y \odot \exp(s(x)) + t(x))$, где $s(x)$ и $t(x)$ векторнозначные функции, а $\odot$ обозначает поэлементное умножение векторов. Покажите, что данное преобразование обратимо. Чему равен определитель матрицы Якоби этого преобразования?

Scaling Laws for Neural Language Models

1. Какие тренды были замечены авторами статьи Scaling Laws for Neural Language Models в обучении трансформеров?
2. Как подбирать гиперпараметры трансформера при обучении языковой модели с фиксированным бюджетом вычисления?
3. Какой размер батча авторы Scaling Laws for Neural Language Models называют критическим? От каких параметров он зависит и как его вычислить?
4. Запишите уравнение зависимости целевой функции языковой модели на тестовой выборке от числа параметров модели и количества обучающих данных L(N, D). Как можно использовать эту зависимость на практике?

The Curious Case of Neural Text Degeneration

1. Какие проблемы наиболее характерны для методов декодирования на основе максимизации правдоподобия (например, лучевого поиска), а какие для метода декодирования на основе сэмплирования из распределения (Pure Sampling)?
2. В чем заключается идея подхода Nucleus Sampling? Как Nucleus Sampling модифицирует распределение $p(x_i \mid x_{1:(i-1)}$ для сэмплирования?
3. Что такое закон Ципфа? Какой эксперимент предложили авторы The Curious Case of Neural Text Degeneration на основе этого законы? Какие они сделали выводы?
4. Что измеряет метрика Self-Blue в экспериментах работы The Curious Case of Neural Text Degeneration?

Electra: Pre-Training Text Encoders As Discriminators Rather Than Generators

1. Опишите принцип работы метода Electra.
2. Выпишите функцию потерь для генератора и дискриминатора в модели ELECTRA.
3. Какой оптимальный размер генератора по отношению к дискриминатору подобрали авторы ELECTRA? Почему плохо использовать одинаковые размеры для генератора и дискриминатора?