Попробовала сделать дистилляцию с MSE.loss, как нашла в источниках - гитхаб и статья (достаточно новая).
Не могу сказать, что это очень сильно сыграло на качестве модели, но улучшения все-таки были, а, на секундочку, число параметров сильно уменьшилось!
Я экспериментировала с дистиллированием достаточно долго: вначале считала MSE на выходах моделей и на выходах моделей и еще выходом с аттеншн-слоя. Долго возилась с изменением с конфигурации модели, чтобы постараться уменьшить число параметров модели в 10 раз.
По итогу лучший результат был достигнут на модели, в которой дистиллировался как выход модели, так и выход с аттеншн-слоя (studen4), общее число параметров-4165, качество - 6.1e-5, эпох обучения - 20 (надо было, наверное, брать больше, чем в базовой модели), число cnn_out_channels - 2, размер hidden - 20,   число слоев в gru - 1.