

УДК 004

**ПРОГРАММНАЯ СИСТЕМА ДЛЯ ИССЛЕДОВАНИЯ СТАТИСТИЧЕСКИХ
ЗАКОНОМЕРНОСТЕЙ, ПРОИСХОДЯЩИХ В СЛОЯХ НЕЙРОННОЙ СЕТИ**Лоточ Е.Ю.¹. аспирант гр. 2.3.1-24, II курсНаучный руководитель: Лемешко Б.Ю.¹. д.т.н. профессор¹Новосибирский государственный технический университет
г. Новосибирск

История развития нейросетей берёт своё начало в 1945 году [1], а область машинного обучения начинает активно внедряться в повседневную жизнь с 90-х годов прошлого столетия. За это время произошёл большой скачок технологий, и сейчас трудно представить современные системы без применения методов компьютерного зрения, распознавания речи, генерации текста и быстрого поиска. Долгое время было принято считать нейронные сети своего рода “чёрным ящиком” с измеримыми входом и выходом, но неизвестным содержанием и поведением, но всё начало меняться с появлением подхода под названием “объяснимый ИИ” (XAI) [2]. Основная идея заключается в том, чтобы заставлять нейросеть вернуть определённые результаты в процессе обучения и на их основе построить взаимосвязи с другими переменными. Так, например, можно узнать иерархию признаков, концептуальные перекрытия, инвариантности, структуру пространства и другие важные особенности.

В настоящей работе речь пойдёт о визуализации скрытых представлений. В данном контексте нейросеть следует воспринимать как динамическую систему. Тогда анализ внутренних состояний превращается из сугубо прикладной в задачу математической статистики. Активации на выходах слоёв нейросети в таком случае рассматриваются, как колебание случайной точки. На этом основании можно считать методы, используемые для анализа стохастических процессов, применимыми для анализа активаций в слоях нейросетей. С целью реализации поставленной задачи был разработан инструмент, позволяющий получать подробную информацию о процессе обучения на основе активаций слоёв и наблюдать изменения в зависимости от эпохи обучения, а также от глубины и структуры слоя. С помощью разработанного программного обеспечения проведено несколько экспериментов, демонстрирующих влияние факторов на успешное обучение нейросети, что позволило выдвинуть ряд гипотез. Основой разработки послужили Python 3.9.13 и библиотека Tensor Flow. Сам же инструмент представляет собой настраиваемую Callback-функцию с возможностью выгрузки активаций в формате, совместимом с программой ISW [3]. Построение графиков и расчёты производятся непосредственно в процессе обучения в конце каждой эпохи. Программа позволяет получить следующие метрики:

- выборочное среднее арифметическое активаций нейронов скрытого слоя по всей валидационной выборке за фиксированную эпоху обучения;
- выборочную дисперсию активаций нейронов скрытого слоя по всей валидационной выборке за фиксированную эпоху обучения;
- спектральную норму (норму оператора) матрицы активаций скрытого слоя за фиксированную эпоху обучения;
- степень разреженности активаций скрытого слоя за фиксированную эпоху обучения;
- корреляционную фрактальную размерность множества точек в пространстве активаций скрытого слоя за фиксированную эпоху;
- взаимную информацию между активациями двух различных слоёв за фиксированную эпоху;
- аппроксимацию нормы якобиана отображения вход-выход для скрытого слоя по вариациям вдоль батча;
- квантильное расстояние между распределениями активаций текущего и предыдущего слоёв (или между текущим слоем и входными данными);
- многомасштабную энтропию (энтропию активаций на масштабах прореживания, возвращая спектр энтропии, как функцию масштаба);

- коэффициент участия (меру эффективной размерности пространства активаций скрытого слоя через сингулярные числа);
- нормированную сложность по Лемполею-Зиву для последовательности активаций скрытого слоя;
- профиль аппроксимации низкого ранга для матрицы активаций скрытого слоя (мера того, насколько хорошо данные представляются подпространствами различной размерности);
- логарифмическую чувствительность нормы активаций к малым случайным возмущениям;
- Дифференциальная энтропия распределения активаций скрытого слоя, аппроксимированная через гистограмму
- энтропию Шеннона нормированного спектра сингулярных чисел якобиана (мера того, насколько равномерно распределена чувствительность слоя по разным направлениям);
- среднюю абсолютную межнейронную корреляцию (меру линейной связности активности и разных нейронов в слое);
- индекс фрустрации (меру конфликта между направлениями обновлений для разных примеров в батче, основанную на косинусном расстоянии между градиентами);
- Проверку нормальности по критериям Андерсона-Дарлинга, Шапиро-Уилка, Колмогорова-Смирнова, Лиллиефорса, Харке-Бера, Д'Агостино, Хи-квадрат Пирсона.

И построить следующие графики:

- информационного потока по слоям;
- стабильности по глубине;
- критических точек преобразований;
- динамики обучения по эпохам;
- многомасштабной энтропии;
- профилей нейронов;
- “сырых” распределений;
- плотностей распределений;
- функций распределений;
- корреляционной матрицы метрик;
- эволюции функции распределения по слоям;
- эволюции функции распределения по эпохам.

С помощью разработанного инструмента произведён анализ 2 архитектур нейросетей: CNN с двумя свёрточными слоями и Dense с 5 плотными слоями на 3 различных наборах данных, реализующих задачу многоклассовой классификации. Полученные распределения активаций на выходах слоёв показаны на рис. 1.

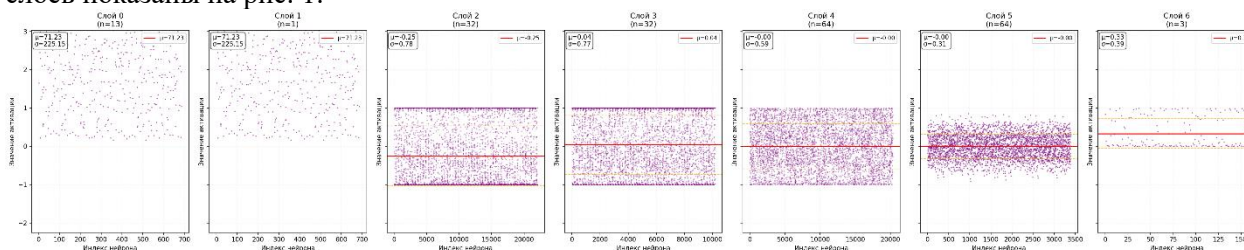


Рис.1. Сырые распределения значений активаций для валидационной выборки CNN – Набор данных: wine.

Слой 0 соответствует входным данным, значения по оси x представляют собой реальные значения признаков. Слой 1 в CNN представляет собой Reshape, поэтому форму данных не изменя-

ет, он необходим для отсеечения лишних размерностей. По распределениям, представленным выше, можно получить функции распределения.

Ниже представлены графики эволюции функций распределения в различных слоях в зависимости от эпохи для различных наборов данных и архитектур (рис.2-6).

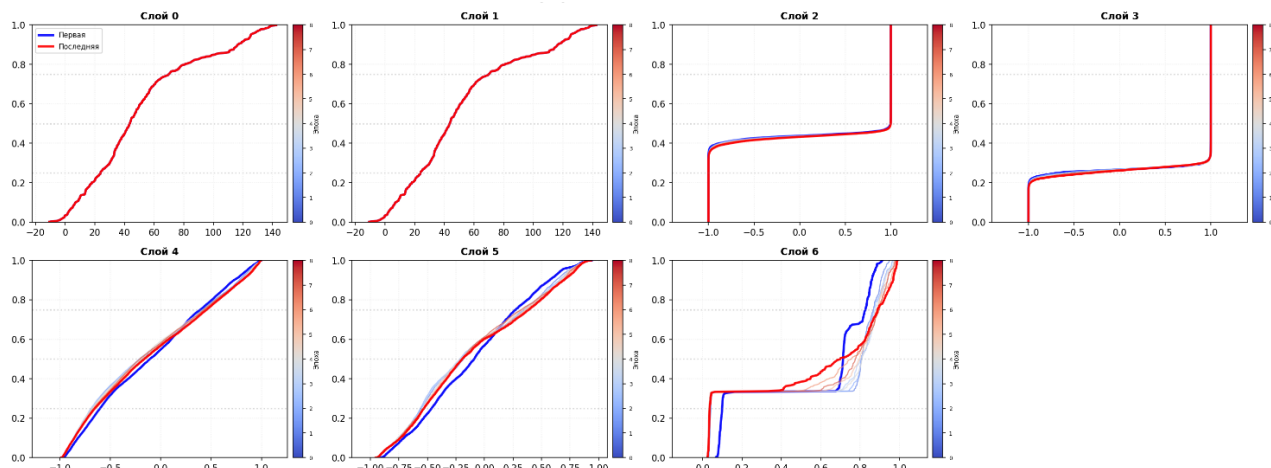


Рис.2. Функции распределения для CNN – Набор данных: spine. Достигнутая точность на валидации: 0.8

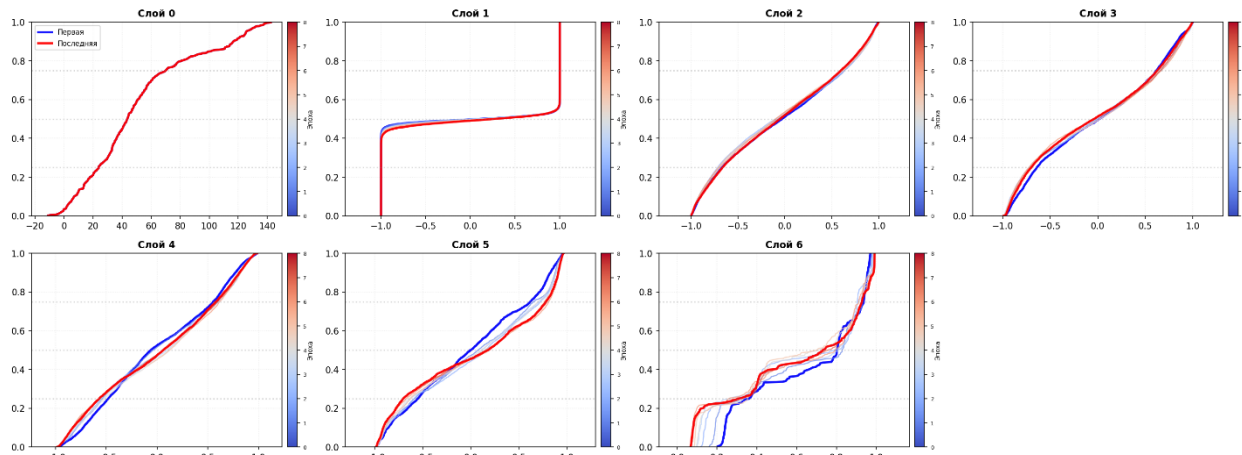


Рис.3. Функции распределения для Dense 5 – Набор данных: spine. Достигнутая точность на валидации: 0.84

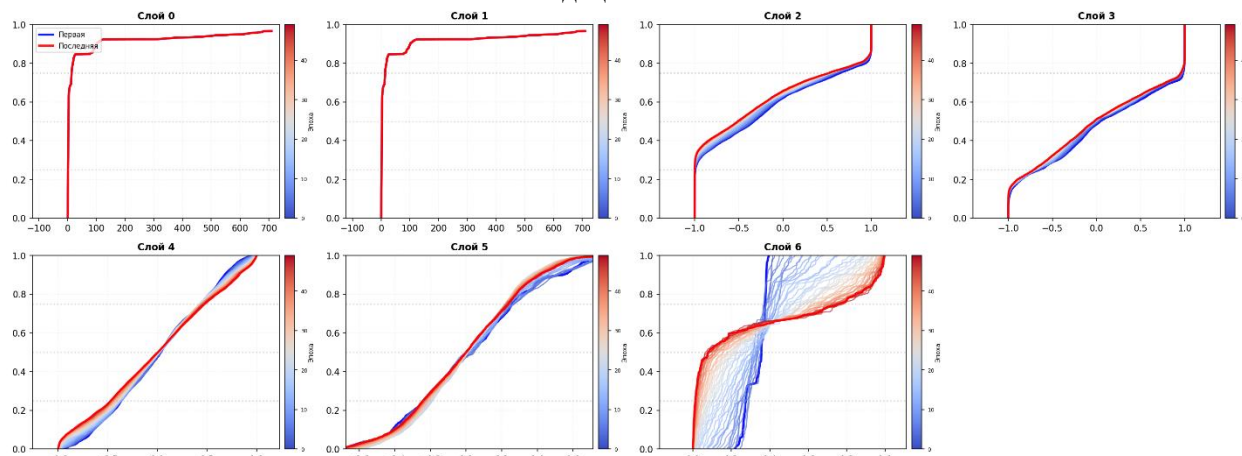


Рис.4. Функции распределения для CNN – Набор данных: wine. Достигнутая точность на валидации: 0.91

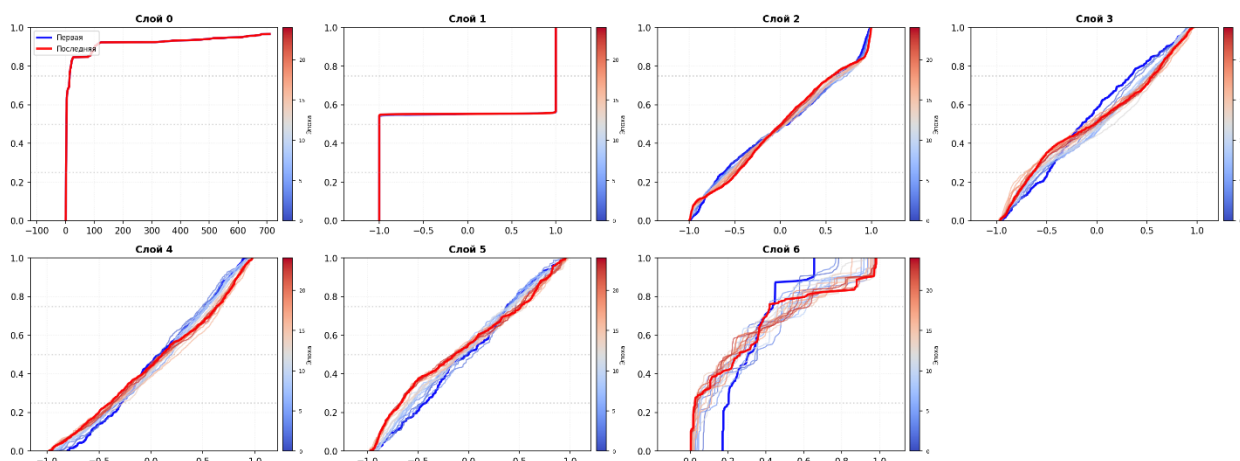


Рис.5. Функции распределения для Dense 5 – Набор данных: wine. Достигнутая точность на валидации: 0.77

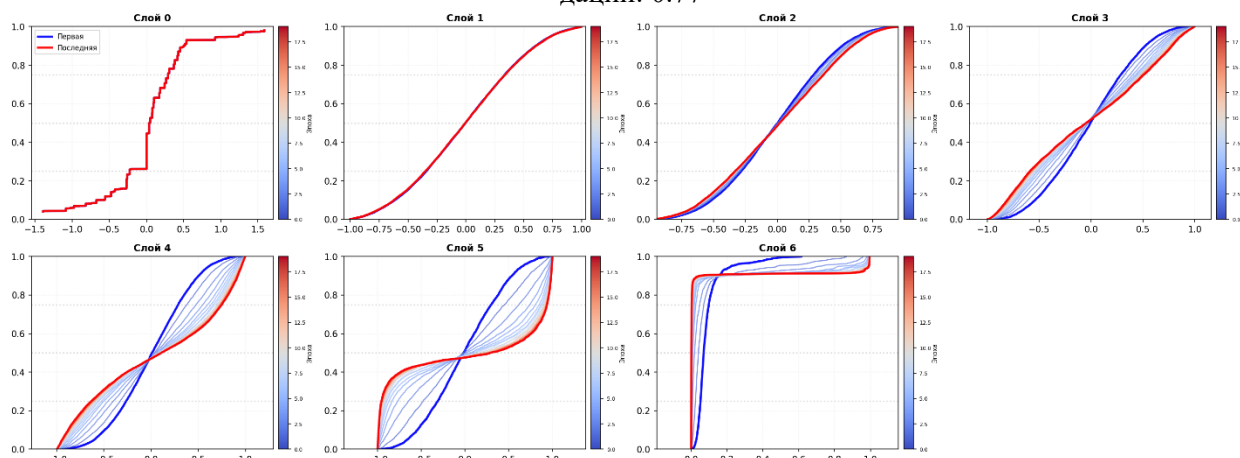


Рис.6. Функции распределения для Dense 5 – Набор данных: anneal. Достигнутая точность на валидации: 0.98

Глядя на приведённые графики, можно сделать эмпирический вывод о том, что слои нейросети реализуют последовательность преобразований, которые переводят произвольное распределение входных активаций в почти вырожденное распределение, либо в распределение Бернулли, соответствующее решению задачи.

Для подтверждения этого факта следует обратить внимание на дисперсию в представленных выходных слоях в начале и в конце обучения (таблица 1). Здесь каждое значение – дисперсия нейрона. Количество нейронов соответствует числу классов. Набор Anneal примечателен тем, что, используемый при обучении остальных нейросетей, он показал себя крайне негативно для CNN (видно по точности). В этой связи был проведен анализ графиков и выдвинута гипотеза о том, что промежуточные слои не успели преобразовать входные данные к выходу достаточным образом и увеличение числа слоёв должно улучшить результат. Как видно в таблице – это действительно привело к необходимому результату.

Наблюдается нормализующий эффект от $\tanh$ в глубоких слоях. Происходит ребалансировка дисперсий. При этом дисперсии часто становятся значительно больше, а не стремятся к 0, значит распределение не вырождается, из чего следует то, что выходное распределение активаций нейронов стремится к распределению Бернулли.

Таблица 1. – Дисперсия в выходном слое нейросети.

Архитектура	Набор данных	Первая эпоха	Последняя эпоха	Разница в процентах (%)	Точность в первой эпохе	Точность в последней эпохе
CNN	Spine	0.0028053943533450365; 0.02287546917796135; 2.8409052902134135e-05	0.0008465559221804142; 0.00019517859618645161; 0.00014864128024782985	-70 -99 -423	0.68	0.80
	Wine	0.00022902652563061565; 2.244710958620999e-05; 0.00026951610925607383	0.1414676159620285; 0.1476612240076065; 0.1559744030237198	61720 657719 57788	0.53	0.91
	Anneal	0.0001514001633040607; 0.0002032575139310211; 0.0015454025706276298; 0.00013146499986760318; 0.00018925136828329414	3.1356077670352533e-05; 0.005506979767233133; 0.006519818212836981; 0.000580330437514931; 0.00024414065410383046	-79 2609 322 341 29	0.77	0.77
CNN-big	Anneal	1.522353159089107e-05; 0.0032595281954854727; 0.005539342761039734; 0.00018286568229086697; 9.312052134191617e-05	0.00010538738570176065; 0.05139503255486488; 0.10628917068243027; 0.06623898446559906; 0.024888725951313972;	592 1477 1819 36120 26621	0.77	0.90
Dense 5	Spine	0.005077083595097065; 0.020051134750247; 0.005016537383198738	0.008107195608317852; 0.042148053646087646; 0.01695171743631363	60 110 238	0.77	0.84
	Wine	0.03630166873335838; 0.005321428645402193; 0.01425782311707735	0.4347308576107025; 0.37884092330932617; 0.186428040266037;	1097 7017 1207	0.53	0.77
	Anneal	0.0037179121281951666, 0.019884271547198296, 0.05258729308843613, 0.016679590567946434, 0.012406956404447556	0.001379130408167839; 0.08254055678844452; 0.15669111907482147; 0.06352964043617249; 0.024312375113368034	-62 315 198 281 96	0.68	0.98

Проверим, насколько это справедливо. Взглянем на графики плотности распределения для первой и последней эпохи самого показательного случая для набора Anneal с архитектурой Dense 5 (рис. 7 - 8).

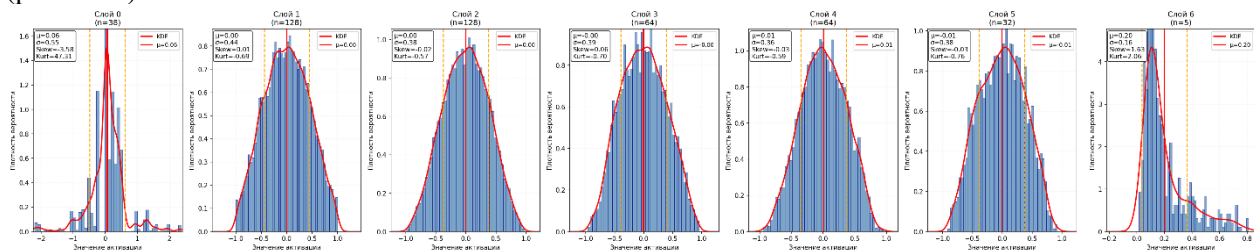


Рис. 7. Плотность распределения - Набор: Anneal; Архитектура: Dense 5. Первая эпоха

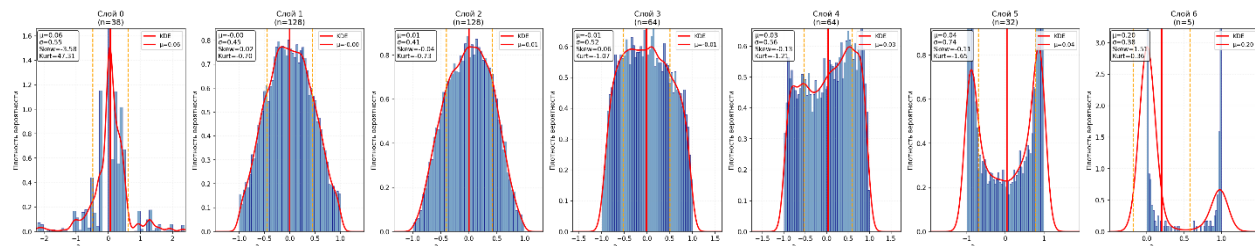


Рис. 8. Плотность распределения - Набор: Anneal; Архитектура: Dense 5. Последняя эпоха
Видно, что хвосты утяжелились.

Таблица 2 – процент бинарных значений для нейросети

Архитектура	Набор дан-ных	Эпоха пер-вая	Эпоха в первой тре-ти	Средняя эпоха	Эпоха в последней трети	Последняя эпоха
CNN	Spine	0%	38.5%	39.6%	39.2%	46.9%
	Wine	0%	0%	16.4%	39.6%	58.5%
	Anneal	0%	49%	32.2%	34.7%	36.0%
CNN-big	Anneal	39.8%	64.3%	70.2%	74.2%	76.8%
Dense 5	Spine	15.4%	15.0%	16.1%	19.4%	19.8%
	Wine	0%	37.7%	12.6%	40.3%	39%
	Anneal	5.5%	80%	85.2%	88%	90.8%

Во всех нейросетях за исключением Anneal на CNN и Spine на Dense 5 прослеживается, что чем выше финальная бинарность, тем выше финальная точность (таблица 2). Исключения обусловлены характером обучения, так как оба варианта имеют плохую динамику (рис.9) (рис.10):

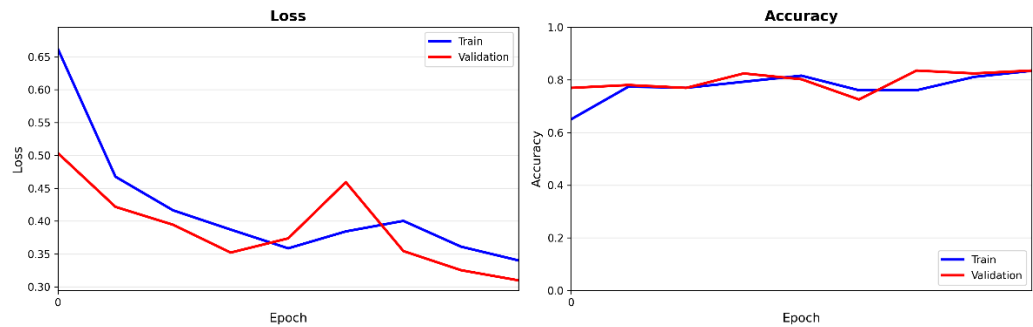


Рис.9. Динамика обучения нейросети Dense_5 на наборе данных Spine

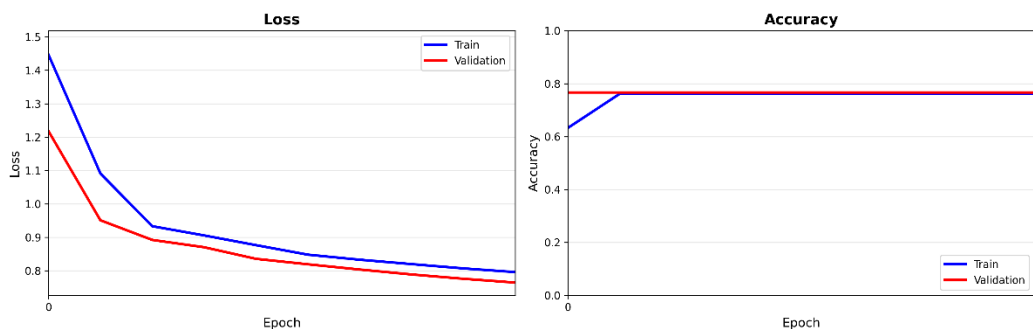


Рис.10. Динамика обучения нейросети Dense_5 на наборе данных Spine

Разработанный инструмент позволяет на глубоком уровне исследовать нейросеть с точки зрения классической статистики, благодаря чему были выдвинуты гипотезы, справедливость которых может быть обоснована с помощью экспериментов на базе разработанной технологии.

Для хорошо обученных нейросетей в задачах многоклассовой классификации с tanh в качестве основной функции активации и softmax на выходе, слои нейросети преобразуют входное распределение, приближая его к распределению, асимптотически стремящемуся к распределению Бернулли. Первый слой реализует сглаживание входной функции распределения, а последующие пошагово приводят к увеличению числа бинарных активаций. Этот процесс обусловлен совместным дей-

ствием нескольких факторов: принципом информационного узкого места, согласно которому каждый слой стремится сохранить только минимально достаточную для классификации информацию, отбрасывая избыточную вариативность; эффектом послонного схождения, при котором нижние слои стабилизируются раньше верхних, создавая каскадное сжатие; требованием положительной пакетной энтропии для прохождения градиентов, что определяет границу, к которой стремятся представления; формой функции потерь (кросс-энтропия), создающей градиенты, направленные на увеличение разделимости классов. Совокупность этих факторов приводит к монотонному убыванию энтропии активаций при прохождении через слои сети, с асимптотическим приближением к распределению Бернулли в выходных слоях.

Полученные результаты согласуются с известными теоретическими положениями [4] о послонном схождении, информационном узком месте и требовании положительной пакетной энтропии. В дополнение к ним, в работе предложена метрика доли бинарных активаций, позволяющая количественно оценить степень приближения распределения к Бернулли, а также показана «прямая корреляция» между ростом бинарности и ростом точности классификации. Это подтверждается экспериментально: рост бинарности с 5.5% до 90.8% сопровождался ростом точности с 68% до 98%; рост бинарности с 39.8% до 76.8% — ростом точности с 77% до 90%; рост бинарности с 0% до 58.5% — ростом точности с 53% до 91%.

Список литературы

1. Фон Нейман Дж. Первый проект отчёта об EDVAC / Дж. Фон Нейман // IEEE Annals of the History of Computing: журнал. – 1993. – Октябрь – Т.15, вып. 4. – с. 27-75
2. Мишра П. Объяснимые модели искусственного интеллекта на Python / П. Мишра; пер. с английского С. Минц. – Москва: ДМК Пресс, 2022. – 298 с. – ISBN 978-5-93700-124-5
3. Лемешко Б.Ю. Программная система ISW (Interactive Statistical Workshop) // Свидетельство о регистрации программы для ЭВМ №2002610178. – М.: Роспатент, 2002. – 1с.
4. Шварц-Зив Р., Тишби Н. Открывая черный ящик глубоких нейросетей с помощью информации // arXiv. – 2017. – № 1703.00810. – URL: <https://arxiv.org/abs/1703.00810> (дата обращения: 24.03.2026).