Can bias be a result of an architecture rather than data.

Факт 1: если взять две нейросети с одинаковыми архитектурами и обучить на одном датасете, то их ответы станут скоррелированы не только на объектах, похожих на датасет, но и на случайных входных данных. Например, я предполагал, что раз у overparamtrised нейросети веса случайные, а потом она обучается только на бесконечно маленьком датасете по сравнению с числом всех возможных входных данных, то нейросеть должна научиться определять

Изучение ответов маленьких нейросетей

на случайные выходные данные.

Актуальность:

Скоррелированность ответов различных нейросетей на случайных выходных данных может говорить о том, что нейросети научились выявлять одни и те же скрытые признаки (hiden features), что приводит к бесполезности использования ансамбля таких нейросетей. Более глубокое понимание причин скоррелированность ответов нейросетей на входных данных, не похожих на обучающую выборку может помочь создавать более эффективные ансамбли из нейросетей.

Что может являться причиной скоррелированности:

1. Архитектура
2. Обучающая выборка (ее погрешности)
3. Решаемая задача

После инициализации параметром, три этих фактора в совокупности с алгоритмом и временем обучения определяют то, как нейросеть отвечает на входные данные. В исследовании я хочу сравнить значимость каждого из этих трех факторов.

Метод:

Возьмем две классических нейросети и обучим их на MNIST (распознавание цифр. Пример входной картинки - рисунок 1).

Построим двухмерную гистограмму их ответов на случайные входные данные (примеры). Рисунок 2. Число у точки показывает число случайных примеров, на которые первая нейросеть ответила x, другая y. Общий процент одинаковых ответов - 81%, если бы корелляции не было, процент был бы где-то около 10%.

Видно, что одинаковые нейросети, обученные на одной и той же выборке имеют одинаковую предвзятость относительно белого шума.

Теперь возьмем другую архитектуру. CNN. Проведем то же самое, получим рисунок 3. Результат похожий (76%), но видно, что обе сети почти в 90% случаев называют белый шум восьмеркой. Это связано с тем, что рисунки из датасета на черном фоне, а цифра 8 часто больше других, поэтому сверточные сети научились асоциировать яркие картинки с 8, если шум затемнить (поставить яркость каждого пикселя от 0% до 50%), то ответ 8 вообще не будет представлен. Из-за этого эффекта дальше, при работе с CNN, вместо белого шума будем использовать картинки из датасета со случайно переставленными пикселями.

На таких случайных примерах скоррелированность 60% - по-прежнему существенная (рис 4)

Однако ответы CNN никак не связаны с ответами классических нейросетей (рис 5)

Расположение пикселя на картинке для классической нейросети не имеет значения, а у CNN информация о расположении пикселя “вшита” в ее архитектуру. Поэтому если переставить случайным но одинаковым образом пиксели во всех картинках обучающего датасета, то эту информацию CNN будет воспринимать по-другому. Таким образом, CNN обучающиеся на таких перемешанных картинках как будто имеют другую структуру\*.

\*- Такое изменение лишает CNN преимуществ перед классическими сетями.

Удивительным оказалось то, что такие нейросети имеют высокую точность (98%), но не имеют корреляции в ответах (21%) на случайные входные данные. При этом этот эффект не может быть связан с числом нейронов, потому что такие же модели без переставления пикселей имеют высокую корреляцию

Вывод:

* Различные архитектуры в разной степени склонны к скоррелированность на случайных входных данных
* Не удалось выявить скоррелированность на случайных входных данных у нейросетей принципиально разных архитектур, что может говорить о том, что такая скоррелированность не вызывается датой или задачей, а скорее возникает как комбинация архитектуры даты

рис 1

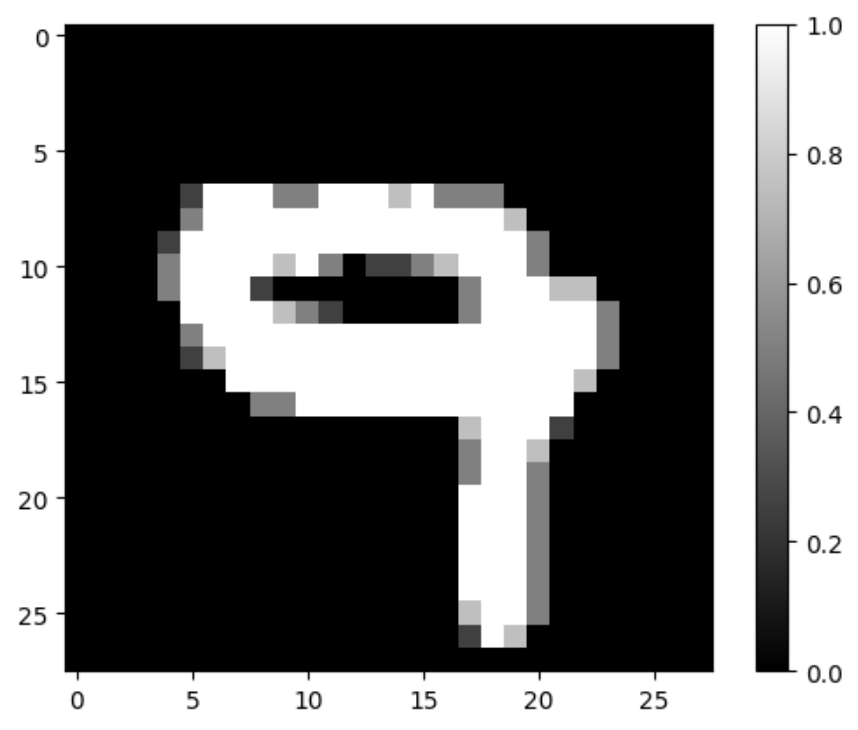


рис 2

net1 vs net2 – очень скореллированы

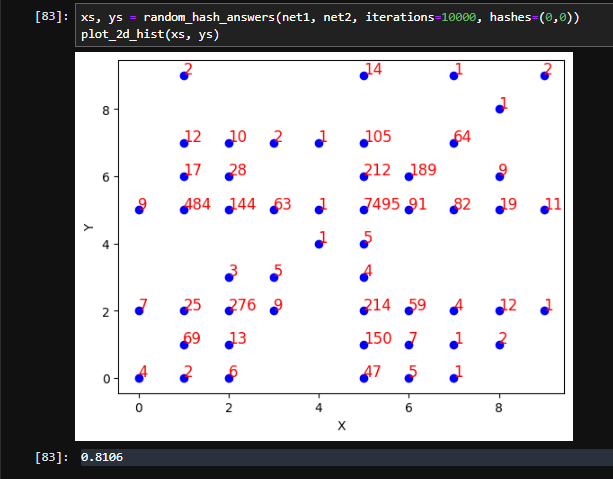


рис 3

cnn vs cnn - скореллированы

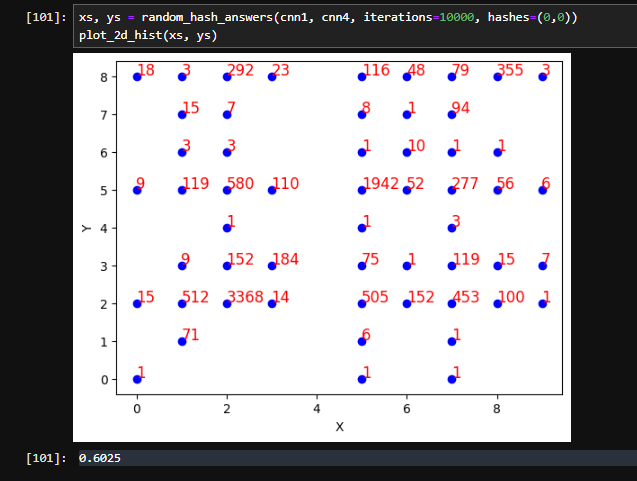


рис 4

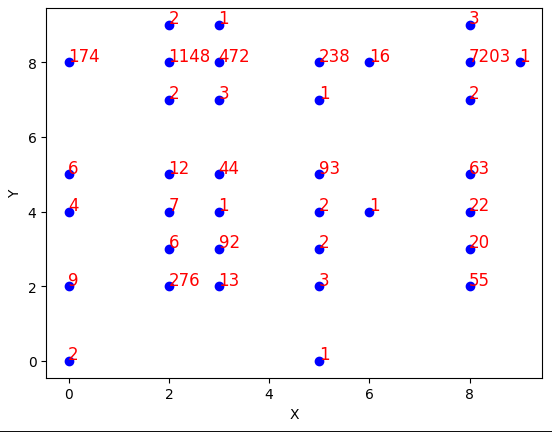
cnn vs cnn - настоящий белый шум. 

рис 5

cnn vs net – почти не скореллированы (совсем было бы 20%)

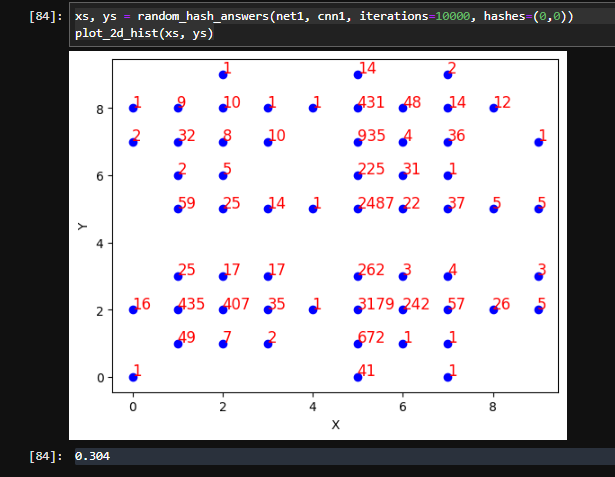


рис 6

разные cnn - нет корреляции

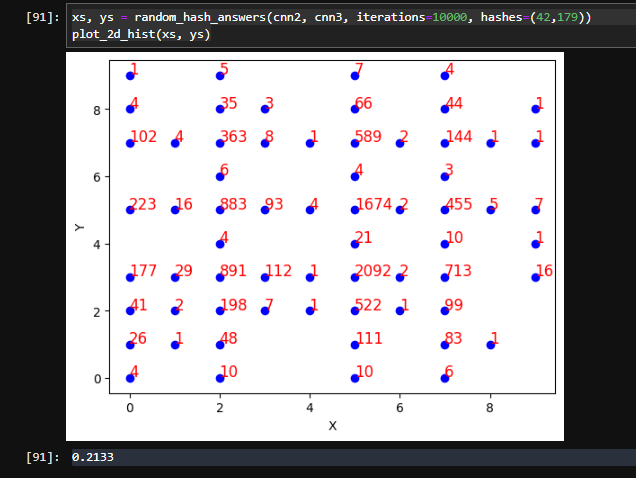


рис 7

одинаковые cnn с переставлением пикселей

