Исследование:

Изучение Ответов Маленьких Нейросетей на Случайные Выходные Данные

## **Актуальность**

Скоррелированность ответов различных нейросетей на случайные входные данные может свидетельствовать о том, что нейросети выявляют одни и те же скрытые признаки (hidden features). Это, в свою очередь, может привести к бесполезности использования ансамбля таких нейросетей, поскольку они будут давать схожие предсказания. Более глубокое понимание причин скоррелированности ответов нейросетей на данные, не похожие на обучающую выборку, может помочь создавать более эффективные ансамбли нейросетей, улучшая их обобщающую способность.

**Возможные причины скоррелированности:**

1. Архитектура нейросети.
2. Обучающая выборка и её погрешности.
3. Тип решаемой задачи.

## **Цель исследования**

Цель исследования — оценить влияние архитектуры нейросети, особенностей обучающей выборки и типа задачи на скоррелированность ответов нейросетей на случайные входные данные.

## **Метод**

Для исследования были выбраны две классические нейросети и сверточная нейросеть (CNN). Обучение сетей проводилось на датасете MNIST, который содержит изображения рукописных цифр.

### **Этапы исследования:**

1. **Обучение нейросетей:**
   * Обучим две классических нейросети на MNIST (задача распознавания цифр). Пример входной картинки приведен на рисунке 1.
2. **Анализ скоррелированности ответов:**
   * Проведем тестирование на случайных входных данных (белый шум) и построим двумерную гистограмму ответов (рисунок 2). Число у каждой точки на гистограмме показывает количество случайных примеров, на которые первая нейросеть ответила значением x, а другая — значением y.
   * Полученный результат показывает, что общий процент одинаковых ответов составляет 81%. Если бы корреляции не было, этот процент составил бы около 10%. Это указывает на высокую предвзятость обеих сетей относительно белого шума.
3. **Исследование влияния архитектуры:**
   * Проведем аналогичный эксперимент с CNN (рисунок 3). Несмотря на смену архитектуры, результат оказался похожим — корреляция составляет 76%. Было выявлено, что обе сети в 90% случаев распознают белый шум как цифру "8". Это связано с тем, что изображения в датасете MNIST имеют черный фон, и цифра "8" часто занимает больше места на изображении. Сверточные сети научились ассоциировать яркие изображения с цифрой "8".
   * Для устранения этой предвзятости, в дальнейшем эксперименты с CNN проводились на изображениях с уменьшенной яркостью каждого пикселя до 50%. В таких условиях ответ "8" отсутствовал.
4. **Сравнение скоррелированности на случайных примерах:**
   * В эксперименте с уменьшенной яркостью, скоррелированность ответов CNN на случайные входные данные составила 60% (рисунок 4).
   * Однако, ответы CNN и классических нейросетей оказались несвязанными (рисунок 5). Это объясняется тем, что для классической нейросети расположение пикселей не имеет значения, в то время как для CNN информация о расположении пикселей «вшита» в архитектуру.
5. **Исследование эффекта случайного переставления пикселей:**
   * Если случайным образом, но одинаково для всех изображений, переставить пиксели в обучающем датасете, CNN будет интерпретировать эту информацию по-другому. CNN, обученные на таких перемешанных картинках, фактически имеют другую структуру, что нивелирует их преимущества перед классическими сетями.
   * Удивительно, что такие нейросети показывают высокую точность на тестовых данных (98%), но не имеют корреляции в ответах на случайные входные данные (21%). Этот эффект не связан с количеством нейронов, так как аналогичные модели без перестановки пикселей демонстрируют высокую корреляцию.

## **Заключение**

Исследование показало, что скоррелированность ответов нейросетей на случайные входные данные может быть обусловлена как архитектурой сети, так и особенностями обучающей выборки. CNN оказываются более чувствительными к изменению структуры данных, в то время как классические нейросети не проявляют зависимости от расположения пикселей. Эти выводы могут быть использованы для разработки более эффективных ансамблей нейросетей, минимизируя взаимную корреляцию их предсказаний и улучшая общую производительность.