

Бустинг глубоких нейронных сетей на основе дисконтированной функции потерь

Шокоров В.А.¹

Южаков Т.А.²

Ветров Д.П.²

¹Московский физико-технический институт

(национальный исследовательский университет)

²Национальный исследовательский университет «Высшая школа экономики»

Мотивация и описание проблемы

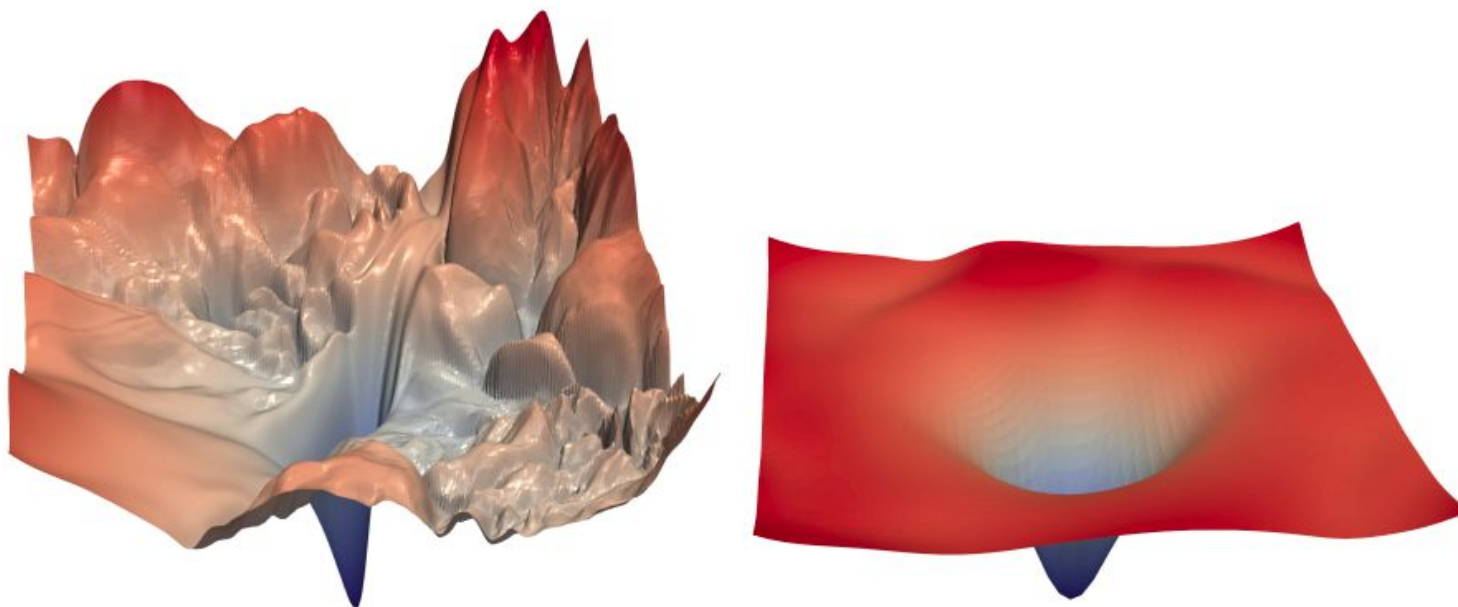
- Глубокие сети перепараметризованы

#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

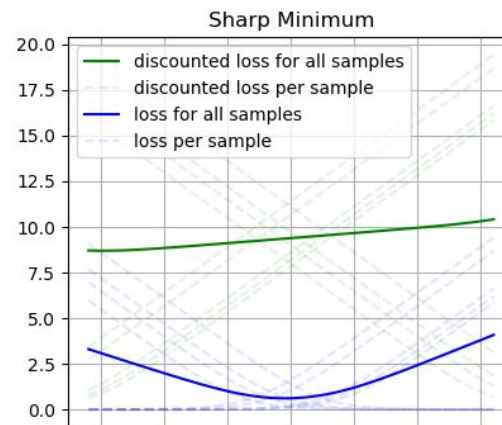
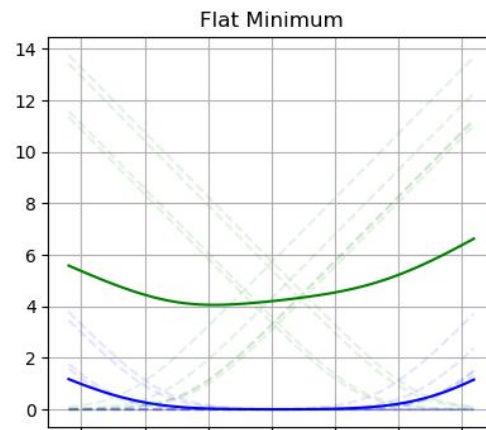
Широкие оптимумы обладают лучшей генерализацией



Дисконтированная функция потерь

Цель: попасть в широкий минимум

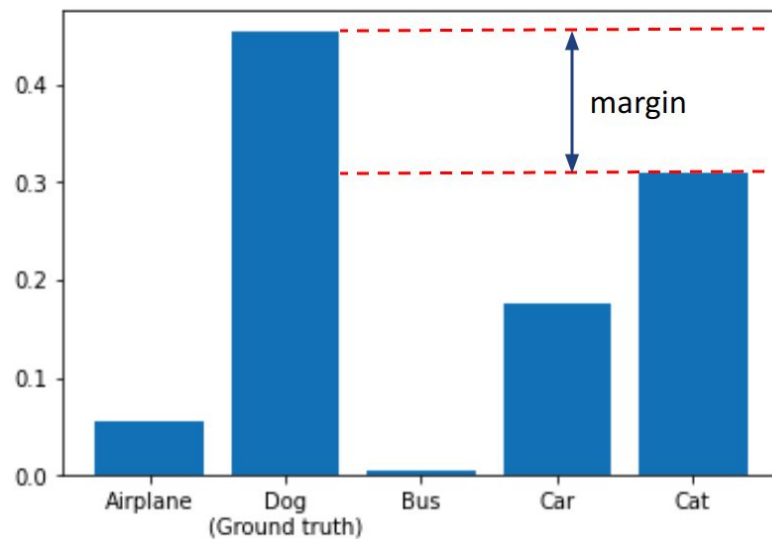
$$L = \frac{1}{N} \sum_i -\log \frac{e^{l_{y_i,i}-g}}{e^{l_{y_i,i}-g} + \sum_{j \neq y_i} e^{l_{j,i}}},$$



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

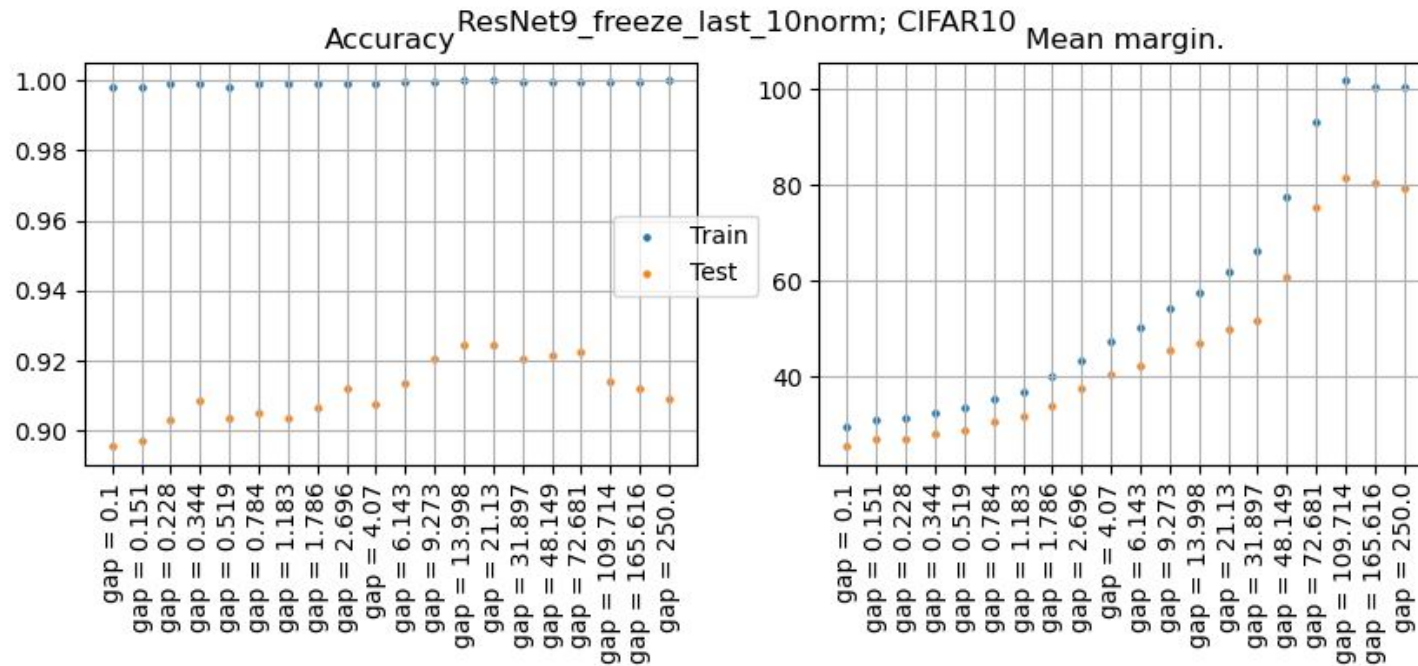
Дисконтированная функция потерь



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

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

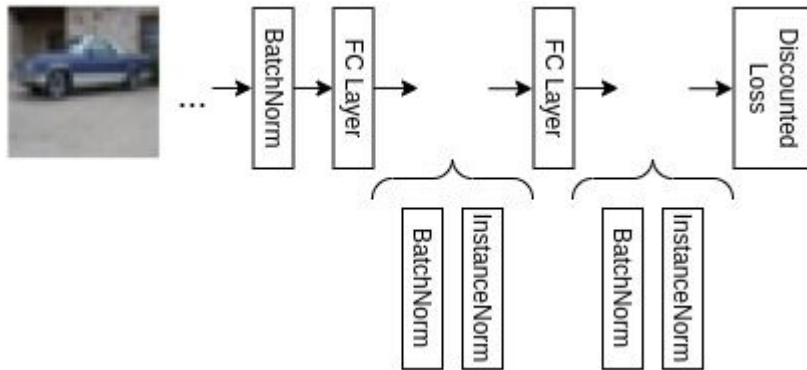
Дисконтированная функция потерь



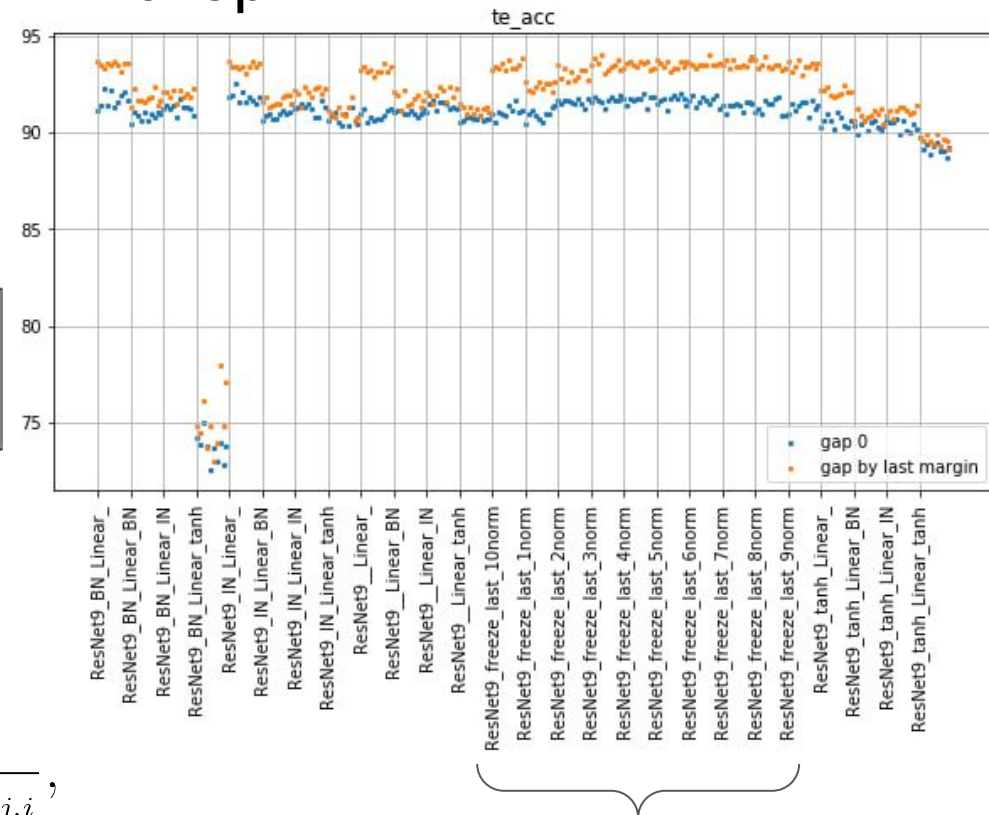
Оптимальный гэп можно задать как средний зазор обученной сети.

Дисконтированная функция потерь

Оптимальный гэп задается как средний зазор обученной сети.



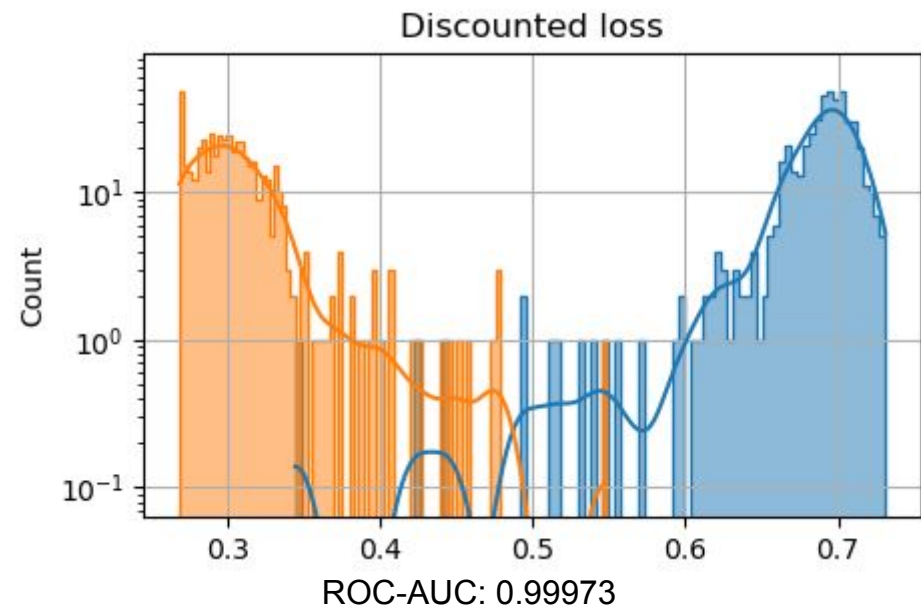
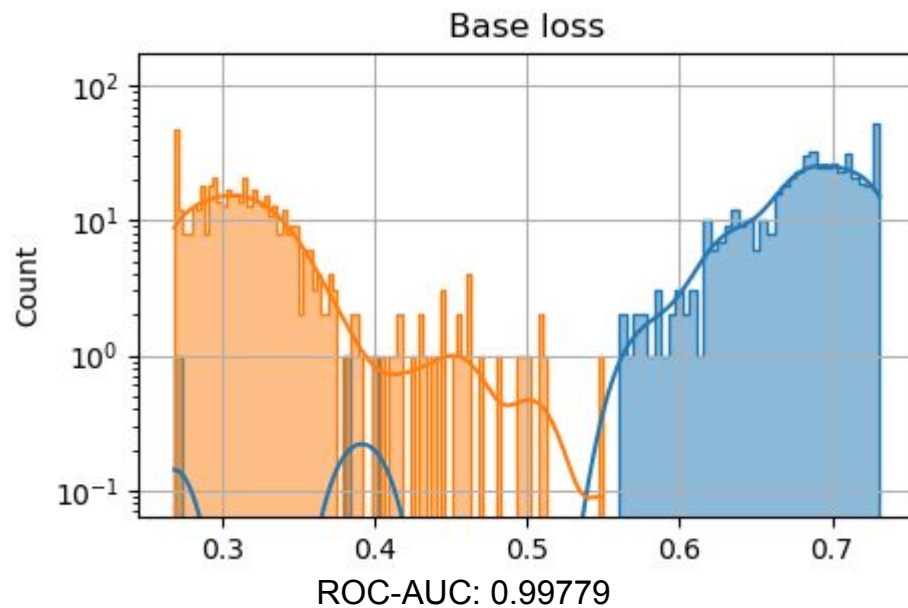
$$L = \frac{1}{N} \sum_i -\log \frac{e^{l_{y_i, i-g}}}{e^{l_{y_i, i-g}} + \sum_{j \neq y_i} e^{l_{j, i}}},$$



Заморожен последний слой модели

Дисконтированная функция потерь

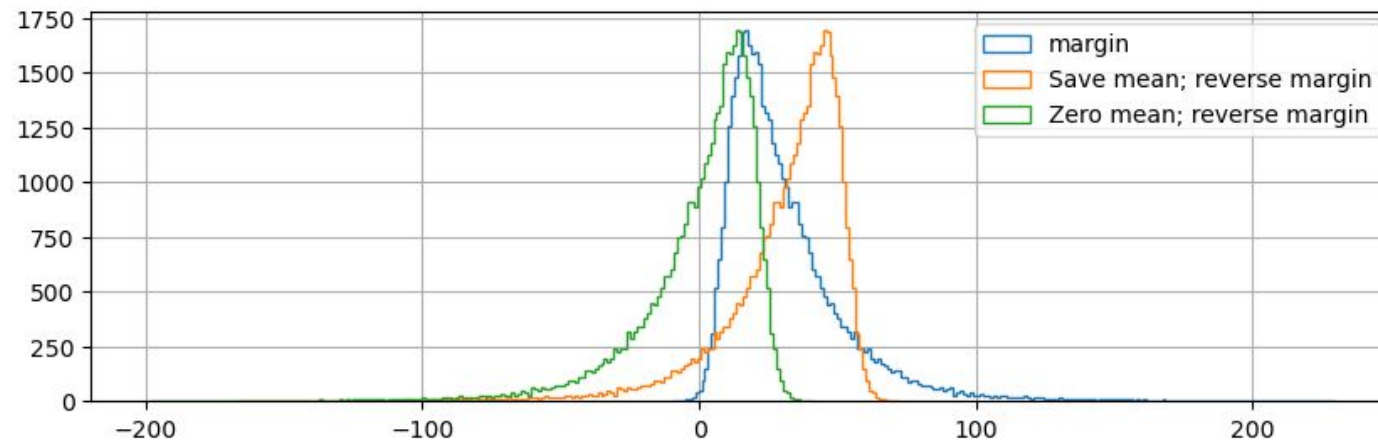
Мы ожидаем, что предлагаемый метод будет улучшать ROC-AUC, так как эта метрика лучше реагирует на изменение в тяжелых хвостах распределений вероятностей, когда модель неуверенно классифицирует объекты



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



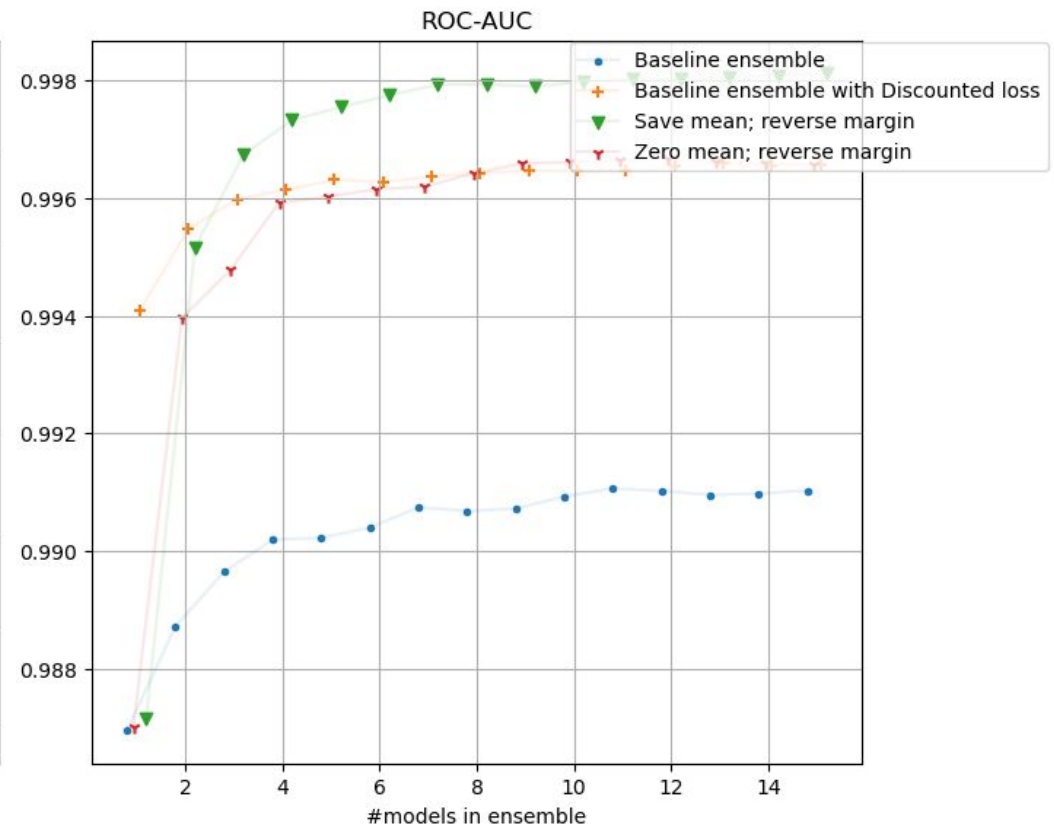
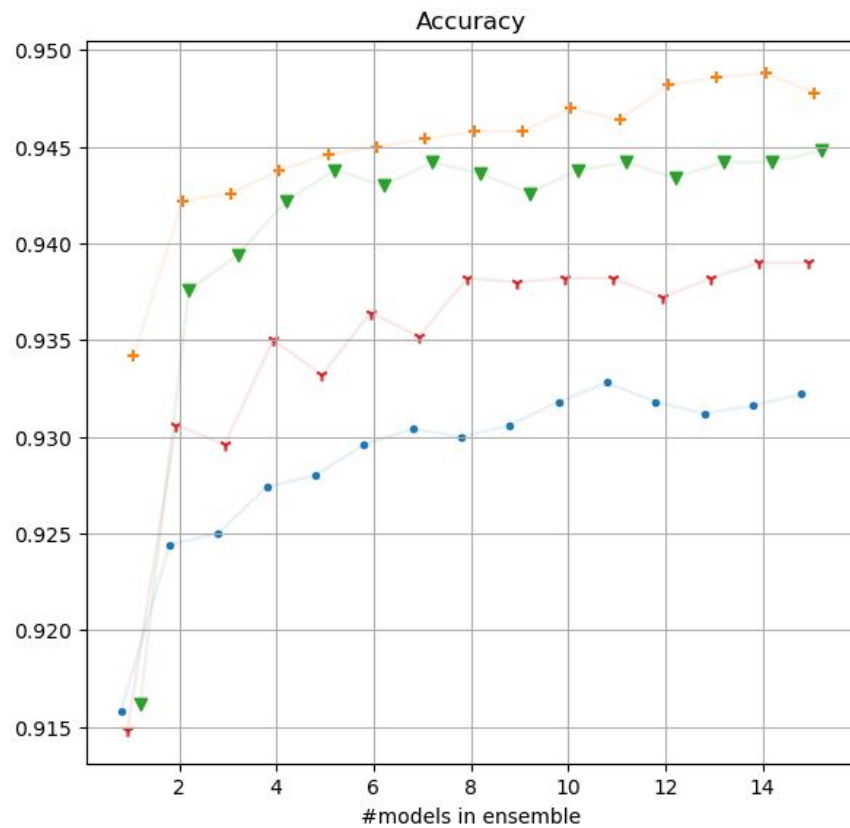
m_i - зазор i -го объекта, g_i - значение гэпа i -го объекта для следующей модели.

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

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

Бустинг глубоких нейронных сетей

Test part



Заключение

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

Дополнение

ResNet9_freeze_last_10norm; CIFAR10

