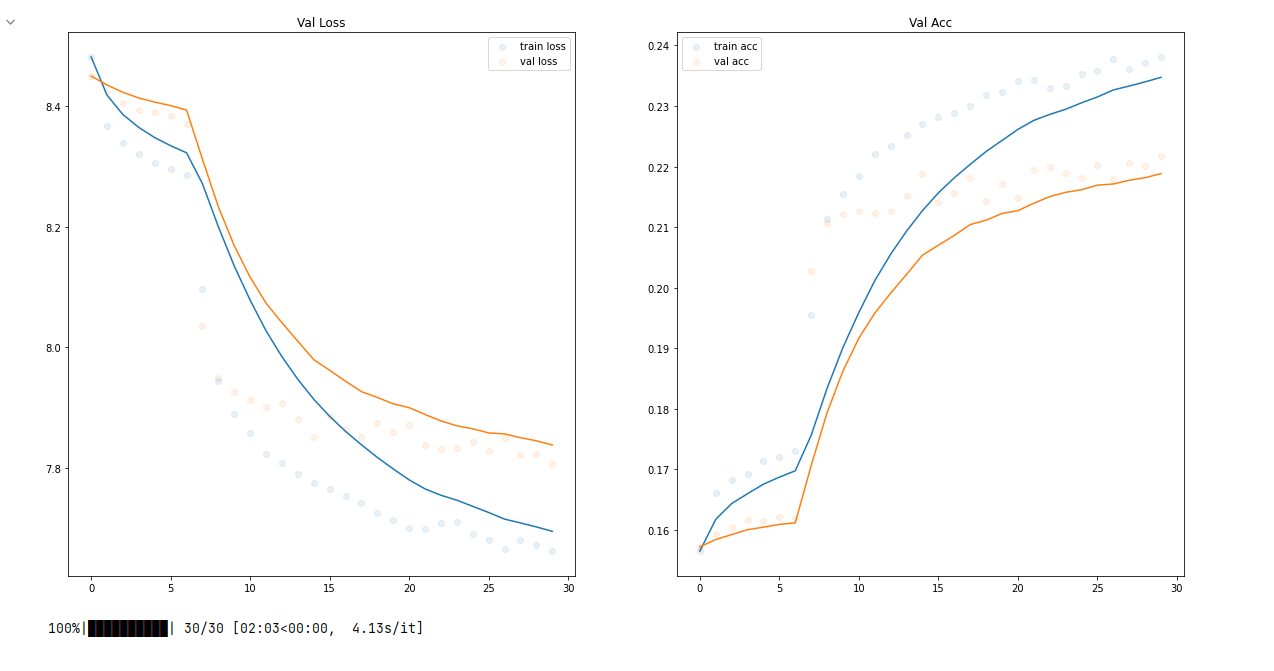
При проведении экспериментов будем фиксировать seed перед инициализацией весов модели.

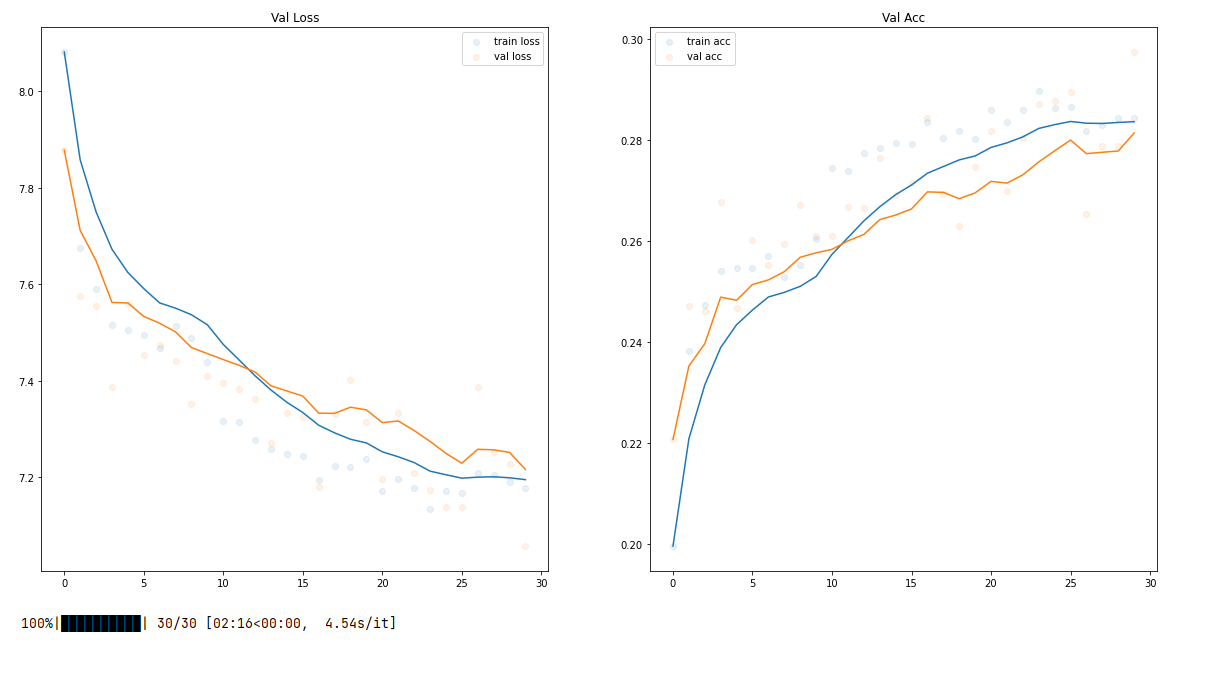
1. **Реализовать каждую из активационных функций - sigmoid, tanh, ReLu (можно взять любую из вариаций). Исследовать влияние каждой функции на обучение модели.**

Проанализируем влияние функции активации на скорость сходимости модели.

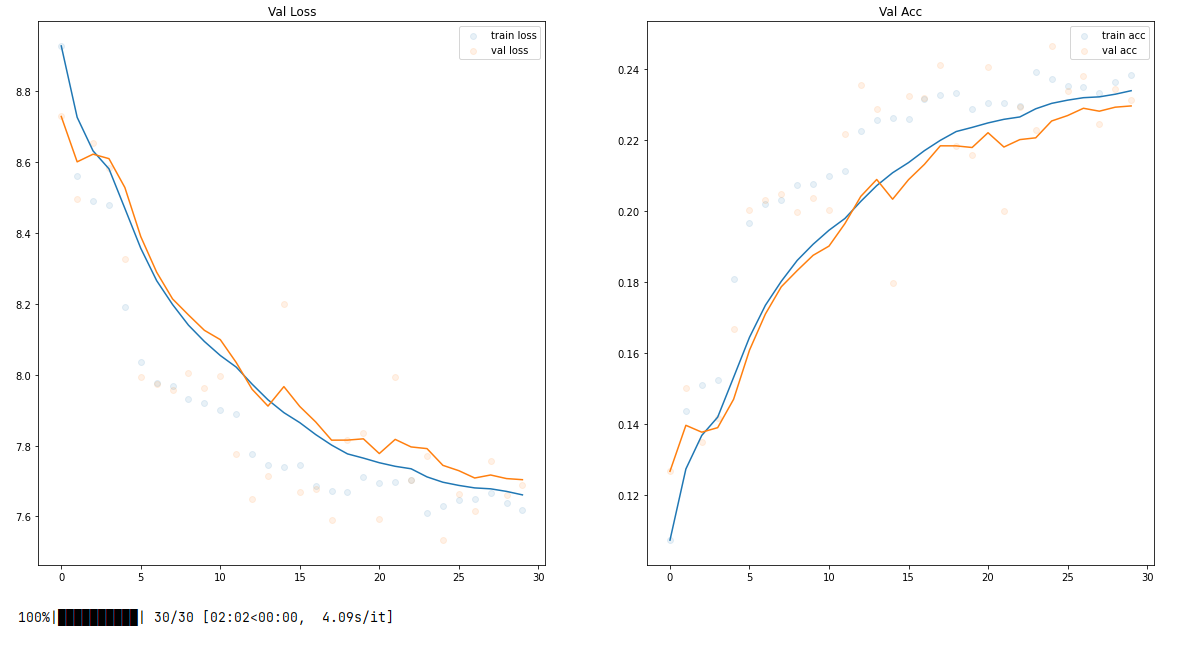
Sigmoid:



Tanh:



ReLU:



Выводы: в данной модели при замене функции активации мы не получили большой разницы в скорости обучения.

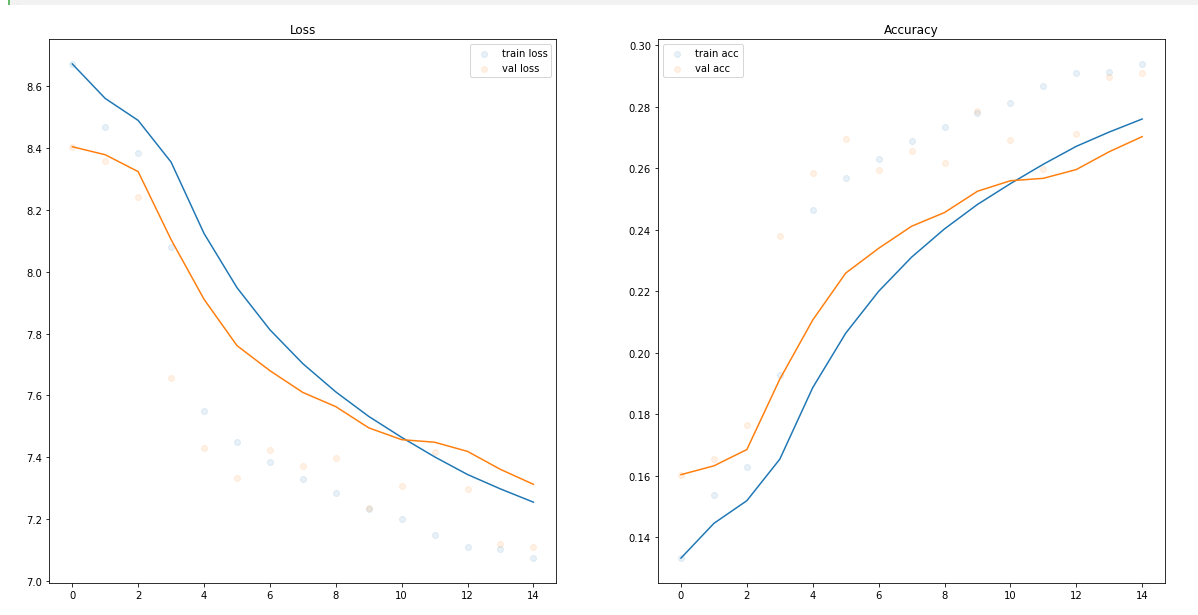
Тем не менее, если сравнить графики обучения для сигмоиды и гиперболического тангенса, то можно увидеть, что у сигмоиды в начале есть более пологий участок, которого нет у тангенса. Возможно, это связано с тем, что сигмоида уменьшает градиент, передаваемый в процессе backpropagation.

1. **Реализовать методы регуляризации - DropOut и L1/L2 и исследовать их влияние на сходимость.**

Зафиксируем параметры модели:

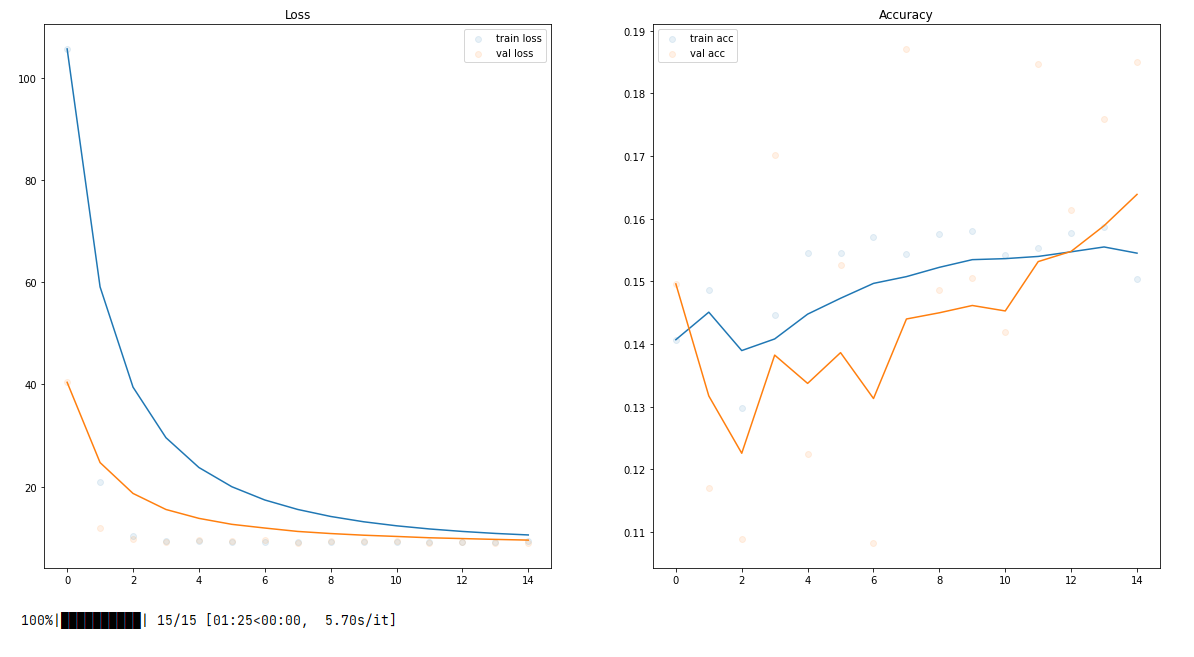
EPOCH\_COUNT = 15  
BATCH\_SIZE = 100  
INPUT\_SIZE = 3072  
HIDDEN\_SIZE = 64  
OUTPUT\_SIZE = 10  
LEARNING\_RATE = 0.03

Без регуляризации:



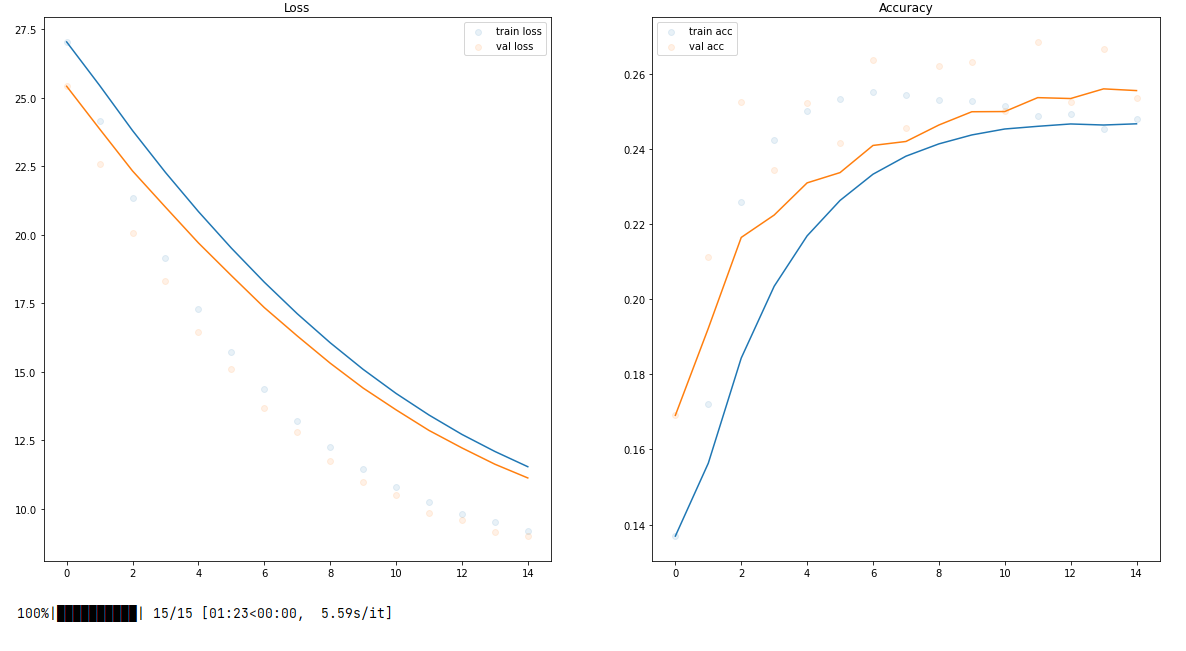
L1-регуляризация:

1. REGULARIZATION\_LAMBDA = 0.01



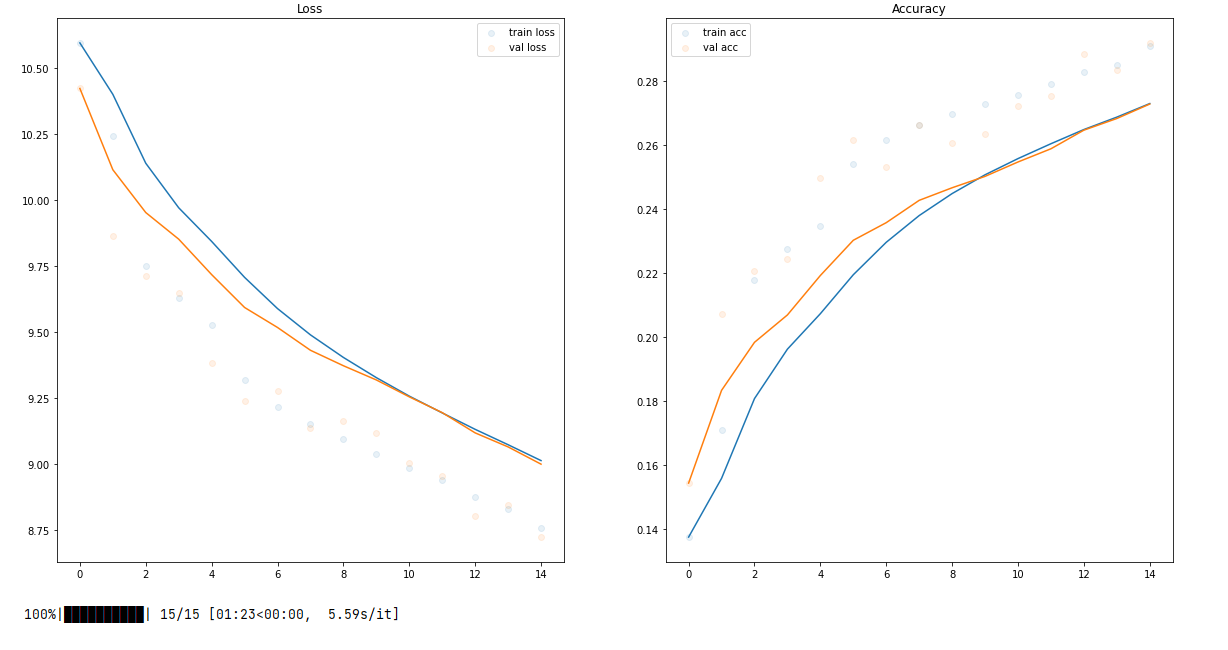
При большом коэффициенте регуляризации модель перестает обучаться.

1. REGULARIZATION\_LAMBDA = 0.001



Все еще слишком большой коэффициент.

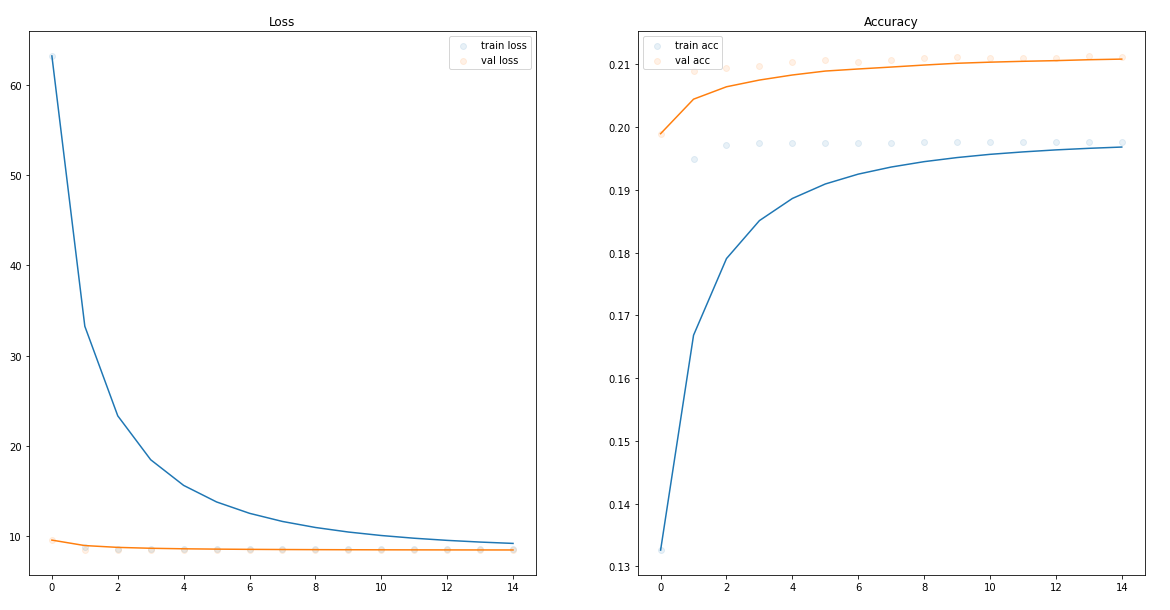
1. REGULARIZATION\_LAMBDA = 0.0001



Если сравнить этот график с обучением без регуляризации, видно положительное влияние на значение функции потерь и accuracy на валидационной выборке. Также значения loss и accuracy становятся ближе друг другу, чем при отсутствии регуляризации. Это можно объяснить, что модель меньше подстраивается под обучающие данные и имеет лучшую обобщающую способность при использовании регуляризации.

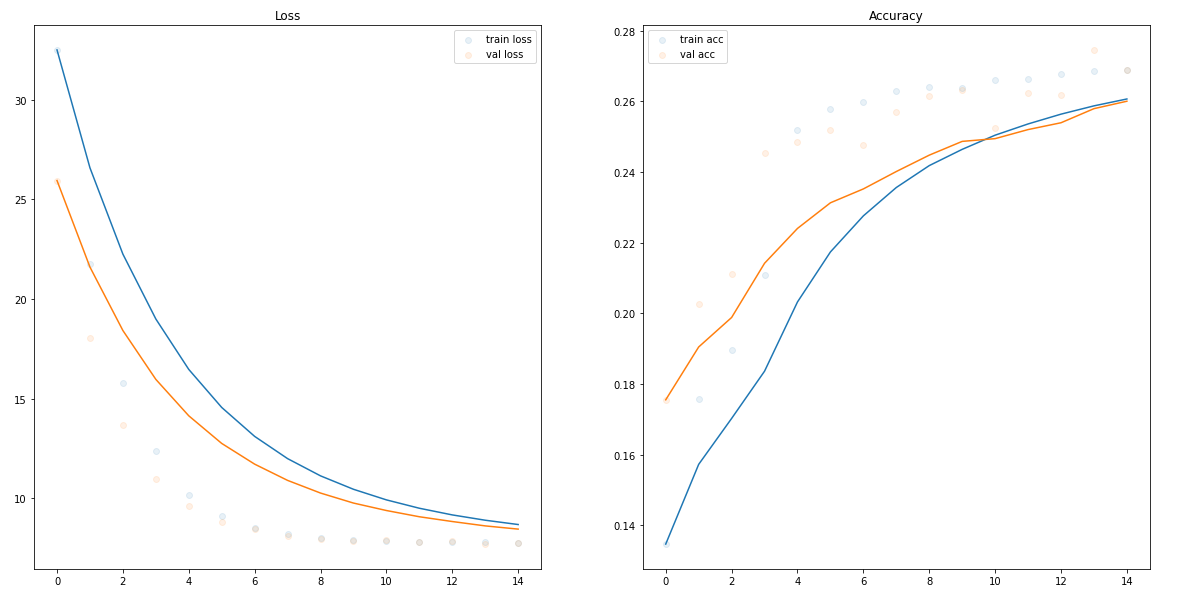
L2-регуляризация:

1. REGULARIZATION\_LAMBDA = 0.1



При большом коэффициенте регуляризации модель перестает обучаться.

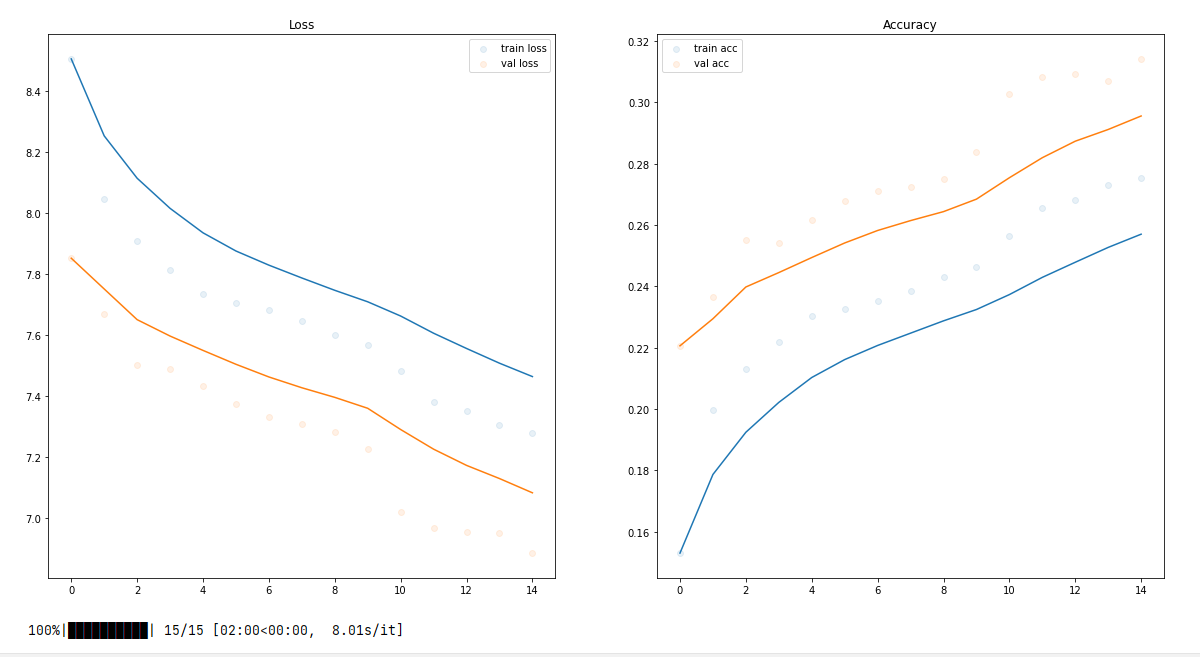
1. REGULARIZATION\_LAMBDA = 0.01



(Аналогично выводам для L1-регуляризации) Если сравнить этот график с обучением без регуляризации, видно положительное влияние на значение функции потерь и accuracy на валидационной выборке. Также значения loss и accuracy становятся ближе друг другу, чем при отсутствии регуляризации. Это можно объяснить, что модель меньше подстраивается под обучающие данные и имеет лучшую обобщающую способность при использовании регуляризации.

DropOut:

REGULARIZATION\_LAMBDA = 0.00  
INPUT\_DROPOUT\_PROBA = 0.8  
HIDDEN\_DROPOUT\_PROBA = 0.5



При использовании DropOut возникает бОльший разрыв между качеством на тренировочной и валидационной выборке. Это можно объяснить тем, что при обучении используются «уменьшенные» версии нейросети с разделяемым набором весов, а при валидации используется их ансамбль.

Качество на валидационной выборке улучшается быстрее, чем без DropOut.

1. **Исследовать гиперпараметры модели - количество слоев, количество нейронов модели, размер батча, коэффициент регуляризации, вероятность в DropOut.**

Зафиксируем параметры модели:

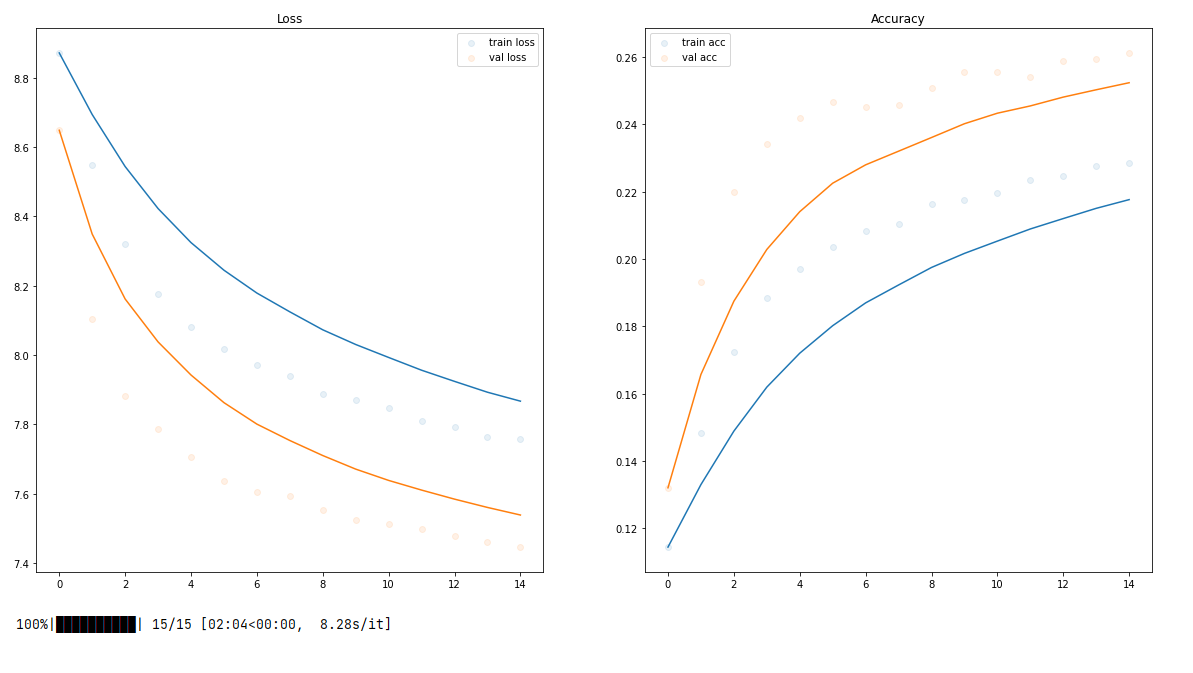
EPOCH\_COUNT = 15  
BATCH\_SIZE = 100  
LAYER\_SIZES = [3072, 64, 10]  
LEARNING\_RATE = 0.01

REGULARIZATION\_LAMBDA = 0.00

INPUT\_DROPOUT\_PROBA = 0.8  
HIDDEN\_DROPOUT\_PROBA = 0.5

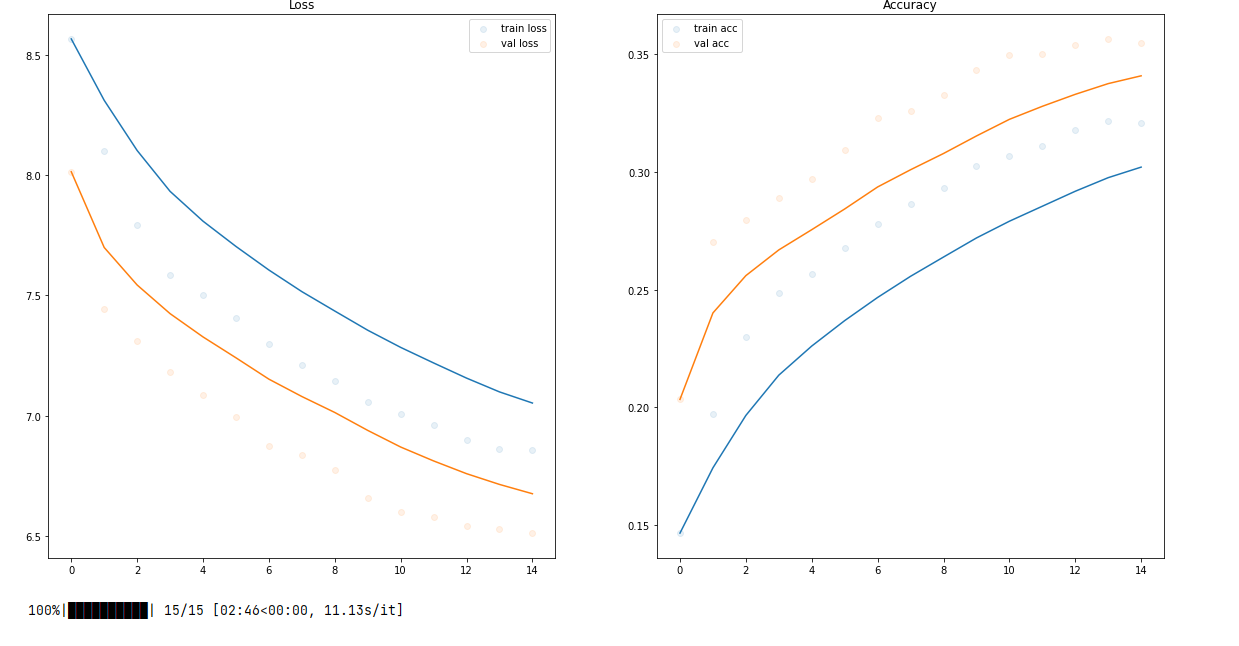
1. 1 слой, 64 нейрона

LAYER\_SIZES = [3072, 64, 10]



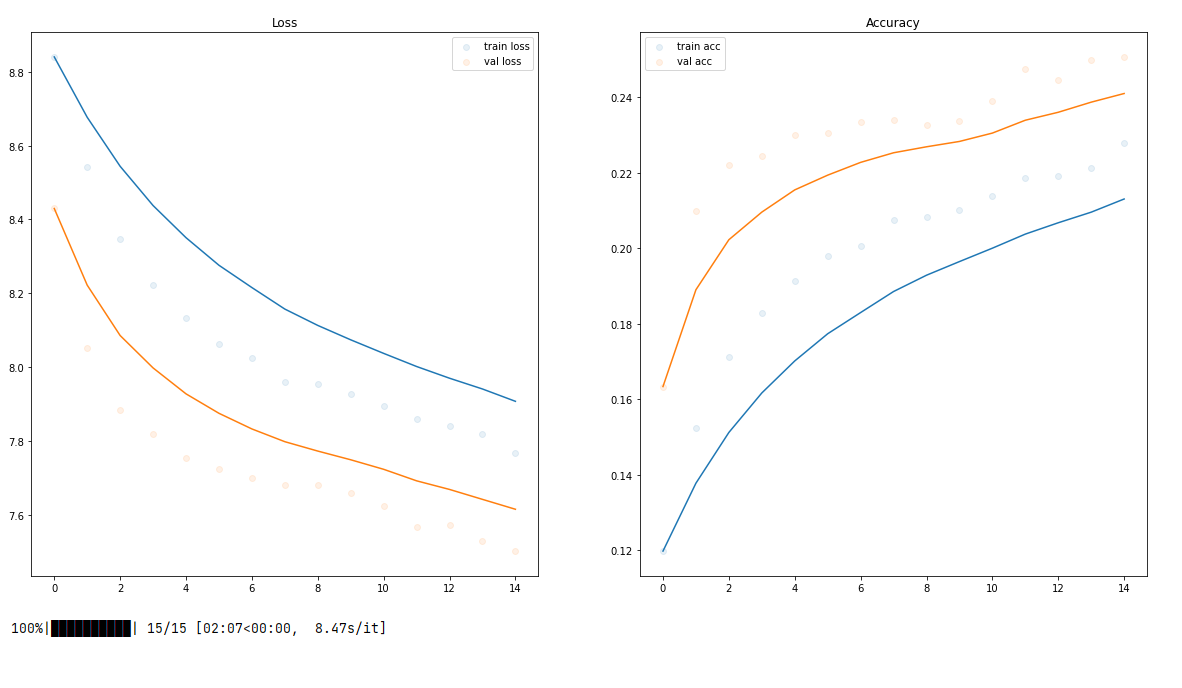
1. 1 слой, 128 нейронов

LAYER\_SIZES = [3072, 128, 10]

Результат лучше, чем при использовании 64 нейронов.

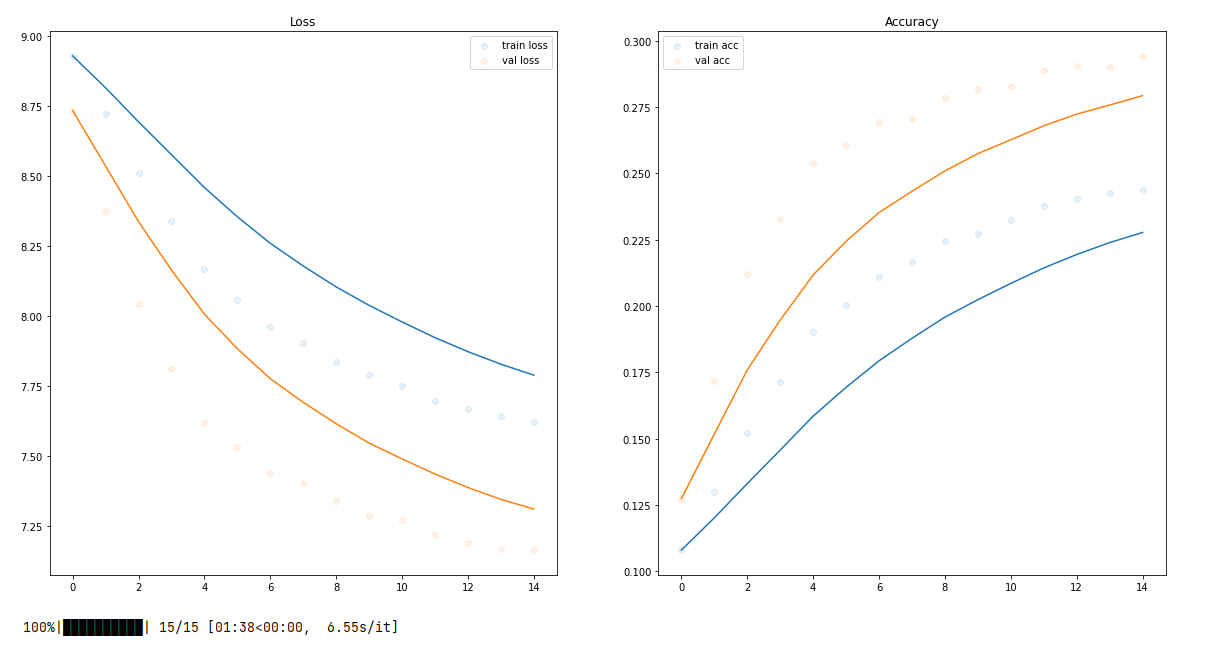
1. 2 слоя, 64 нейронов в каждом

LAYER\_SIZES = [3072, 64, 64, 10]

Нельзя сказать, что результат улучшился.

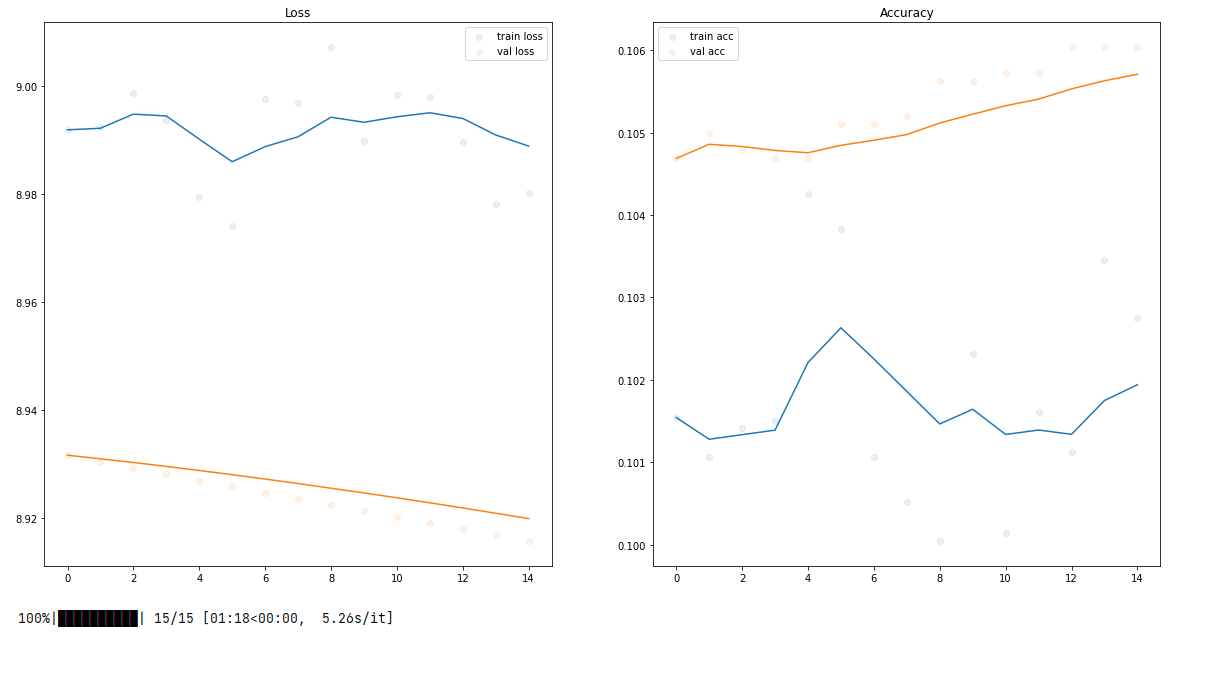
1. 1 слой, 64 нейрона, батч 200

BATCH\_SIZE = 200  
LAYER\_SIZES = [3072, 64, 10]

После 15 эпох обучения результат на валидационной выборке примерно на 2% лучше, чем при использовании батча размером в 200 объектов.

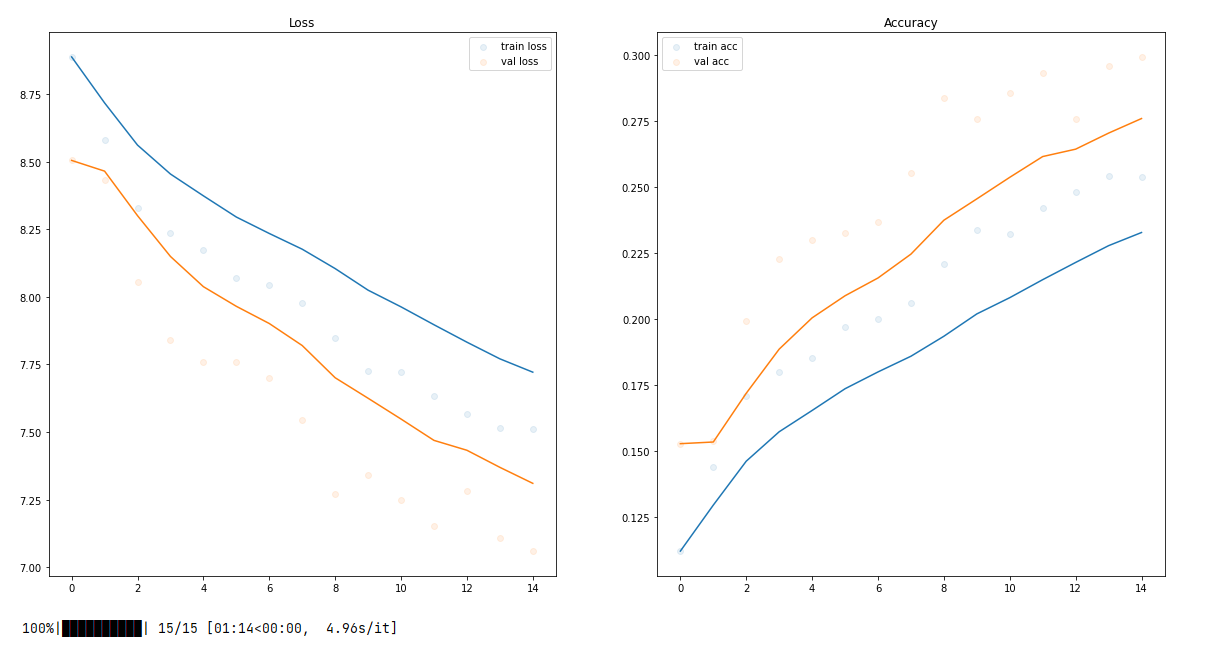
1. 1 слой, 64 нейрона, батч 5000

BATCH\_SIZE = 5000  
LAYER\_SIZES = [3072, 64, 10]



Если не менять learning rate, модель перестает сходиться.

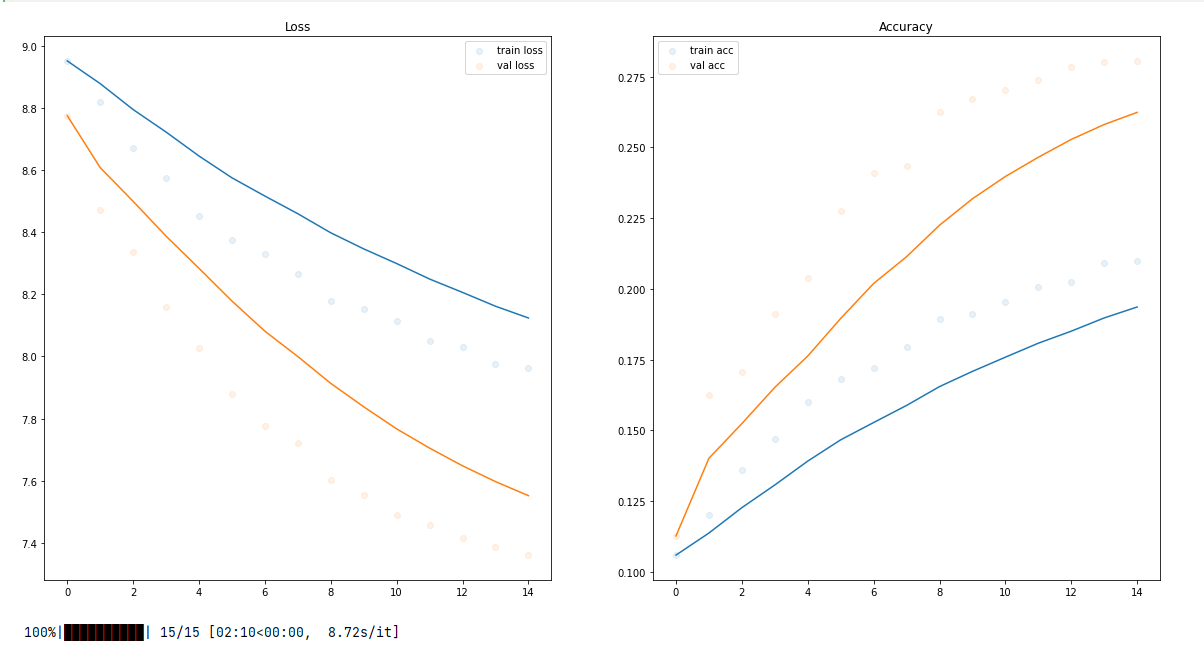
Увеличим learning rate линейно в зависимости от batch size (до 0.5).



Результат примерно такой же, как при батче размером 200.

1. 1 слой, 64 нейрона, уменьшим вероятность сохранения нейронов в результате DropOut

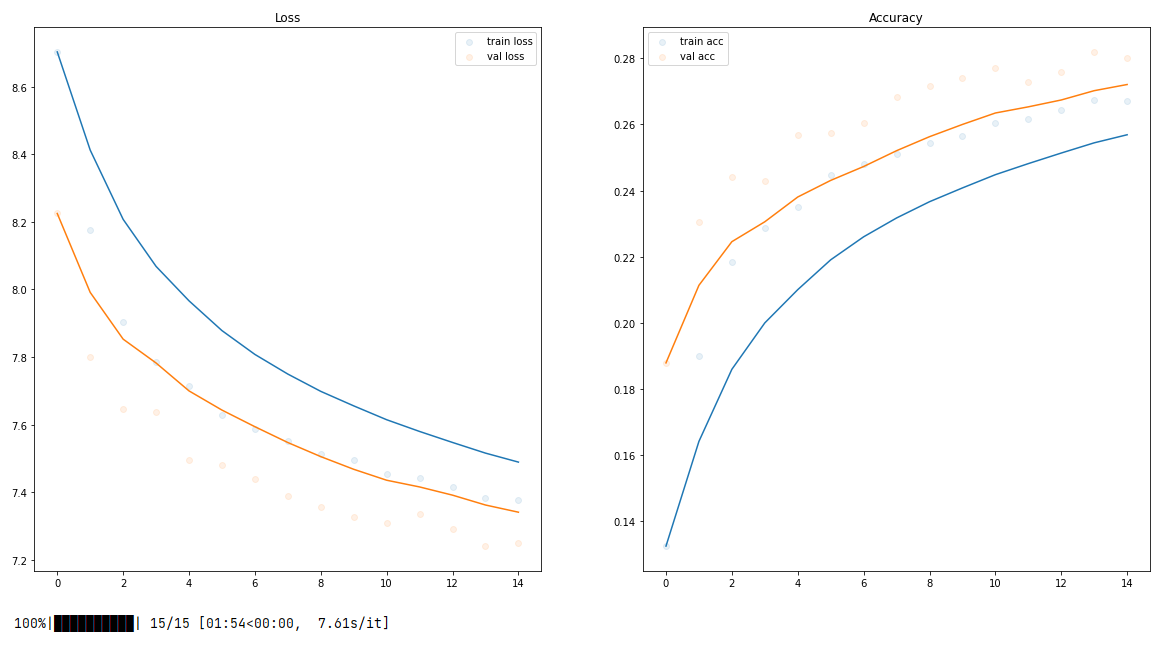
INPUT\_DROPOUT\_PROBA = 0.6  
HIDDEN\_DROPOUT\_PROBA = 0.3



Качество на валидационной выборке изменилось не сильно. Увеличился разрыв между качеством на тренировочной и валидационной выборке.

1. 1 слой, 64 нейрона, увеличим вероятность сохранения нейронов в результате DropOut

INPUT\_DROPOUT\_PROBA = 1.0  
HIDDEN\_DROPOUT\_PROBA = 0.7



Качество на валидационной выборке изменилось не сильно. Уменьшился разрыв между качеством на тренировочной и валидационной выборке.

**Выводы:**

1. В данном случае при прочих равных оказалось лучше увеличить кол-во нейронов на скрытом слое, а не кол-во скрытых слоев.
2. Прямое влияние размера батча на сходимость выявить не удалось. Небольшое увеличение батча (в 2 раза) при неизменном learning rate приводит к небольшому улучшению сходимости. Сильное увеличение батча при неизменном learning rate приводит к ухудшению сходимости (хотя в теории должно быть наоборот, т. к. при использовании большего размера батча функция потерь на всем датасете аппроксимируется лучше). Скорее всего, нужно использовать либо адаптивный learning rate, либо рассматривать разные комбинации этих 2 параметров и выбирать наилучшую.
3. Изменение вероятностей «отключения» нейронов в DropOut не показало сильного влияния на скорость сходимости.