Бустинг глубоких нейронных сетей на основе дисконтированной функции потерь

Шокоров В.А.[1] Южаков Т.А.[2] Ветров Д.П.[2]

[1]Московский физико-технический институт (национальный исследовательский университет)

[2]Национальный исследовательский университет «Высшая школа экономики»

Мотивация и описание проблемы

- Глубокие сети перепараметризованы

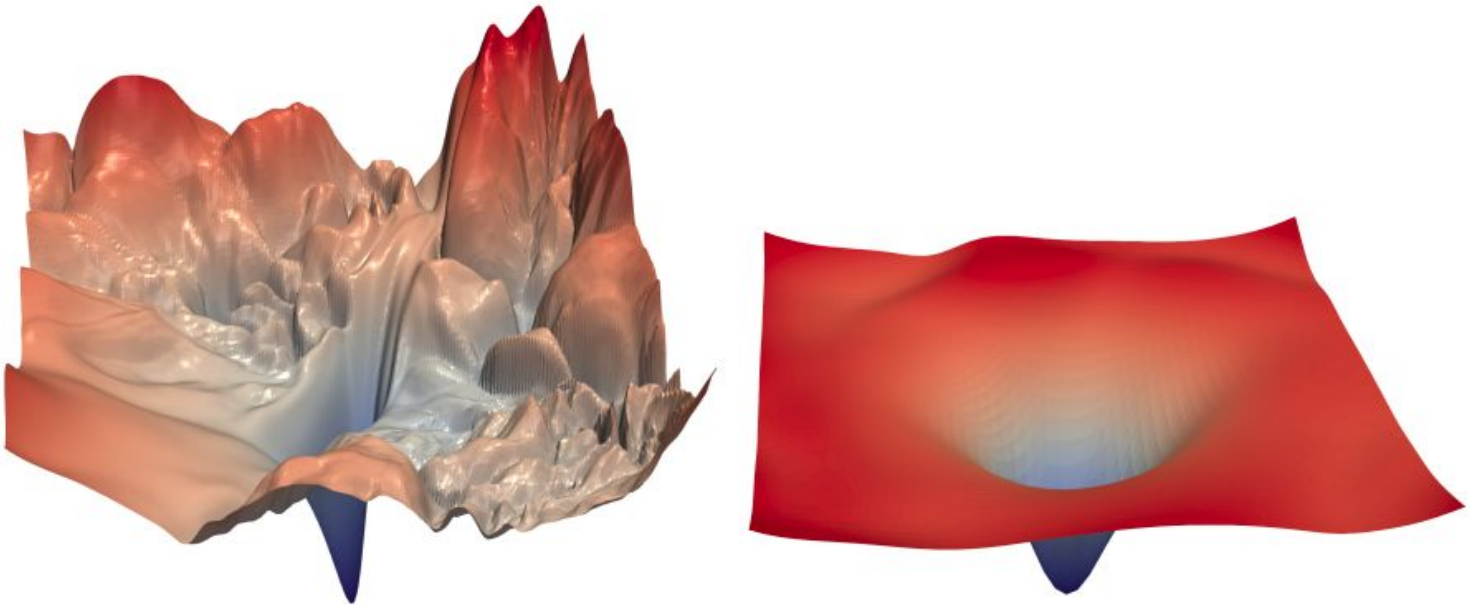
  #параметров >> #семплов

- Очень много глобальных минимумов
  - Какие-то из них обладают лучшей генерализацией чем другие

- Ключевой вопрос, как обеспечить сходимость модели к минимуму, который обладает лучшей генерализацией?

Дисконтированная функция потерь

Широкие оптимумы обладают лучшей генерализацией

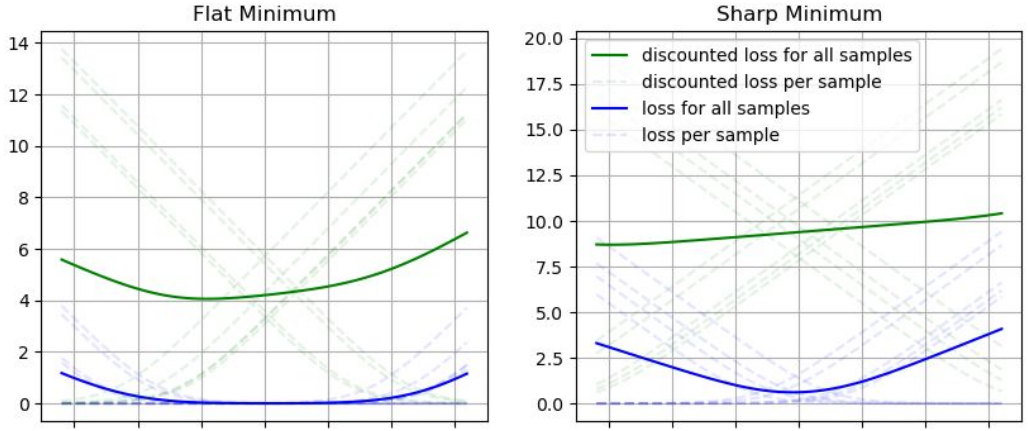


Hao Li, Zheng Xu, Gavin Taylor, Tom Goldstein: Visualizing the Loss Landscape of Neural Nets. CoRR abs/1712.09913 (2017)

Дисконтированная функция потерь

Цель: попасть в широкий минимум

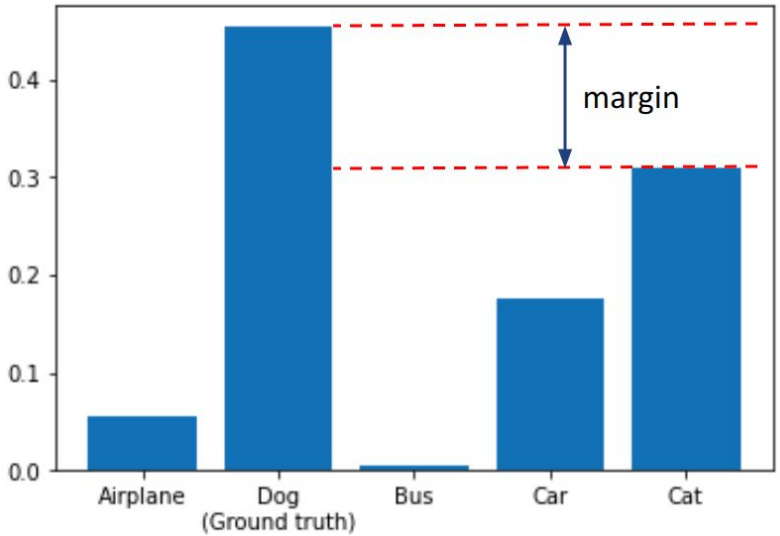
$$L = \frac{1}{N}\sum_i -\log \frac{e^{l_{y_i,i}-g}}{e^{l_{y_i,i}-g} + \sum_{j \neq y_i} e^{l_{j,i}}},$$



где l - логит (выход модели), y_i - метка правильного класса, g - значение задаваемого гэпа (гиперпараметр).

Таким образом ожидаем, что узкие минимумы перестанут быть минимумами, а широкие минимумы останутся

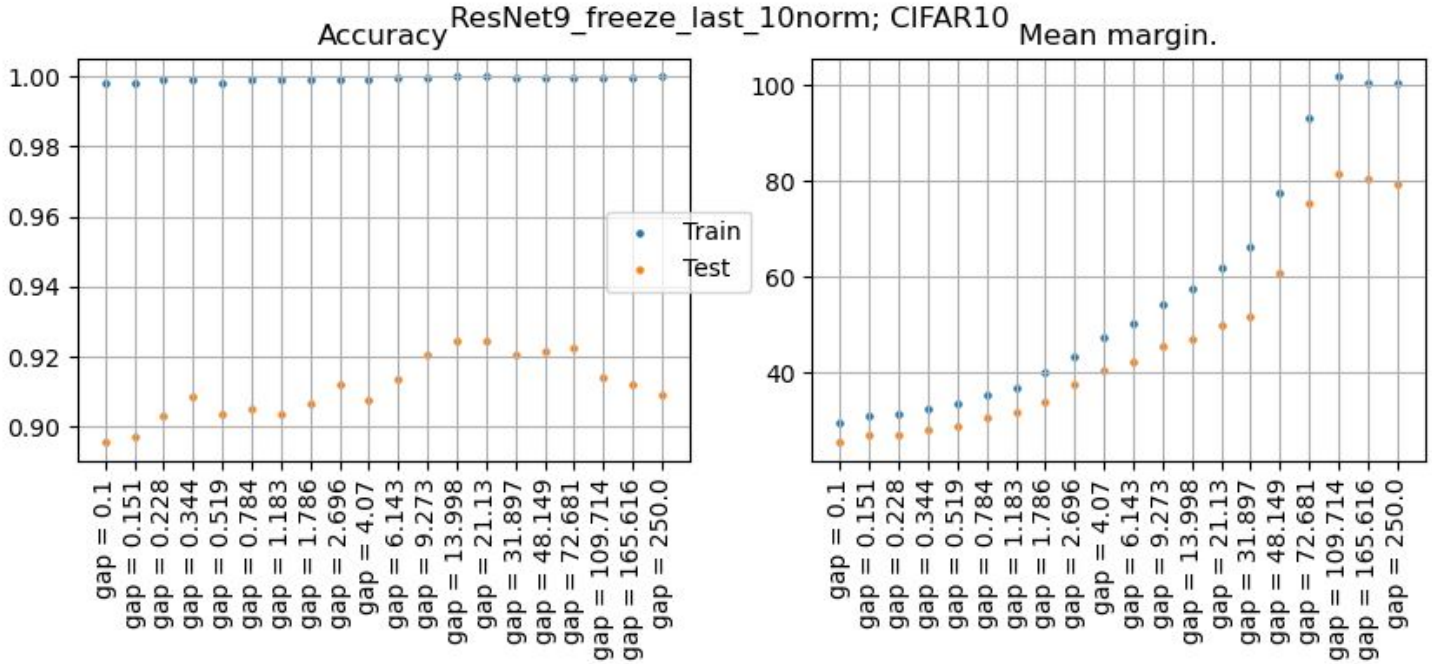
Дисконтированная функция потерь



Пример определения зазора (марджина) в домене вероятностей. В работе мы считаем марджин в домене логитов

margin = score of correct label -
- maximum scores of other label

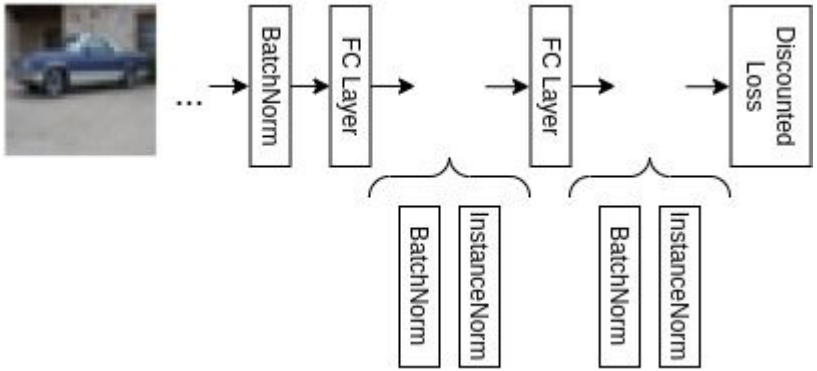
Дисконтированная функция потерь



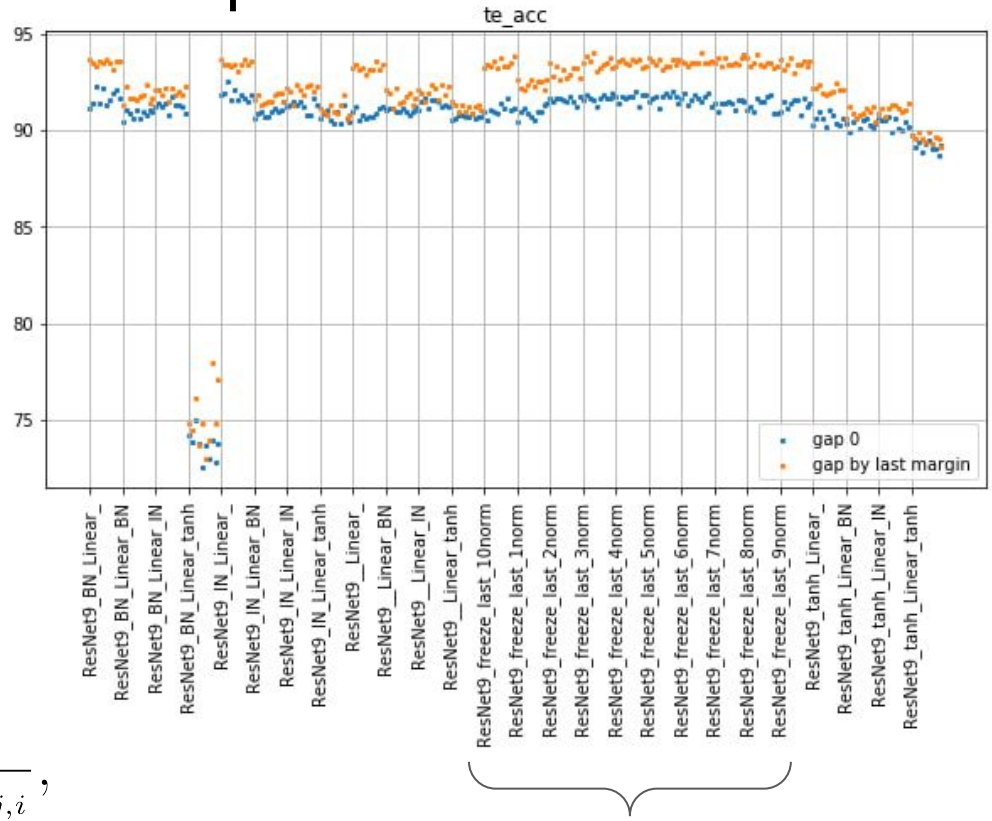
Оптимальный гэп можно задать как средний зазор обученной сети.

Дисконтированная функция потерь

Оптимальный гэп задается как средний зазор обученной сети.



$$L = \frac{1}{N}\sum_i -\log \frac{e^{l_{y_i,i}-g}}{e^{l_{y_i,i}-g} + \sum_{j\neq y_i} e^{l_{j,i}}},$$



Заморожен последний слой модели

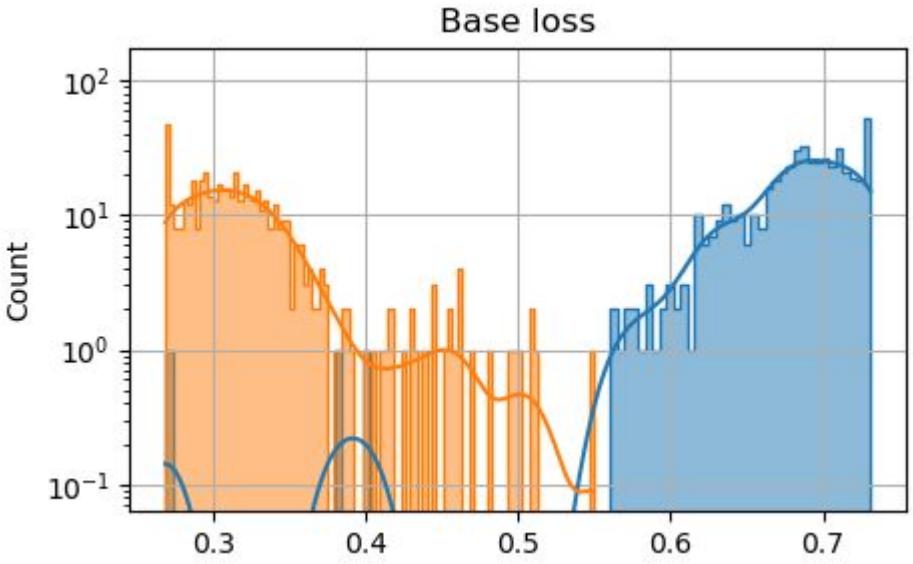
Теоретическое обоснование

Теорема 1: Пусть модель обучена с дисконтированной функцией потерь, с заданным значением гэпа *g*, до 100% точности на обучающей выборке. Тогда на всех объектах обучающей выборки зазор (марджин) для данной модели будет не меньше, чем g.

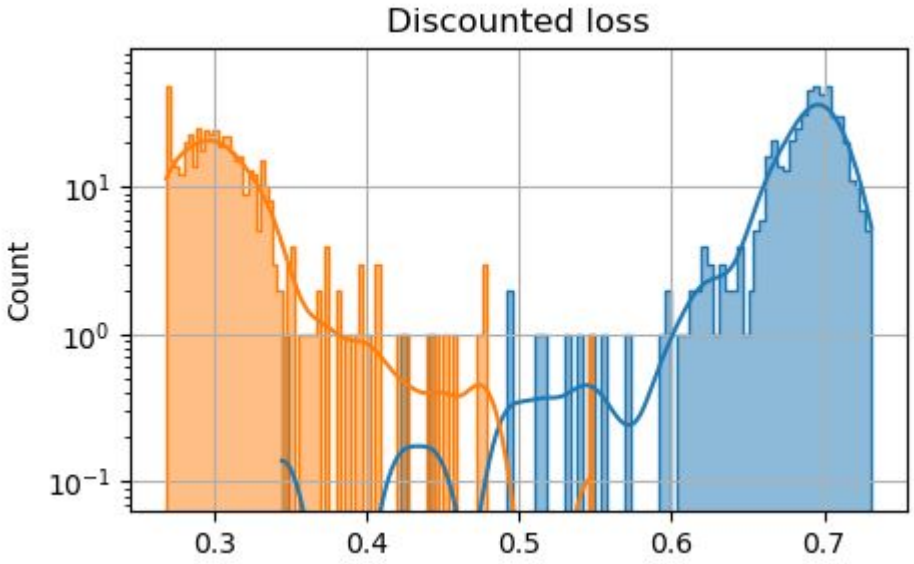
Теорема 2: Существует максимальное значение гэпа, при обучении на котором, модель может показывать 100% точности на обучающей выборке.

Дисконтированная функция потерь

Мы ожидаем, что предлагаемый метод будет улучшать ROC-AUC, так как эта метрика лучше реагирует на изменение в тяжелых хвостах распределений вероятностей, когда модель неуверенно классифицирует объекты



ROC-AUC: 0.99779



ROC-AUC: 0.99973

Бустинг глубоких нейронных сетей

- Ансамблирование моделей приводит к улучшению качества алгоритма

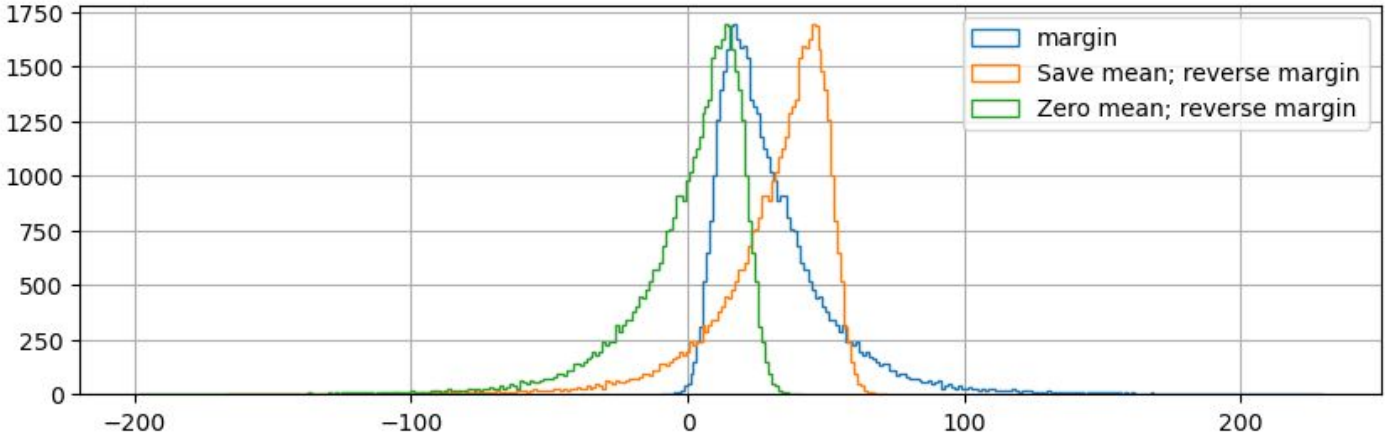
- Бустинг - метод построения ансамбля:

  следующая получаемая модель исправляет ошибки предыдущих

- Глубокие нейронные сети имеют нулевую ошибку.

- Ключевой вопрос, как построить бустинг на глубоких нейронных сетях?

Бустинг глубоких нейронных сетей

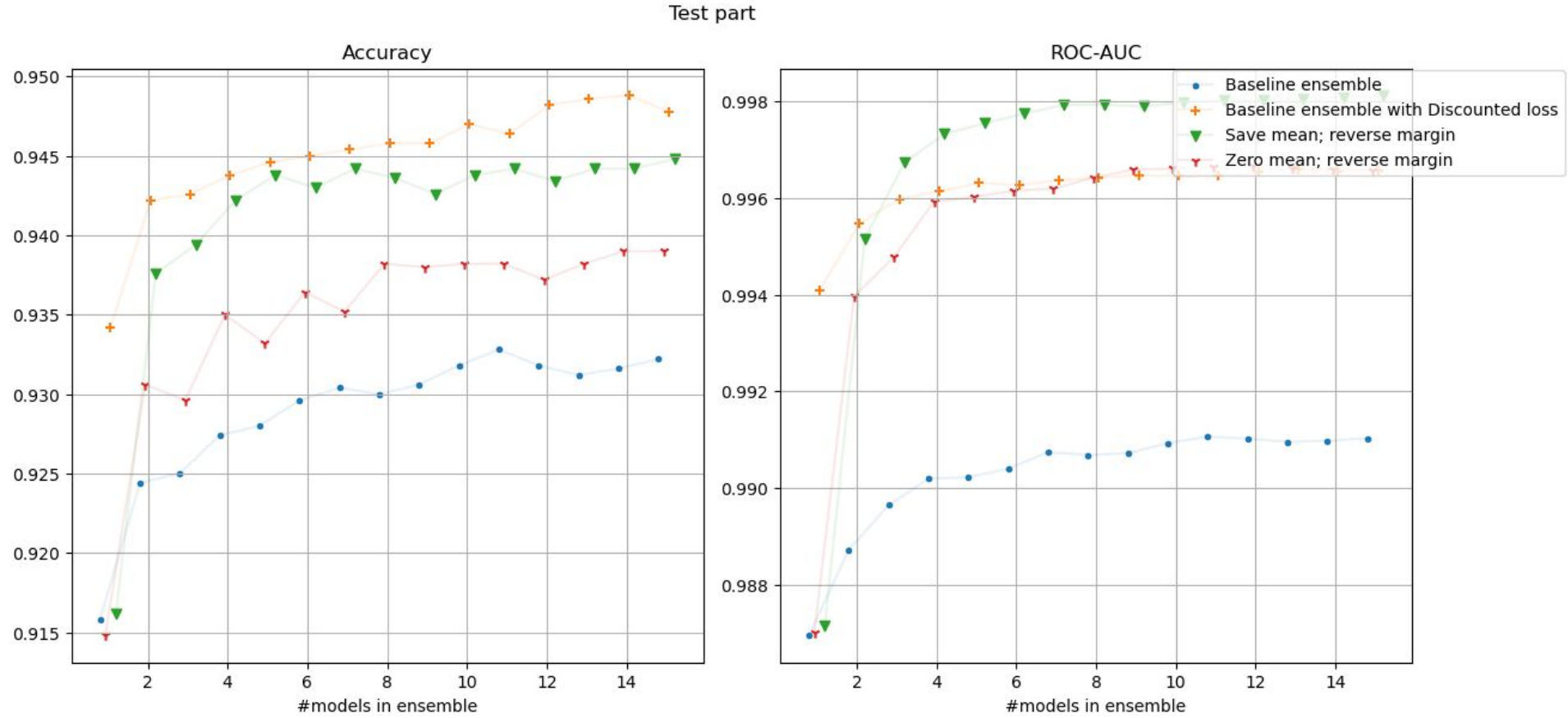


m_i - зазор *i*-го объекта, g_i - значение гэпа *i*-го объекта для следующей модели.

$$\text{Save mean; reverse margin} : g_i = 2 \cdot \sum_j m_j - m_i$$

$$\text{Zero mean; reverse margin} : g_i = \sum_j m_j - m_i$$

Бустинг глубоких нейронных сетей



Заключение

- Предложена дисконтированная функция потерь, повышающая генерализацию модели
- Показано, что подход не зависит архитектуры модели
- Предложены методы построения бустинга моделей

Обо мне и моих планах

Кто сейчас ваш научный руководитель?

- Ветров Дмитрий Петрович (консультант)
- Стрижов Вадим Викторович (формальный научный руководитель)

Кто планируется вашим научным руководителем в аспирантуре (если такой есть)?

- Планируется сохранить нынешний формат работы (договоренности с руководителями и кафедрой имеются).

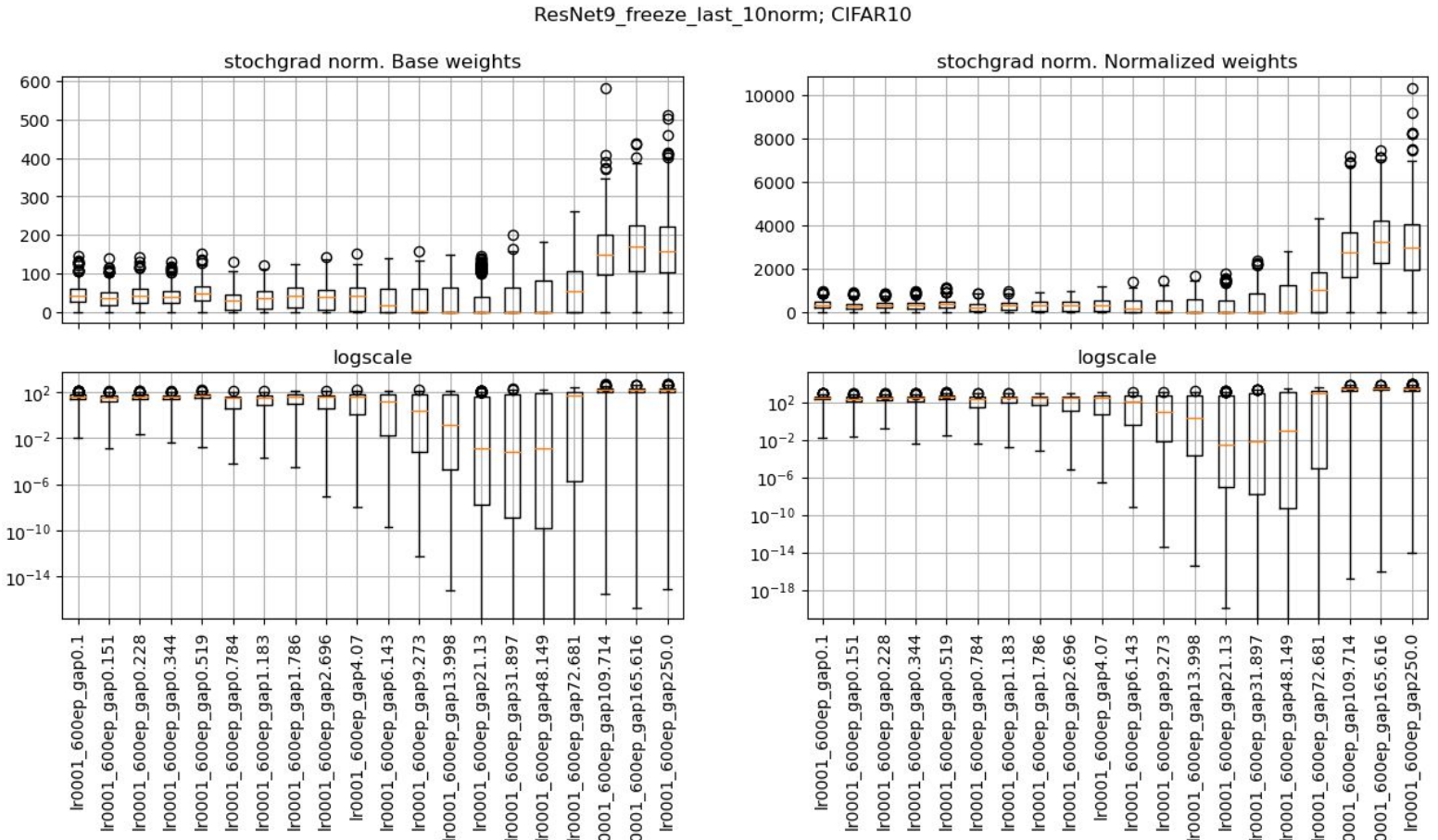
Какой темой вы планируете заниматься в аспирантуре (какой занимаетесь сейчас - увидим на презентации)?

- Цель продолжить нынешнее исследование

Есть ли у вас статьи (и в какой стадии - пишутся/поданы/опубликованы?), которые пригодятся для планируемой защиты диссертации?

- Оформляем, пишем статью посвященную данному исследованию.

Дополнение


ResNet9_freeze_last_10norm; CIFAR10
stochgrad norm. Base weights
stochgrad norm. Normalized weights
logscale
logscale
lr0001_600ep_gap0.1
lr0001_600ep_gap0.151
lr0001_600ep_gap0.228
lr0001_600ep_gap0.344
lr0001_600ep_gap0.519
lr0001_600ep_gap0.784
lr0001_600ep_gap1.183
lr0001_600ep_gap1.786
lr0001_600ep_gap2.696
lr0001_600ep_gap4.07
lr0001_600ep_gap6.143
lr0001_600ep_gap9.273
lr0001_600ep_gap13.998
lr0001_600ep_gap21.13
lr0001_600ep_gap31.897
lr0001_600ep_gap48.149
lr0001_600ep_gap72.681
lr0001_600ep_gap109.714
lr0001_600ep_gap165.616
lr0001_600ep_gap250.0