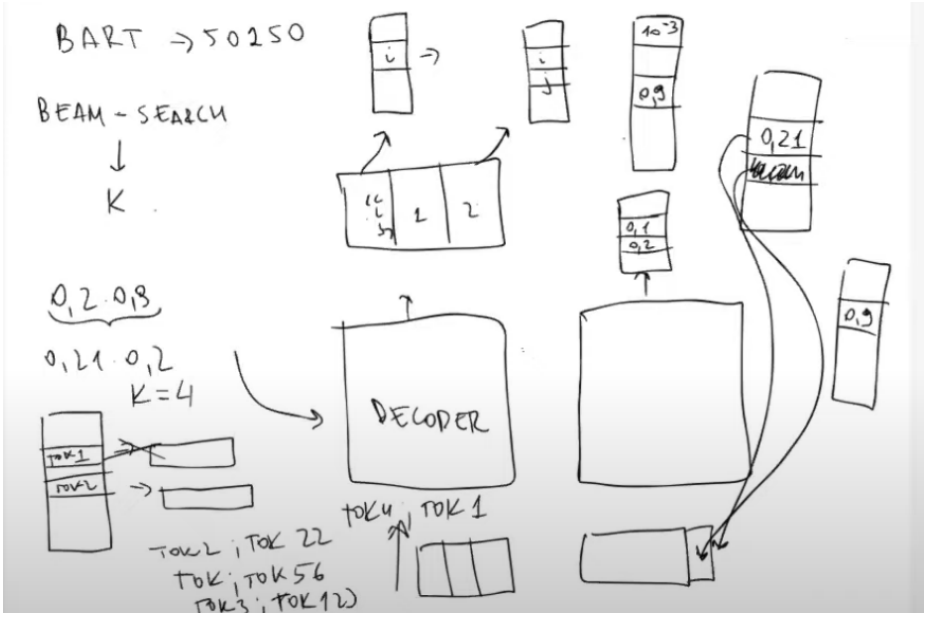
1. Какие два глобальных преимущества transformer-ов над RNN в задаче seq2seq и за счёт чего они возникают? (не теряют информацию так как "вытягивают" её из матрицы, а не из вектора; гораздо быстрее обучаются). Трансформер смотрит на все слова, обучается быстрее так как декодеру подается для оценки весь текст а не по одному
2. Перечислите методы сэмплирования следующего слова decoder-based моделью (argmax (no sampling), probability sampling, sampling if p(word) > x%, cumulative sampling берем столько токенов чтобы вероятность их была больше).
3. Как работает beam-search? За что отвечает параметр?



За кол-во последовательностей которые хранятся, самые оптимальные по кумулятивной вероятностью

4. Какие слои attention-а используются в декодере в Transformer и как они устроены? (masked attention и cross attention)

cross attention мы смотрим и на сгенерированые токеты и на обработаные енкодером
masked attention мы закрываем маской слова которые будут впереди