Анализ устойчивости нейронных сетей к шумам

Абстракт.

Анализ возможностей нейросетей давать правильные ответы в шумных условиях, все ещё остается критически важной задачей в различных областях такие как: медицина, промышленность, беспилотные устройства. Для решений задачи анализа устойчивости нейронных сетей к шумам мы интерпритируем нейронную сеть, как иерархическую динамическую систему. Главное открытие нашей работы заключается в том, что мы не разделяем нейронные сети и Динамические системы. Это означает то, что к нейросети можно применить математический аппарат, который применим к динамическим сетям. Весь код находится на Github: https://github.com/companys1234/stability-analysis-of-neural-networks

Введение.

До нашей работы было несколько работ посвященных анализу устойчивости нейронных сетей[3,4]. Новаторство нашего подхода заключается в том,что для анализа мы используем математический аппарат теории динамических систем, теории катастроф, и теории бифуракций. Это может быть уместным поскольку мы считаем, что нейросеть можно описать как иерархическую динамическую систему. Мы анализируем устойчивость с помощью 5 методов, которые будут описанные далее. Мы тестируем наши методы на 4 моделях нейросетей. Мы предполагаем, что качество наших 5 метрик не уступает другим методам (SSIM, PSNR, и т.д) анализа устойчивости нейронных сетей, а также может открыть новые горизонты в сфере этих исследований

Связанные работы.

Transformer[2] - классическая архитектура для обработки последовательностей.

VIT[1] - трансформенная архитектура для обработки изображений.

SSIM - учитывает структуру изображения и лучше для восприятия. Формула: $SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{x,y}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$, где μ -среднее значение, σ - стандартное отклонение, $\sigma_{x,y}$ - ковариация, $C_{1,2}$ - константы стабилизации. Минусы: менее точен для сильных шумов

Accuracy drop - прямая оценка влияния шума на модель. Формула: $AccuracyDrop = Accuracy_{clean} - Accuracy_{noisy}$

Минусы: нужно знать точность на чистых данных, зависит от задачи.

PSNR - довольно простая метрика, проста для вычисления, хорошо отображает шумы, формула: $10 * \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$, где MAX_I - максимальное значение пикселя. MSE - среднеквадратическая ошибка. Минусы: плохо работает с локальными признаками, зависит от MSE .

Динамическая система - это математическая модель описывающая эволюцию во времени точки в множестве элементов(состояний), по определенному закону эволюции

Теория бифуракций динамических систем - это раздел математики, который изучает резкое изменение поведения динамических систем при небольшом изменении их параметров

Теория хаоса -

Методы.

Мы предлагаем класс методов основанных на теории хаоса, теории бифуракций, и теории динамических систем. Методы: Энтропийный анализ, анализ сингулярных значений.

Рассмотрим работу принцип работы этих методов.

Энтропийный анализ:

$h_{KS} = \sup_{P} \lim_{n\to\infty} \frac{1}{n} H(P \vee f^{-1}P \vee \ldots \vee f^{-n+1}P)$, где P - разбиение фазового пространства, H - энтропия шеннона.

Если, $h_{KS} > 0$ система в состоянии хаоса.

Энтропийный анализ иерархий позволяет: Количественно оценить степень хаотичности на каждом уровне, выявить критические точки перехода между порядком и хаосом, оптимизировать структуру системы.

Анализ сингулярных возмущений:

Система представляется в виде раздельных переменных:

$\epsilon \frac{dx}{dt} = f(x, y)$ (быстрые), $\epsilon \frac{dy}{dt} = g(x, y)$ (медленные)

Где $\epsilon \leq 1$ малый параметр.

Суть метода: позволяет анализировать где в системе процессы протекают медленнее, или быстрее.

Каскадные бифуракции:

Последовательное возникновение бифуракций, при изменение параметра, ведущее к усложнению динамики.

Логистическое отображение:

$x_{n+1} = rx_n(1 - x_n)$, где $x_n \epsilon [0,1]$, а $r \epsilon [0,4]$ - управляющий параметр

Этот метод позволяет анализировать бифуракции как в подсистемах, так и на глобальном уровне.

Многоуровневые экспоненты Ляпунова:

Метод анализа хаотичности в системе с помощью расчёта спектров Ляпунова для каждого подуровня.

Спектр Ляпунова(λ_i)- характеризует среднюю скорость схождения/ расхождения траекторий в фазовом пространстве.

Алгоритм Бенеттина(адаптированный для иерархий):

1 разделение системы на уровни

2 для каждой подсистемы строится матрица Якоби

3 QR-разложение матриц Якоби (локальный спектр Ляпунова)

4 расчёт спектра через предел:

$$\lambda_i^k = \lim_{t \to \infty} \frac{1}{t} \ln \left| \delta x_i^k(t) \right|$$

Фрактальная размерность - количественная мера сложности аттрактора динамической системы.

Наш метод расчёта - корреляционная размерность(D_2).

Формула:

$D_2 = \lim_{r \to 0} \frac{\log \ C(r)}{\log \ r}$ где: $C(r) = \frac{1}{N^2} \sum_{i \neq j} \theta \left(r - \left(x_i - x_j \right) \right)$ - кореляционный интеграл, θ - функция Хевисайда, r - радиус окрестности,

N - количество точек траектории.

Алгоритм:

1 Встраивание временного ряда(метод Такенса)

2 Построение графиков log C(r) и log r

3 Наклон линейного графика → D_2

Результаты.

Мы проверили на устойчивость 4 архитектуры нейронных сетей:

1 CNN-подобную модель(2 слоя свёртки с пулингом, 2 линейных слоя, активация ReLU)

2 Двухслойную MLP модель

3 VIT

4 Transformer

Интерпретация результатов:

Результаты CNN модели:

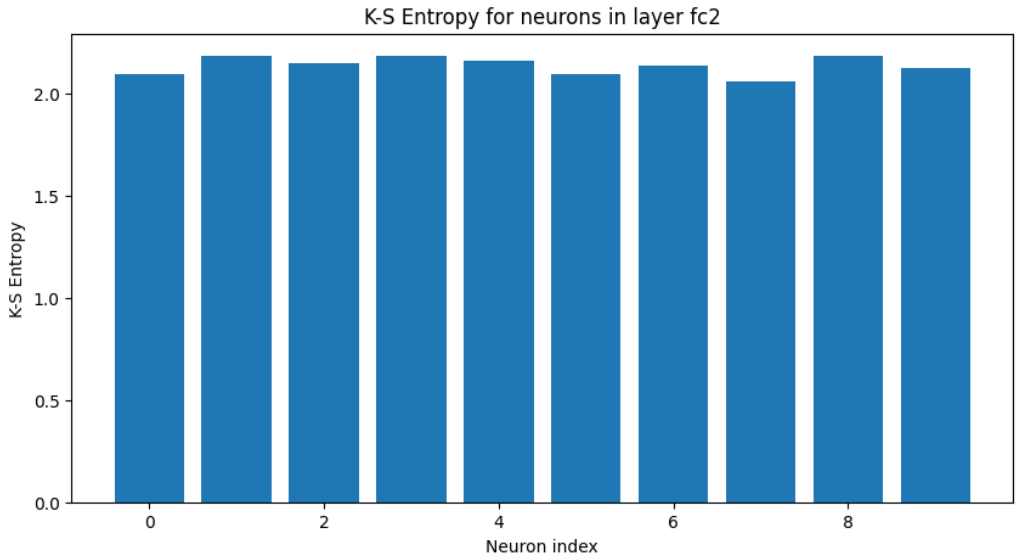


Рисунок 1. вычисление энтропии для CNN модели

Средняя энтропия для слоя FC2: 2.1398

Анализ сингулярных возмущений:

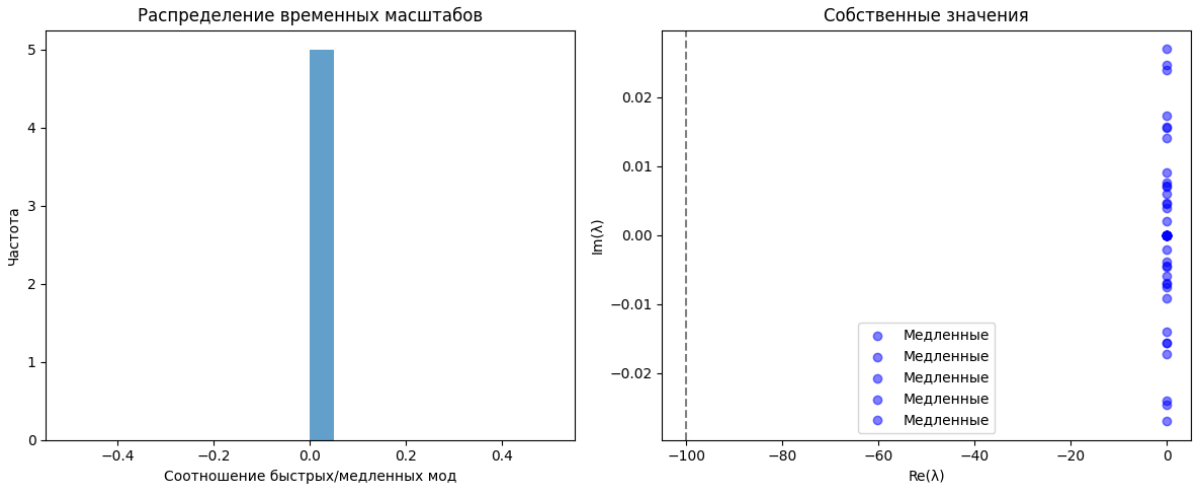


Рисунок 2. Анализ сингулярных возмущений для CNN модели.

Каскадные бифуракции:

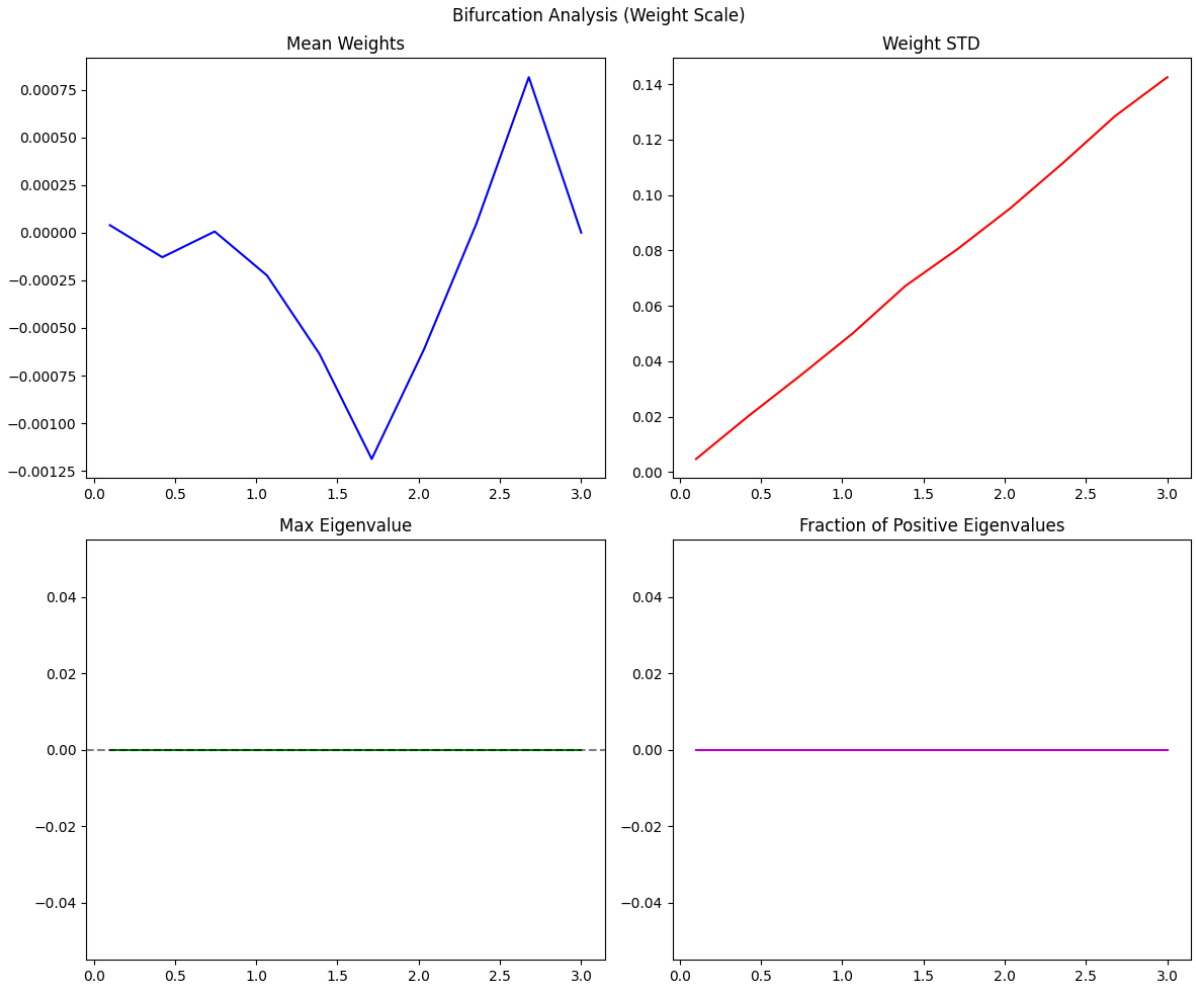
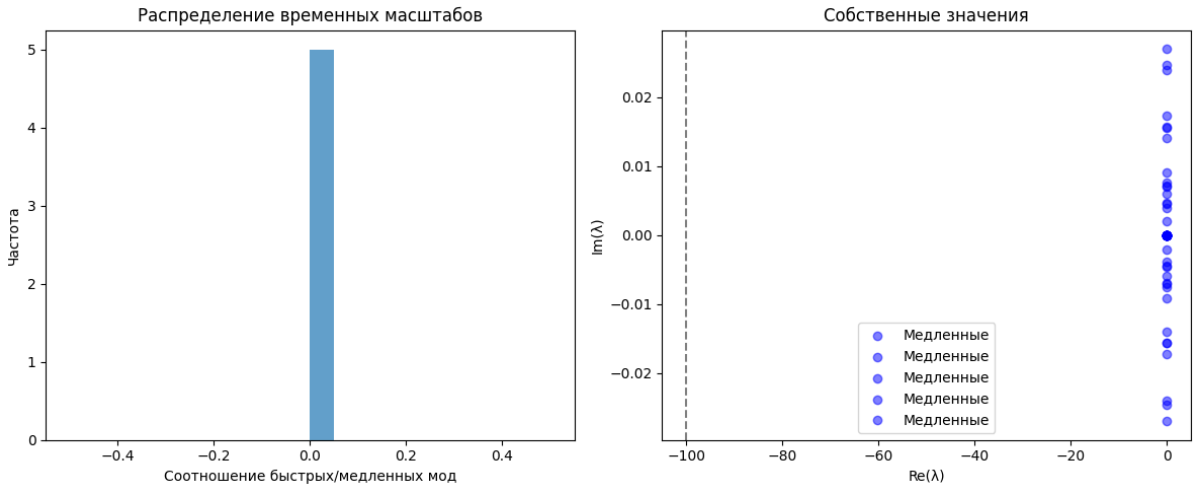


Рисунок 3. Каскадные бифуракции для CNN модели

Двухслойный MLP:

Анализ сингулярных значений:



Многоуровневые экспоненты Ляпунова и фрактальная размерность:

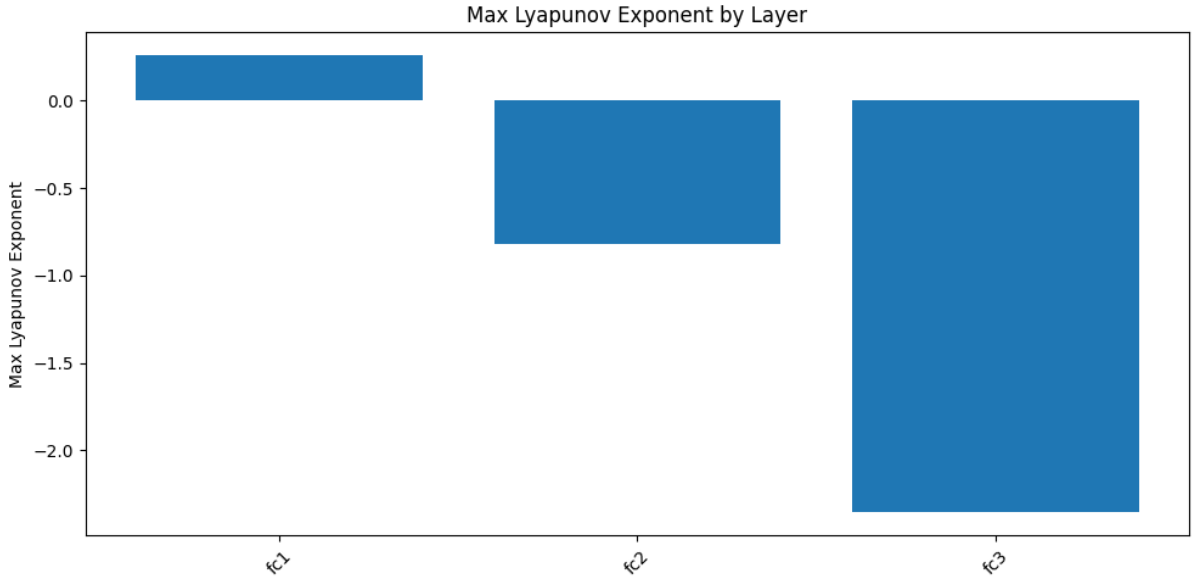


Рисунок 4. Многоуровневые экспоненты Ляпунова для CNN

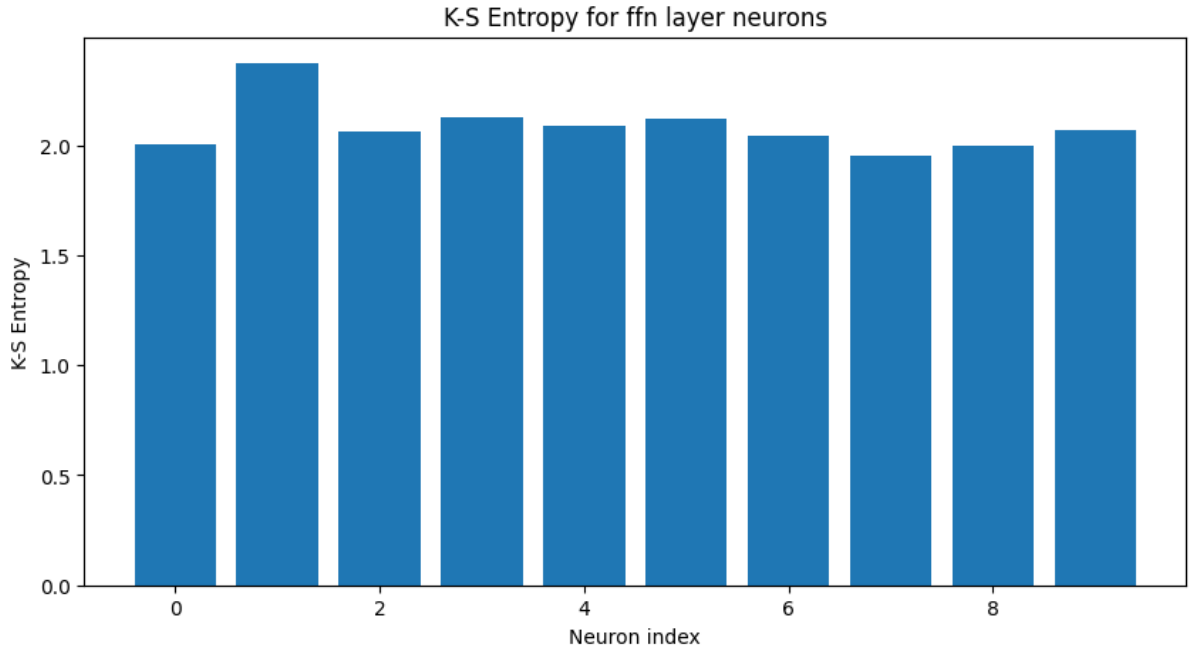
Фрактальная размерность:

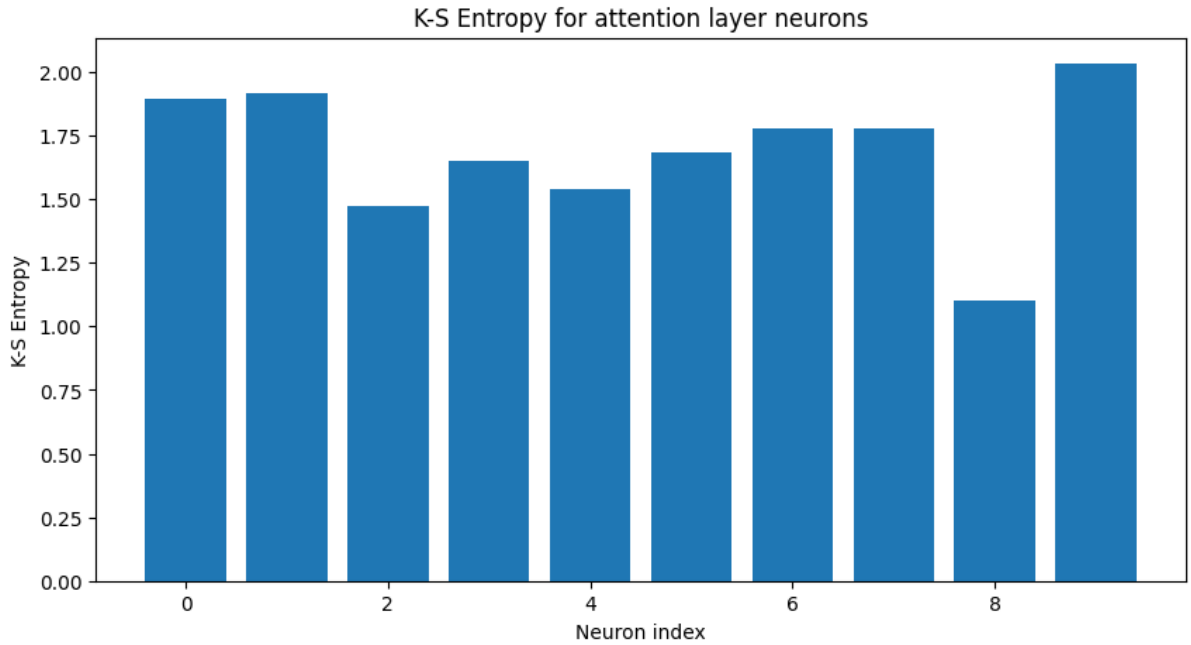
Слой 1: 6,90

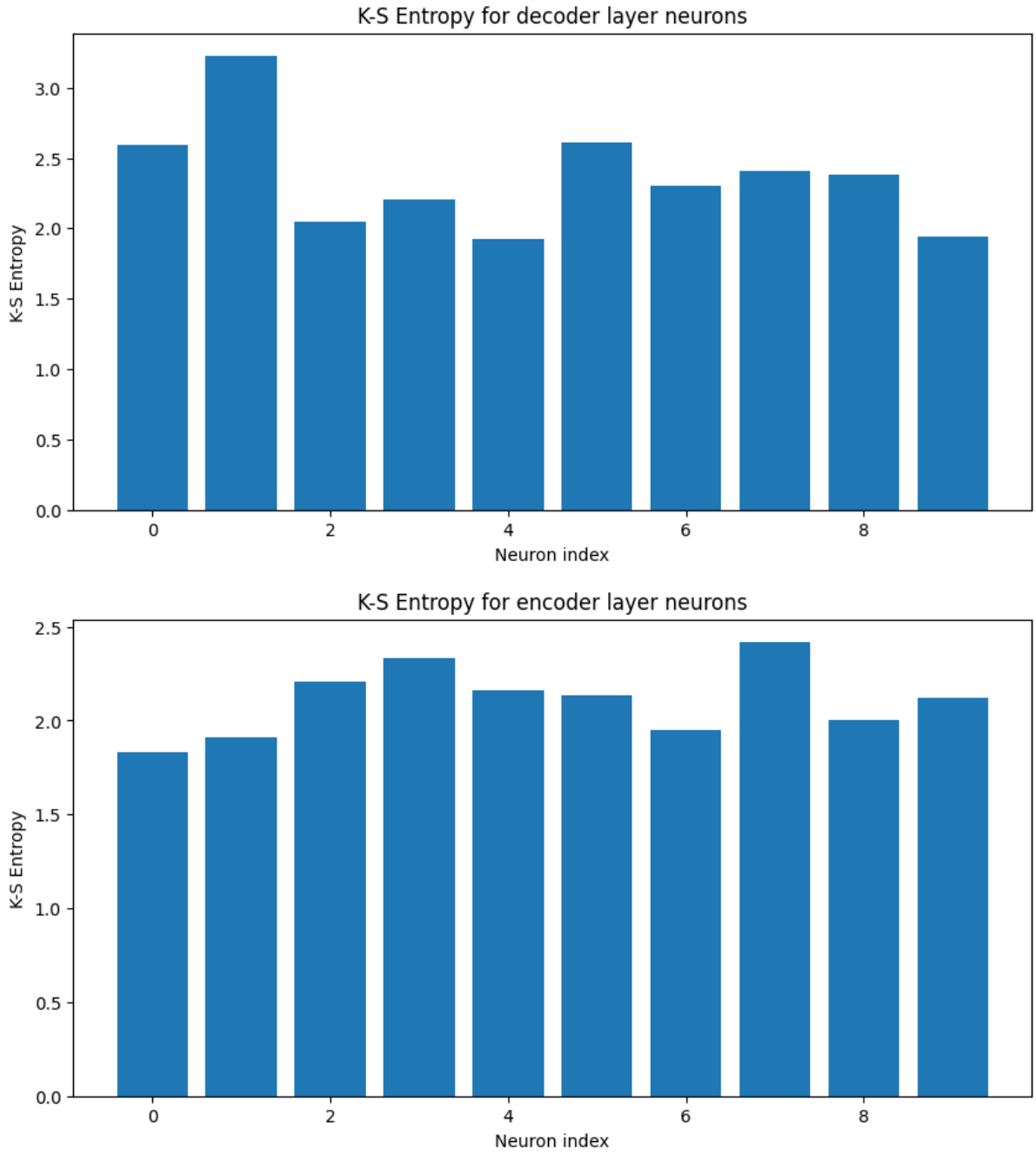
Слой 2: 8,10

Слой 3: 8,48

Transformer:



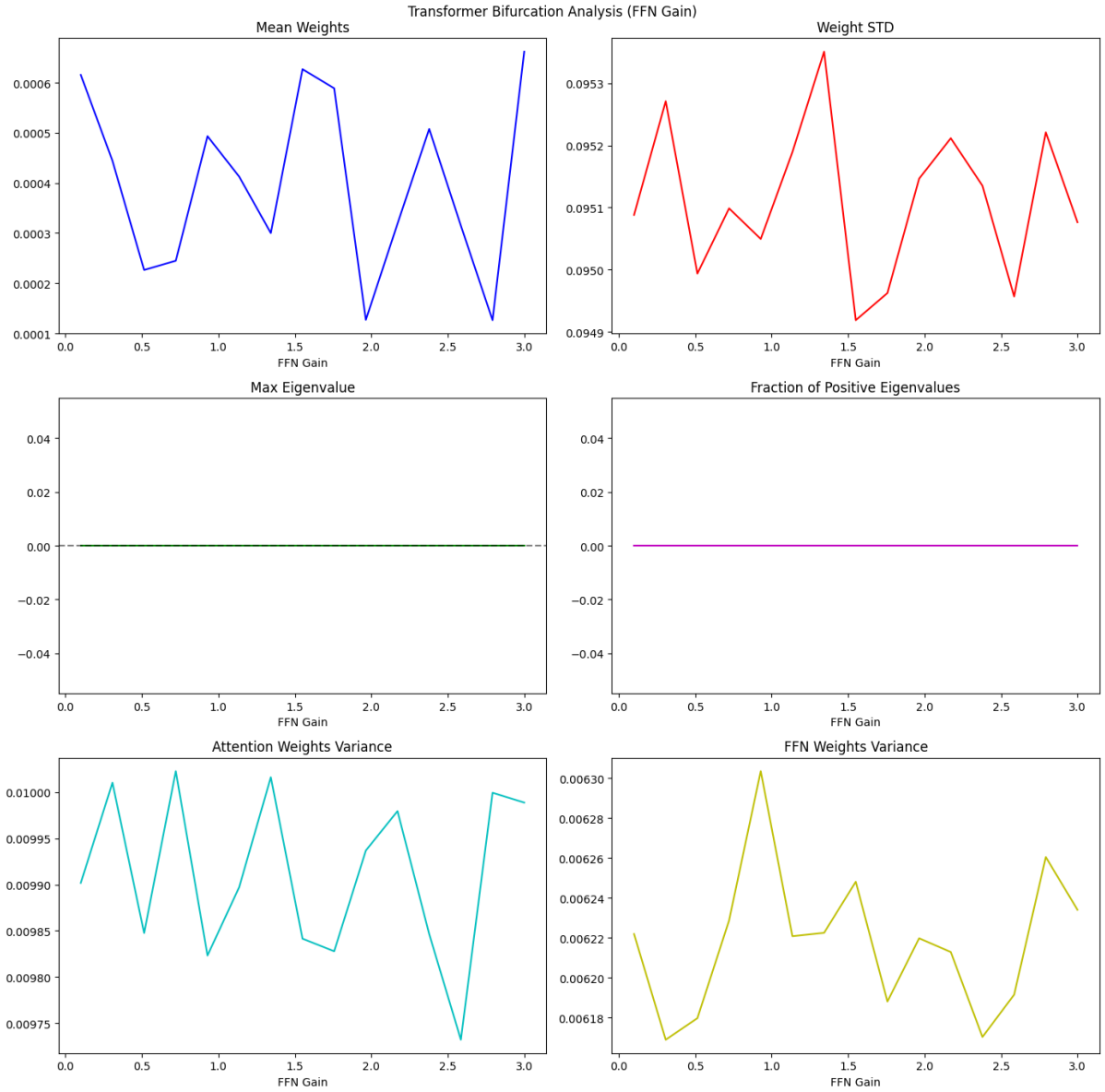


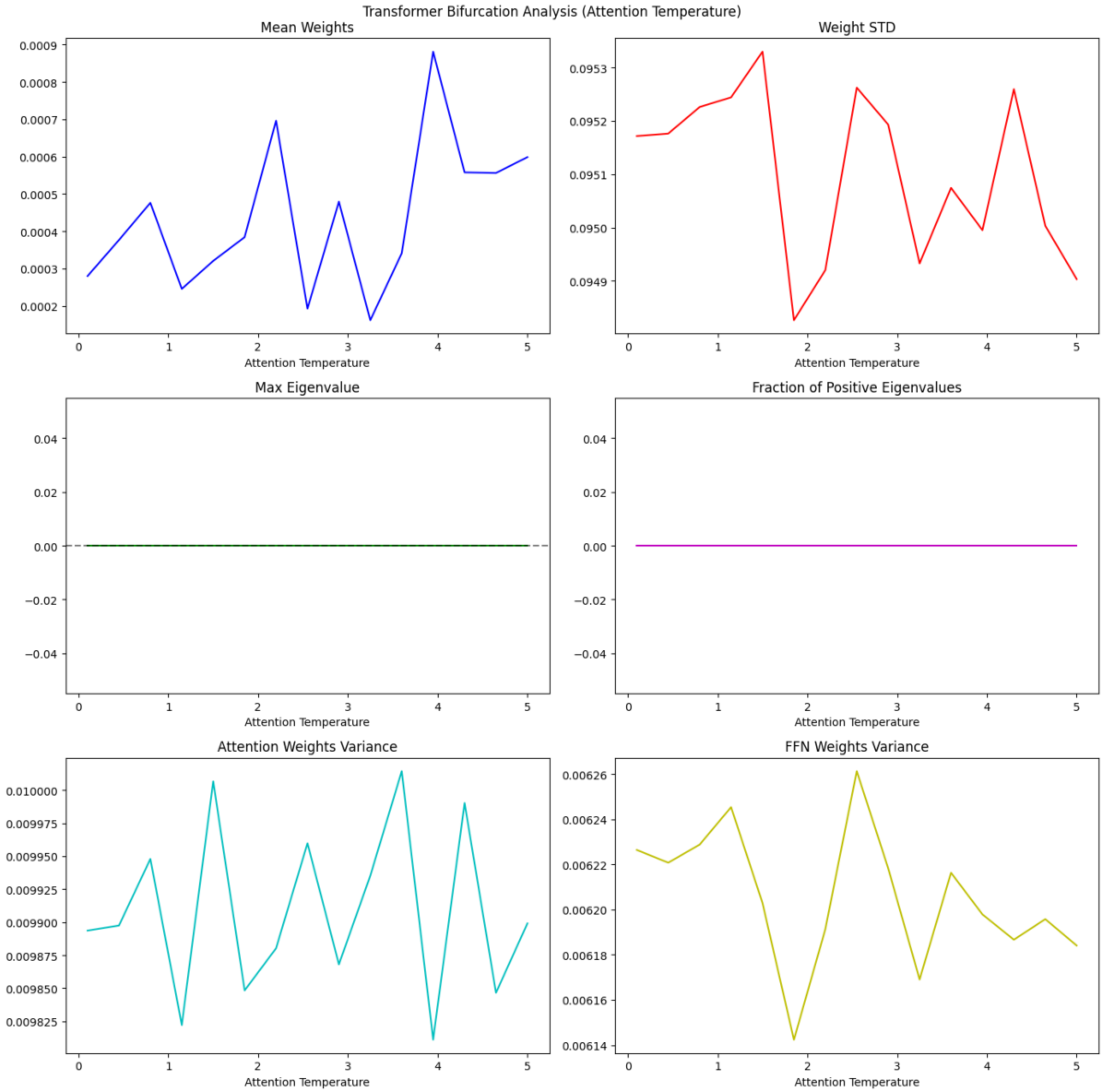


Рисунки 6,7,8,9. Энтропийный анализ для разных слоёв Transformer

Средние значения энтропии для слоёв: 2.1081, 2.3641, 1.6852, 2.0846

Каскадные бифуракции:


Transformer Bifurcation Analysis (FFN Gain)
Mean Weights
Weight STD
Max Eigenvalue
Fraction of Positive Eigenvalues
Attention Weights Variance
FFN Weights Variance
FFN Gain



Рисунки 10,11,12. Анализ каскадных бифуракций для transformer.

Многоуровневые экспоненты Ляпунова и фрактальная размерность:

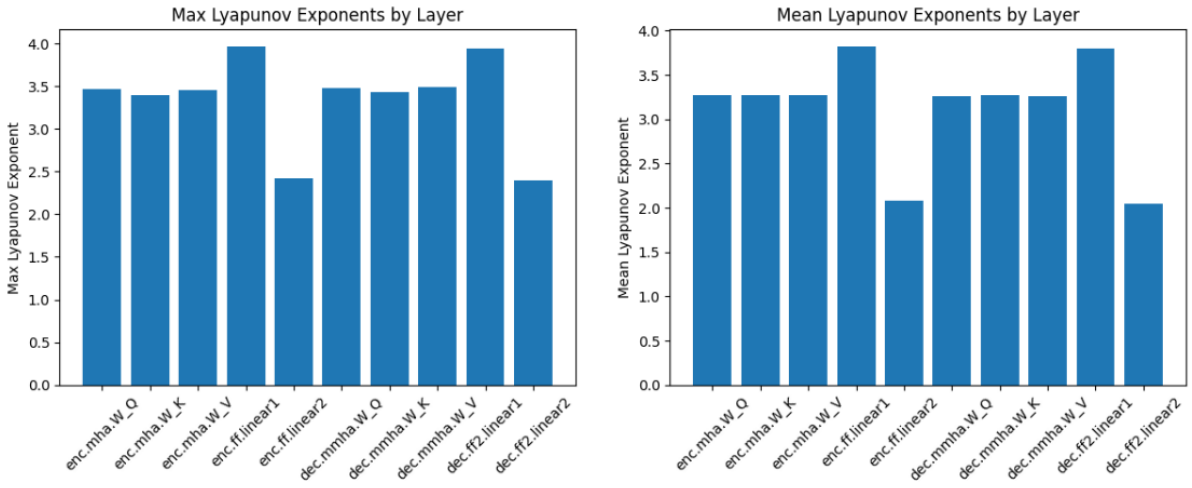


Рисунок 13. Многоуровневые экспоненты Ляпунова для Transformer.

VIT:

Энтропийный анализ:

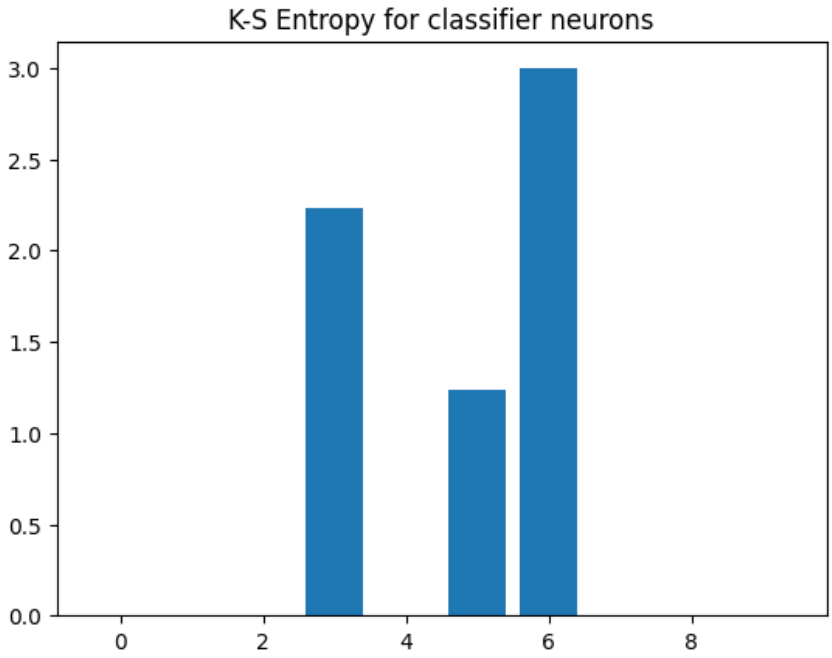


Рисунок 14. Энтропийный анализ для VIT

Анализ сингулярных значений:

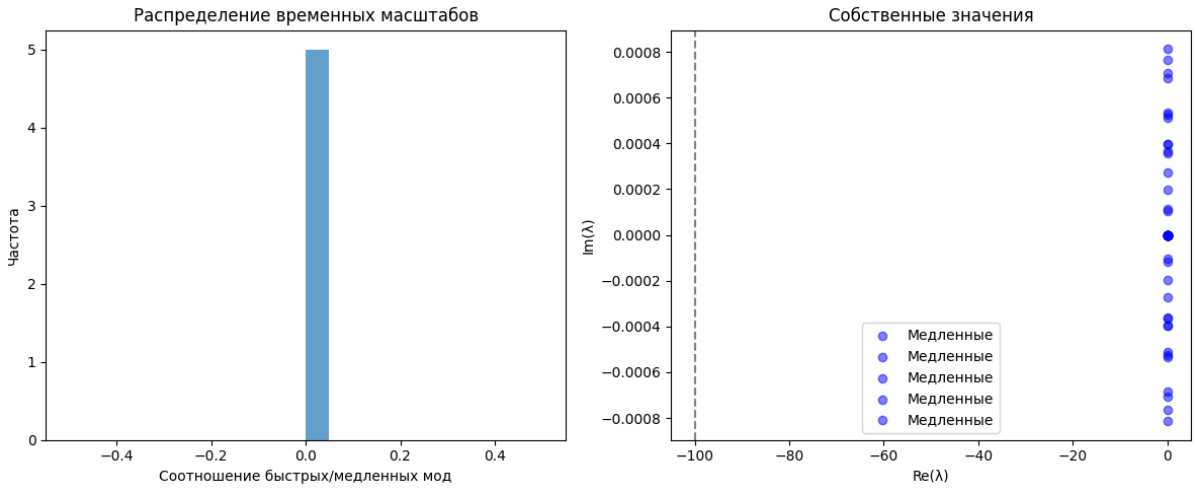
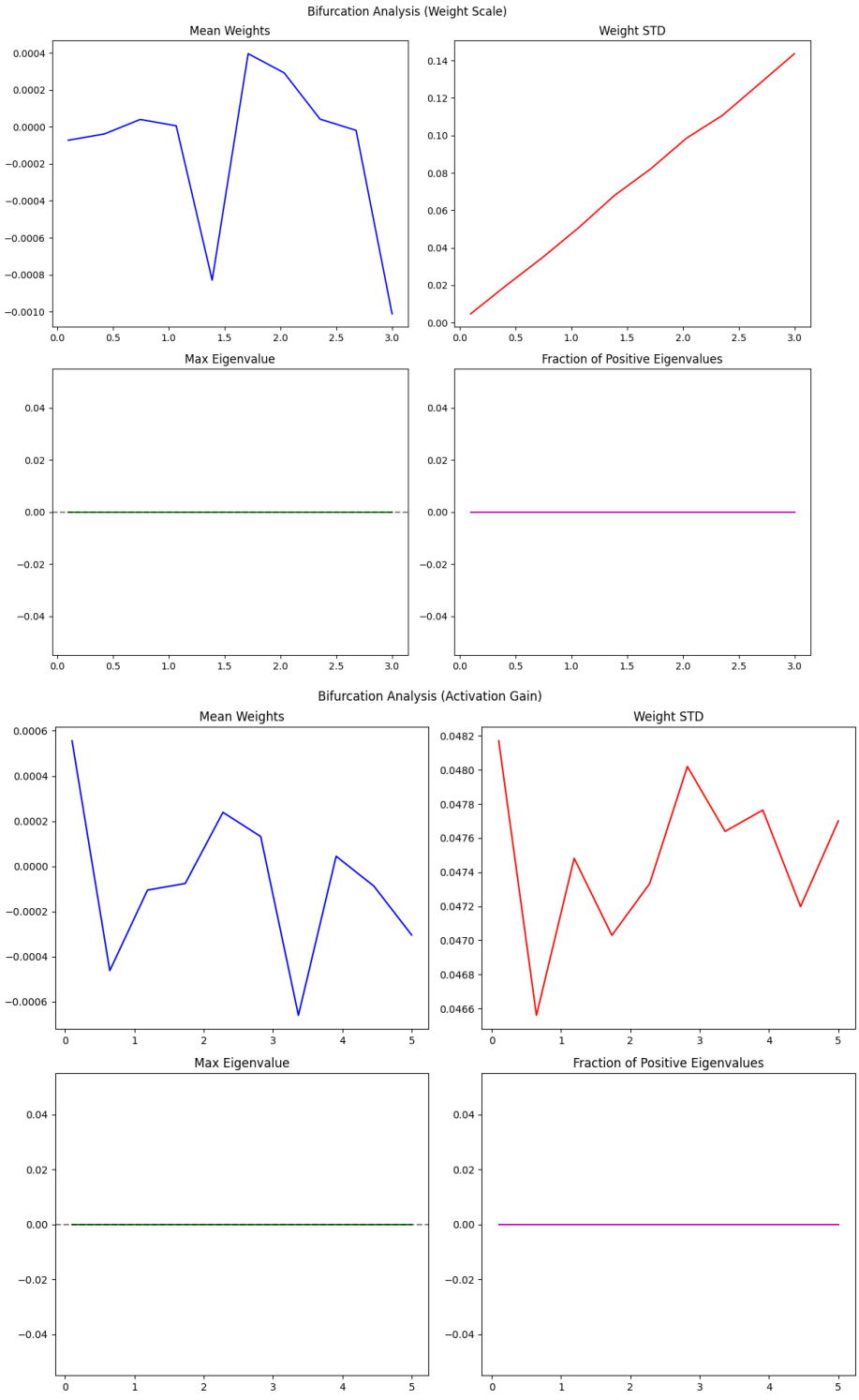


Рисунок 15. Анализ сингулярных значений для VIT

Каскадные бифуракции:



Рисунки 15,16. Каскадные бифуракции для VIT

Многоуровневые экспоненты Ляпунова:

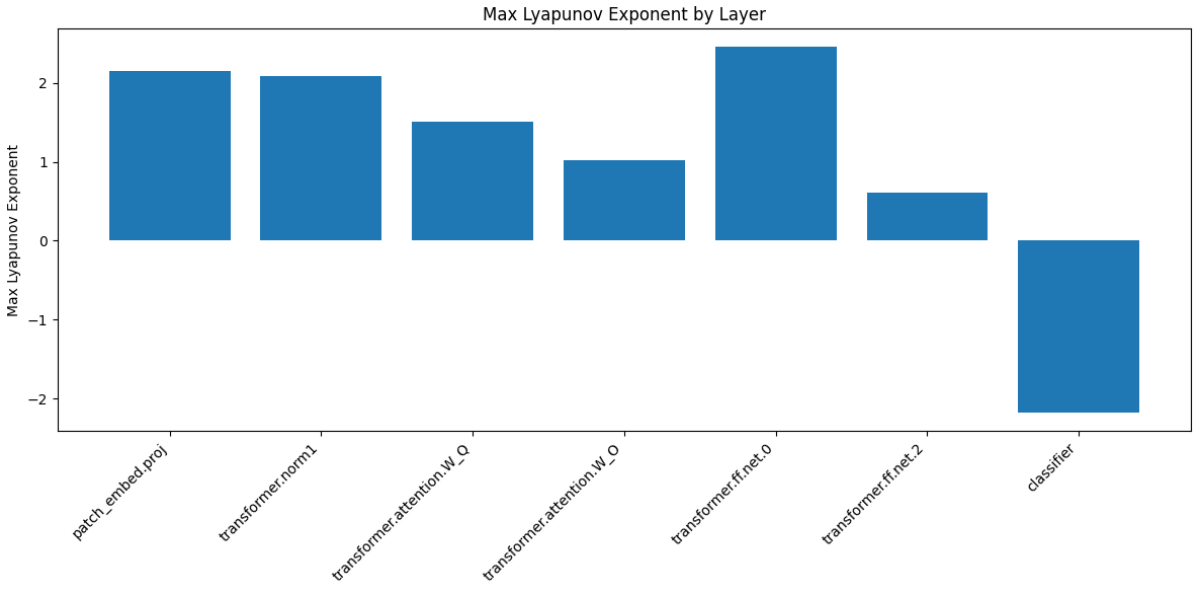


Рисунок 17. Максимальные экспоненты Ляпунова для VIT.

Заключение.

Методы анализа устойчивости нейронных сетей к шуму имеют большой потенциал, но при этом методы все ещё не универсальны. Каждый метод нужно подстраивать под определённую модель и архитектуру. Мы надеемся что в будущих работах будут использованы более новые и разнообразные методы анализа из теории хаоса, теории бифуракций или из теории динамических систем.А также будут проведен анализ с другими популярными архитектурами

Список литературы.

[1] - Dosovitskiy A. et al. An image is worth 16x16 words: Transformers for image recognition at scale //arXiv preprint arXiv:2010.11929. – 2020.

[2] - Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.

[3] - Хохлова Татьяна Наилевна Устойчивость полносвязной и звёздной структур нейронных сетей // Вестник ЮУрГУ. Серия: Математика. Механика. Физика. 2012. №34. URL:

https://cyberleninka.ru/article/n/ustoychivost-polnosvyaznoy-i-zvyozdnoy-struktur-neyronnyh-setey

[4] - Клестов Роман Андреевич, Клюев Андрей Владимирович, Столбов Валерий Юрьевич АЛГОРИТМИЧЕСКАЯ УСТОЙЧИВОСТЬ НЕЙРОННЫХ СЕТЕЙ ГЛУБОКОГО ОБУЧЕНИЯ ПРИ РАСПОЗНАВАНИИ МИКРОСТРУКТУРЫ МАТЕРИАЛОВ // Вестник ЮУрГУ. Серия: Компьютерные технологии, управление, радиоэлектроника. 2021. №1. URL: https://cyberleninka.ru/article/n/algoritmicheskaya-ustoychivost-neyronnyh-setey-glubokogo-obucheniya-pri-raspoznavanii-mikrostruktury-materialov (дата обращения: 14.08.2025).