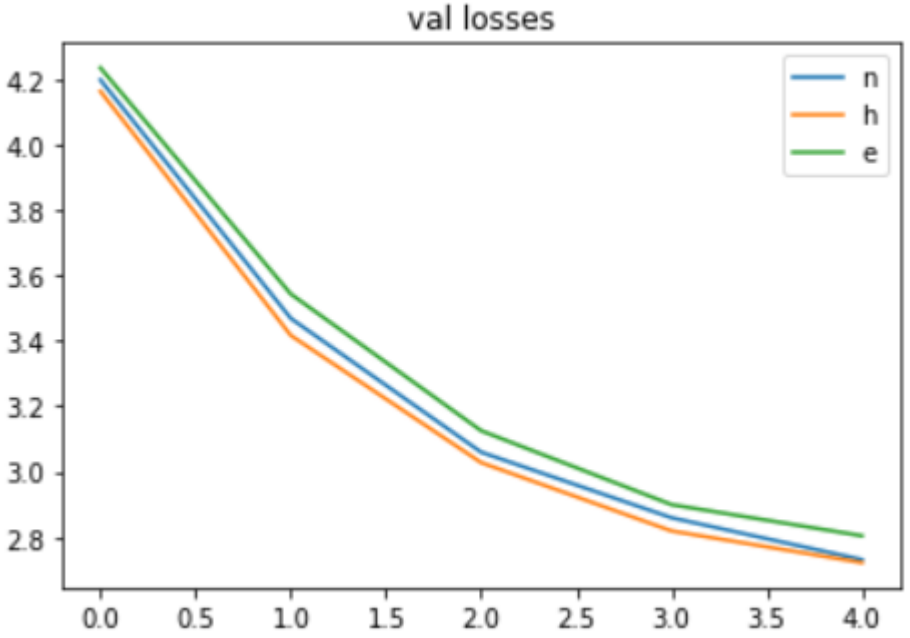
Дисклеймер: я понимаю, что это очев, но как бы пробовать тут много чего не выйдет. Модель - трансформер. Гипотезы: какая-то фича увеличит BLEU на тесте/лосс на вале. Поскольку больше 3 трансформеров не залезает на куду, то результаты будут не всегда все вместе, а скорее попарно сравниваемые (обычно что-то с дефолтным вариантом).

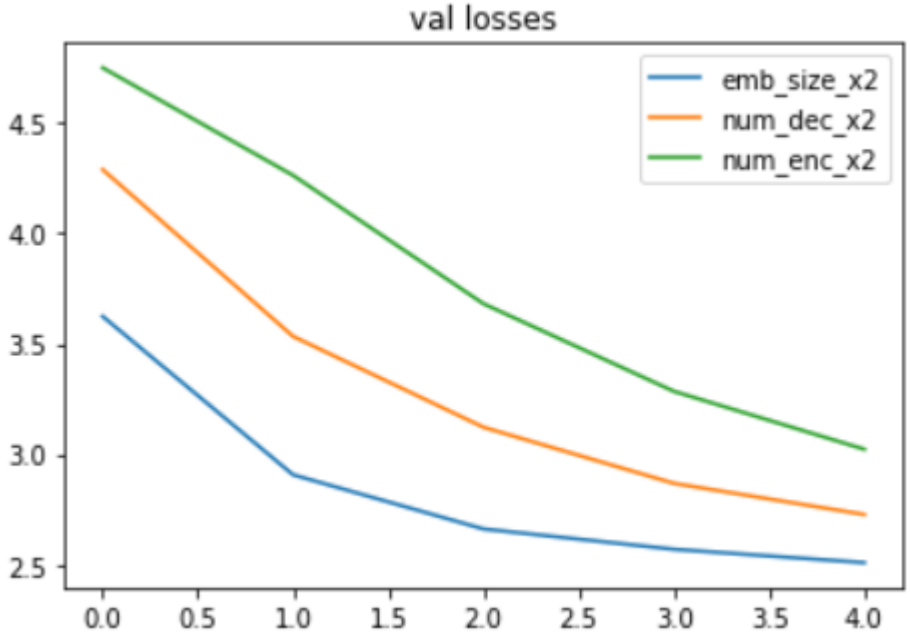
На всех графиках я буду рисовать loss на валидации в зависимости от эпохи. Поэтому оси без подписей)

Пробуем гиперпараметры

1) сначала я сравнил на 5 эпохах дефолтный трансформер (из чекпоинта); тот же, но с nhead в 2 раза больше (x2 nhead); тот же, но hidden_dimention в 2 раза больше. (Легенда на графике вывелась криво, а данные слетели, но порядок такой: синий, оранжевый, зеленый). Но видно, что результаты не сильно разняться, так что оставим дефолтный

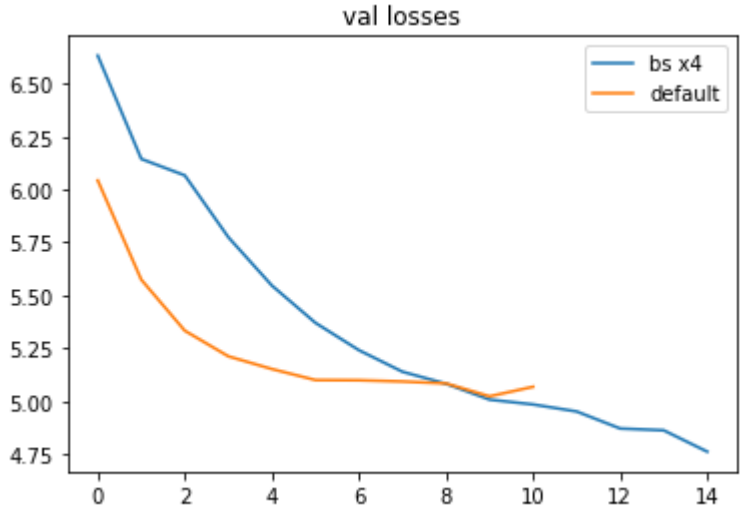


2) Теперь попробуем тоже на 5 эпохах обучить трансформер из чекпоинта в 3 модификациях: x2 embedding size, x2 кол-во декодеров, x2 кол-во энкодеров. Видно, что сильный прирост качества дало только emb size x2. Возьмем тогда эту конфигурацию за базовую модель



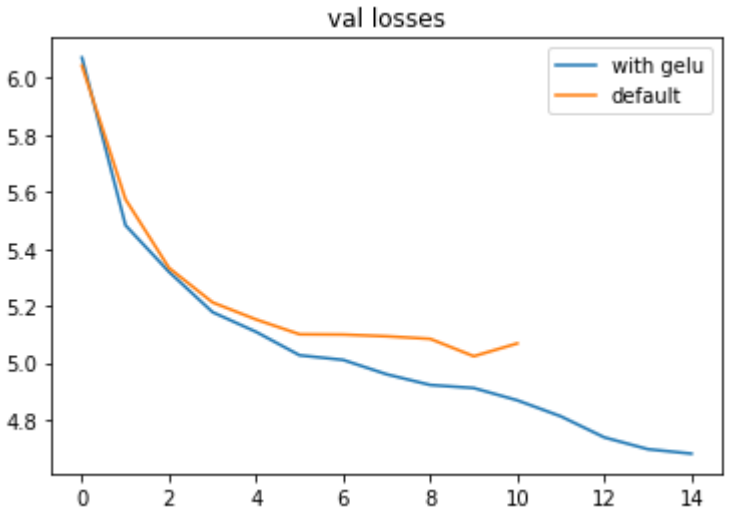
Далее после 2 дней решания проблем с тем, что трансформер не лезет в куду, не хочет просто сохраняться и т.п. я решил сделать трейн-датасет в 10 раз меньше и тестировать гипотезы на нем.

Пробуем batch accumulation
Ну тут нечего объяснять, +3 строчки кода, более долгое обучение, зато явно лучшие результаты



bs x4 - batch size в 4 раза больше

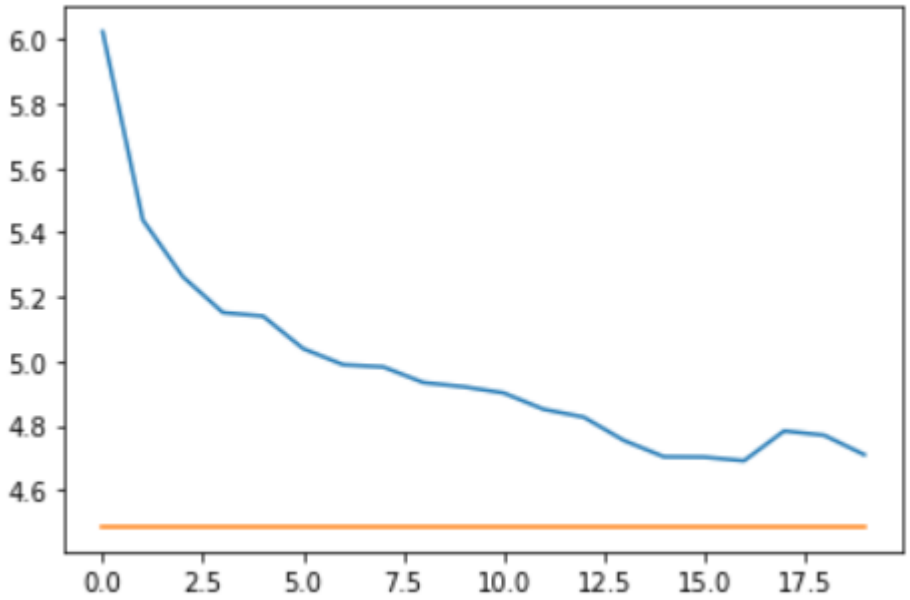
Пробуем gelu



Оказалось, что гелу в качестве активации тоже сильно круче. Но штош, берем.

Пробуем swa

подрубаем swa на последних 25% эпохах обучения (я учил всего 20 эпох)



оранжевый - итоговый результат swa
синий - обучений модели (все на валидации)
Тоже берем

Подбираем scheduler
Я попробовал обычный экспоненциальный с gamma=0.9, дало улучшение качества, поэтому я решил на данном этапе собрать все вместе и обучить.

Бахнул x8 batch accumulation (т.е. по сути 1024 размер батча), swa сделать не смог, так как как вызывать функцию encode для нескольких последних моделей не ясно. А руками уже будет ансамбль. Заслал - +1BLEU :(

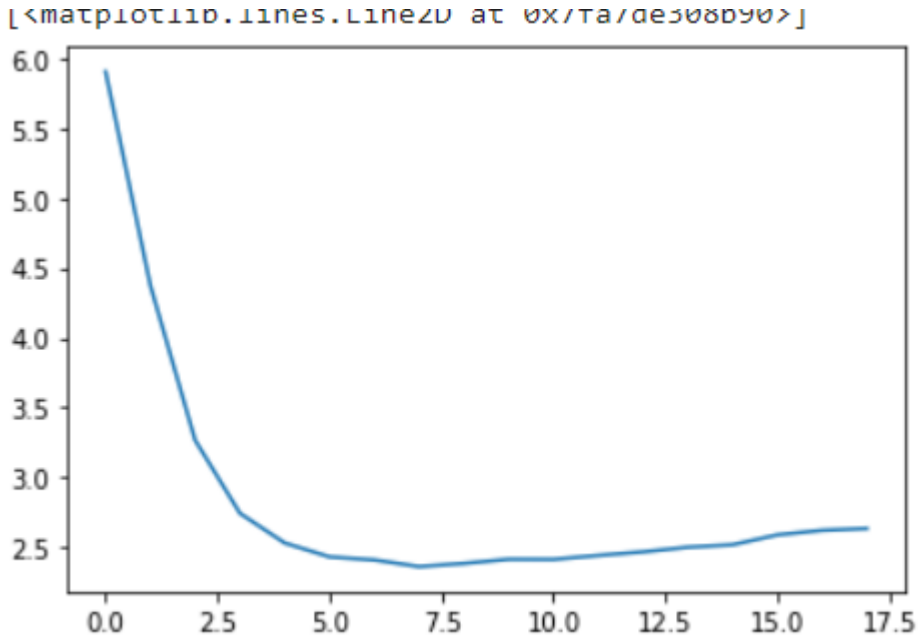
Дальше занимался тем, что пробовал разные расписания, SGD после Адама, но ничего не пробивало результаты СОТЫ.

Тогда я отчаялся и пошел в лекцию.

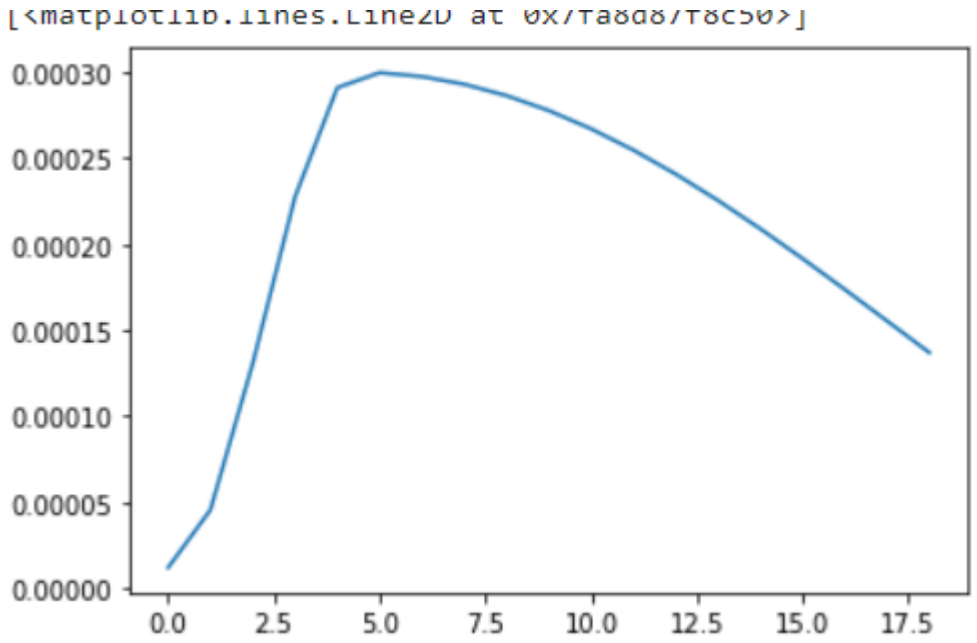
Пробуем расписание из лекции и label smoothing

1)запилил самопильный smoothCE
2) отдебажил его (чтоб совпадал с обычным CE) ((графичек в самом конце))
3)запилил линейный scheduler в форме галочки (возрастает, потом убывает)
4) подобрал к нему оптимальный lr (оказалось, что дефолтный лучший, но я использовал в 3 раза выше)

Поставил учиться. На 8 эпохе качество текущей СОТЫ было пробито, но дальше модель начала переобучаться. Я остановил досрочно.



качество при обучении



lr расписание

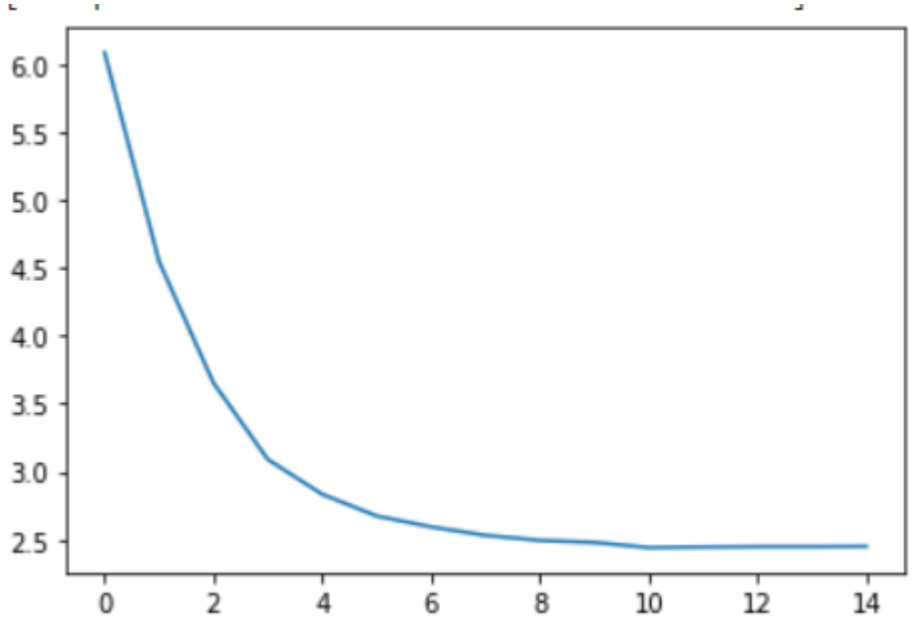
Но видно было, что успех есть

Финальное обучение

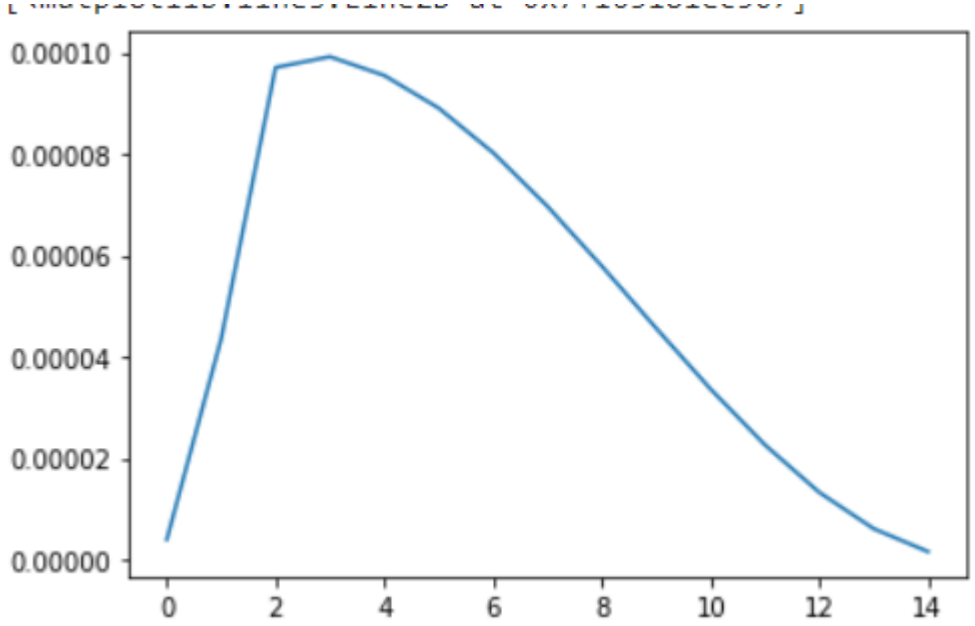
Тут я решил собрать все, что работает хорошо

1) label smoothing (я брал 0.1)
2) batch accumulation x4
3) расписание, как в лекции (см картинку ниже) с максимумом в том, что норм работало на дефолтном адаме (1e-4 вроде).
4) 15 эпох. Т.к. раньше на 10 начинал переобучаться, а тут шаг поменьше
5) reduceonplatou с patience=3 эпохи, на случай, если пойдет переобучениеили не будет улучшений
6) Еще была идея дообучить с SGD на симметрично возрастающем-убывающей lr-расписании, но тут прислали расходы датасферы)0) (поэтому без этого пункта)

Обучил, получил качество не сильно выше, чем было на одной из первых попыток после чекпоинта…



Обучение финальной модели



расписание

Итог: 26.4BLEU

Я не знаю, что я делаю не так в этой жизни
Я использовал примерно все идеи, которые мы обсуждали на лекции и семинарах (кроме того, что поковыряться внутри кода трансформера)

Можно было, конечно, поперебирать всякие гиперпараметры типа dropoutов, label smoothing, и тп, но это бы не дало сильных приростов.

Другие оптимайзеры кроме SGD я не стал пробовать - опыт прошлой дз показал, что там где что-то норм рабит, там и адам норм рабит.

SGD я пробовал после оптимизации дефотным адамом на разных lr - результатов не дало - или переобучение, или нет улучшения.

Как-то так(В прошлый раз не хватило 0.5% аккураси до порога, в этот раз 0.5BLEU люблю жызьнь

Красивый график на конец: пруф, что моя SmoothCE рабит