Факультет Компьютерных Наук

Институт искусственного интеллекта и цифровых наук

Москва, 2025

Методы оценки ошибок генерации диффузионных моделей

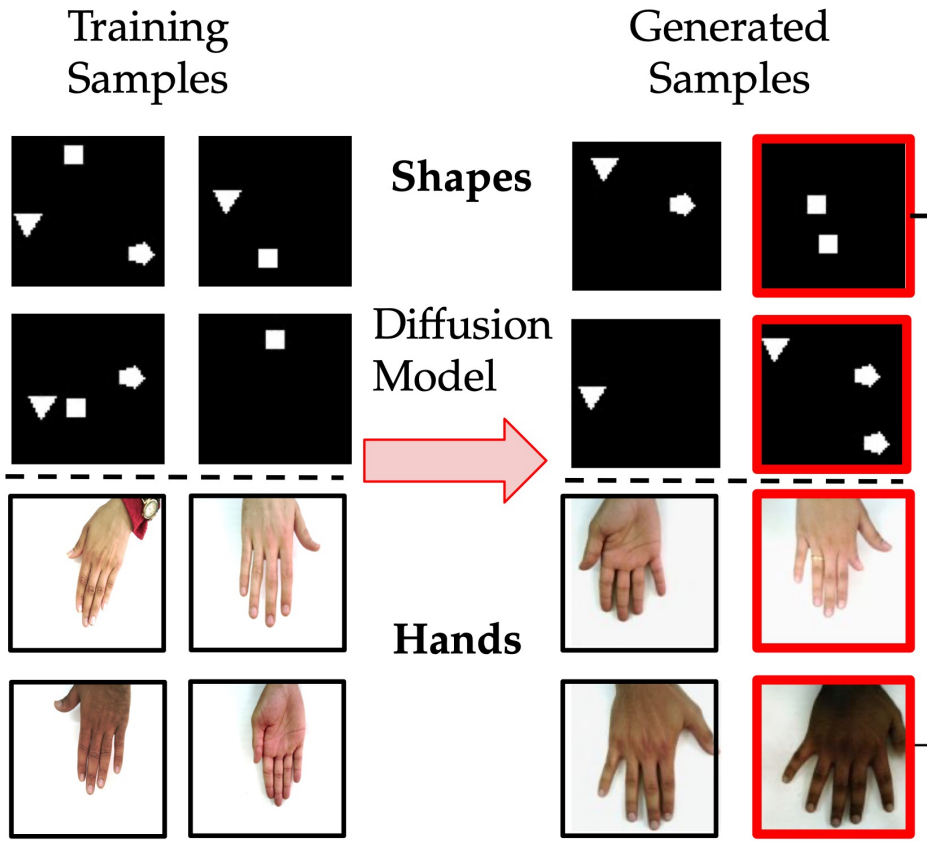
Фёдор Пахуров, стажер-исследователь, проектно-учебная лаборатория «Искусственный интеллект в математических финансах»

План доклада

- Проблема «галлюцинаций»
- Постановка задачи
- Обзор существующих подходов
- Выбранная методика (Complexity-Entropy)
- **Этап 1:** анализ временных рядов
- **Этап 2:** применение на изображениях
- Результаты и выводы
- Будущие направления

Проблема галлюцинаций

- Диффузионные модели –> фотореализм, НО генерируют несуществующие детали
- Искажение данных в прикладных задачах
- Нужен легкий детерминированный способ оценки



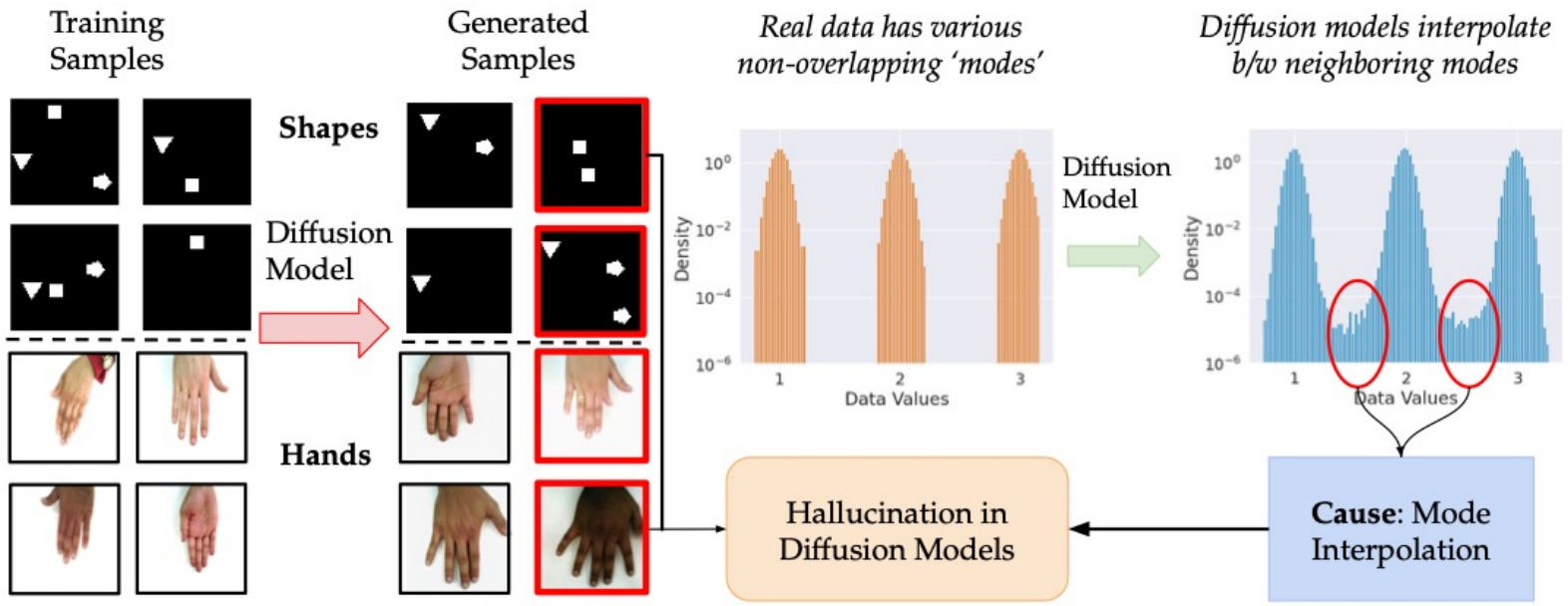
Examples of hallucinations from Sumukh K Aithal et al. Understanding hallucinations in diffusion models through mode interpolation.

(https://proceedings.neurips.cc/paper_files/paper/2024/file/f29369d192b13184b65c6d2515474d78-Paper-Conference.pdf.)

Планируемый результат

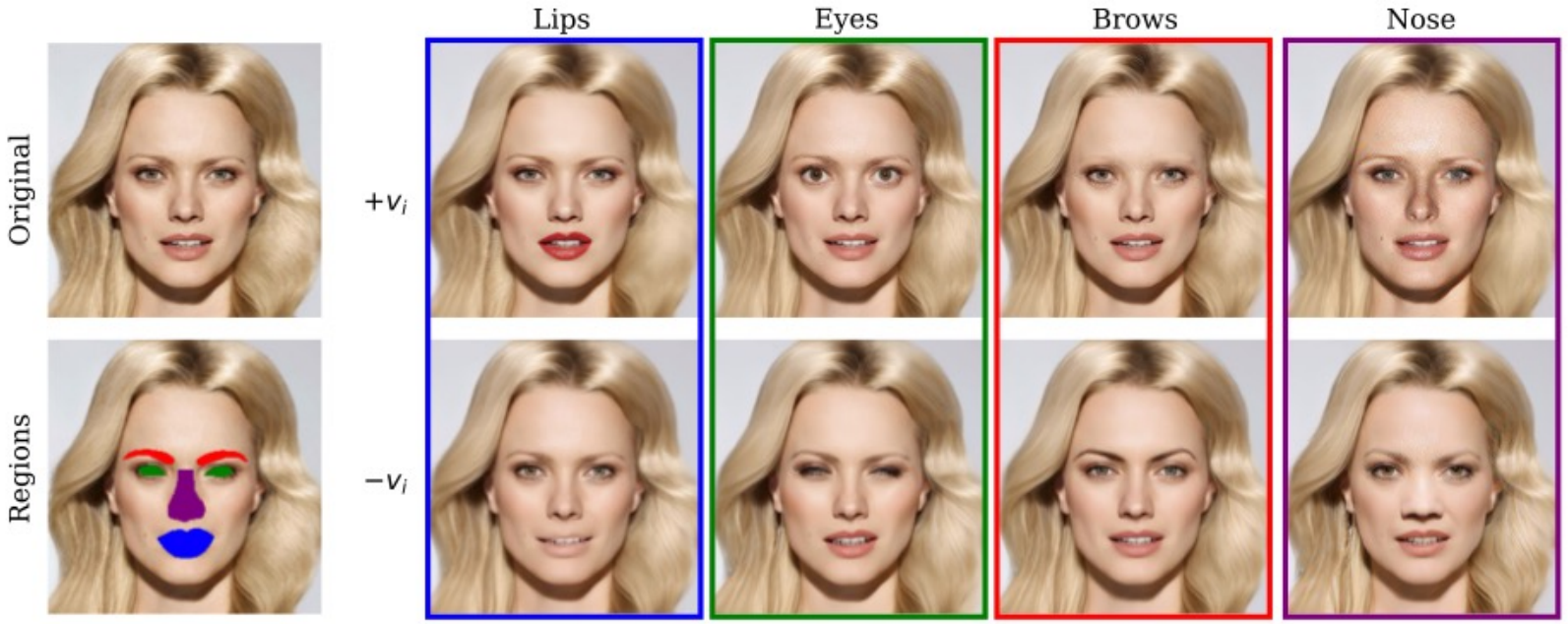
- Оценить «уверенность» в корректной генерации
- Научиться считать интерпретируемую метрику
- Делать это быстро

Существующие методы: Интерполяция мод



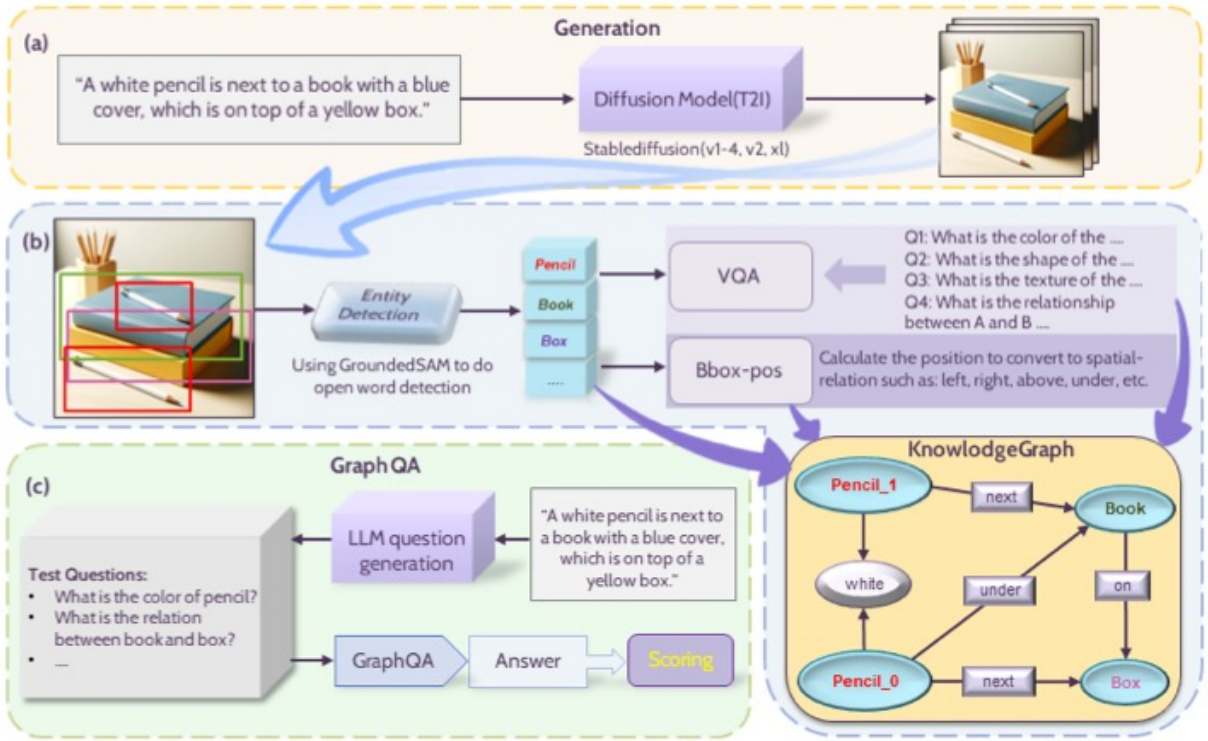
Understanding Hallucinations in Diffusion Models through Mode Interpolation, https://proceedings.neurips.cc/paper_files/paper/2024/file/f29369d192b13184b65c6d2515474d78-Paper-Conference.pdf

Существующие методы: Локальная диффузия



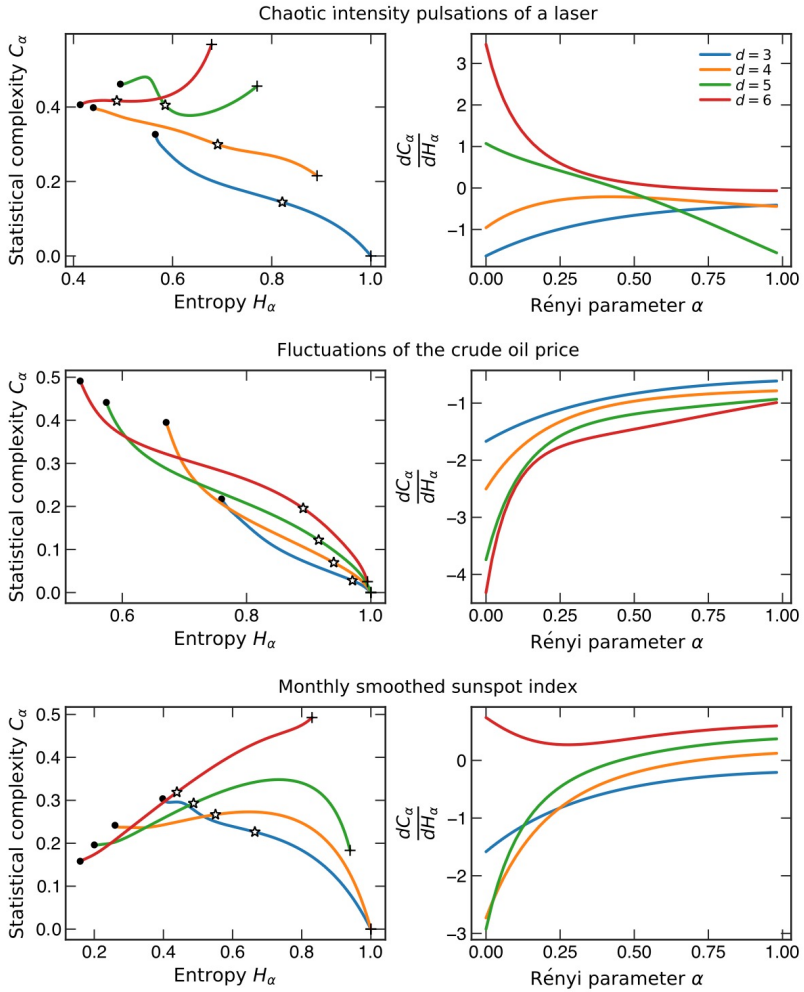
Tackling Structural Hallucination in Image Translation with Local Diffusion (ECCV'24 Oral),
https://arxiv.org/abs/2308.11223

Существующие методы: QA–Агент



Evaluating Hallucination in Text-to-Image Diffusion Models with Scene-Graph based Question-Answering Agent, https://arxiv.org/pdf/2412.05722

Существующие методы:
Кривые Комплексии–Энтропии



Characterization of Time Series via Renyi Complexity-Entropy Curves,
https://ideas.repec.org/a/eee/phsmap/v498y2018icp74-85.html

Наш выбор: Complexity-Entropy (CE)

- Permutation Entropy h — мера беспорядка
- Jensen-Shannon Complexity C — структурная сложность
- Точка (h, C) –> отображается на CE-плоскости
- Быстро, линейно по длине сигнала
- Rényi CE –> кривые вместо точек

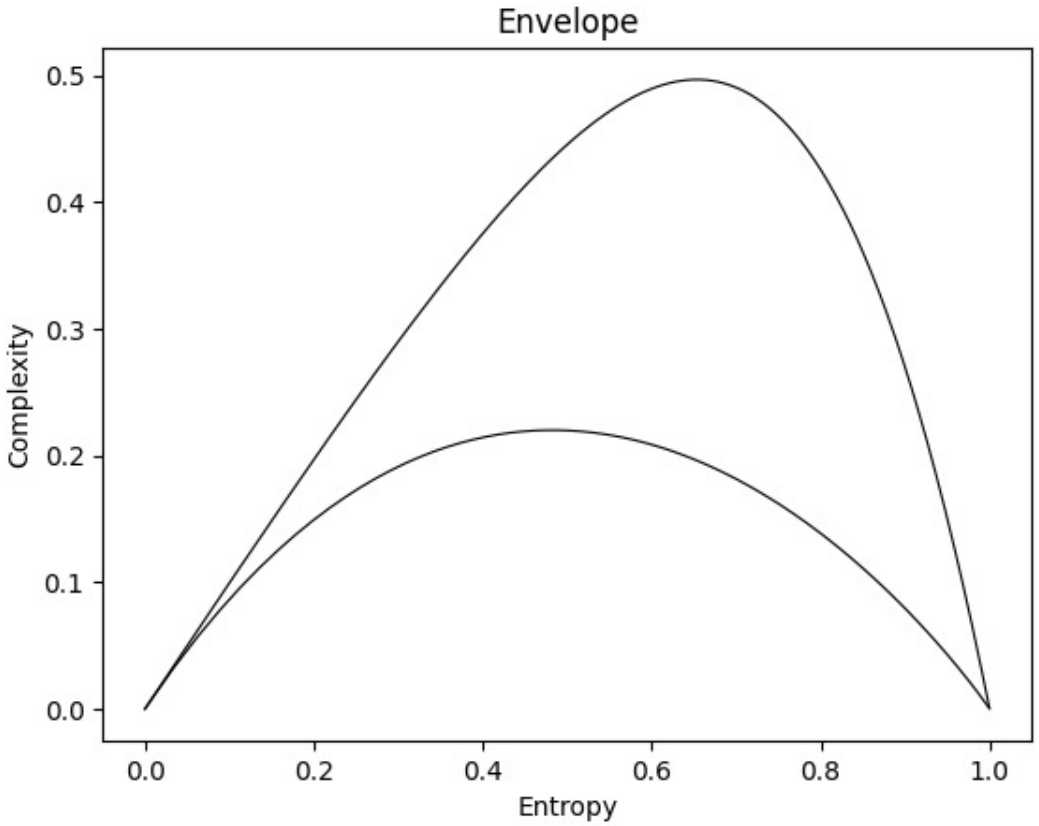
$$H_\alpha(P) = \frac{1}{1-\alpha}\log\Big[\sum_{i=1}^{d!} p_i^{\alpha}\Big],$$

$$C_\alpha(P) = \Big[\frac{D_{\mathrm{JS},\alpha}(P, P_u)}{D_{\mathrm{JS},\alpha}^{\max}}\Big]\ \frac{H_\alpha(P)}{H_\alpha^{\max}},$$

Как считаем энтропию и комплекцию (на формулах):
1. Сигнал –> упорядоченные отрезки длины d, шаг τ
2. Частоты перестановок –> h
3. Сравниваем с равномерным распределением –> C

Кривая обусловленности

- Аналитические верхние и нижние границы
- Можно посчитать в явном виде для всех *d*
- Уже имплементировано в *ordpy* :)



Causality and the Entropy-Complexity Plane:
Robustness and Missing Ordinal Patterns
https://arxiv.org/pdf/1105.4550

План:

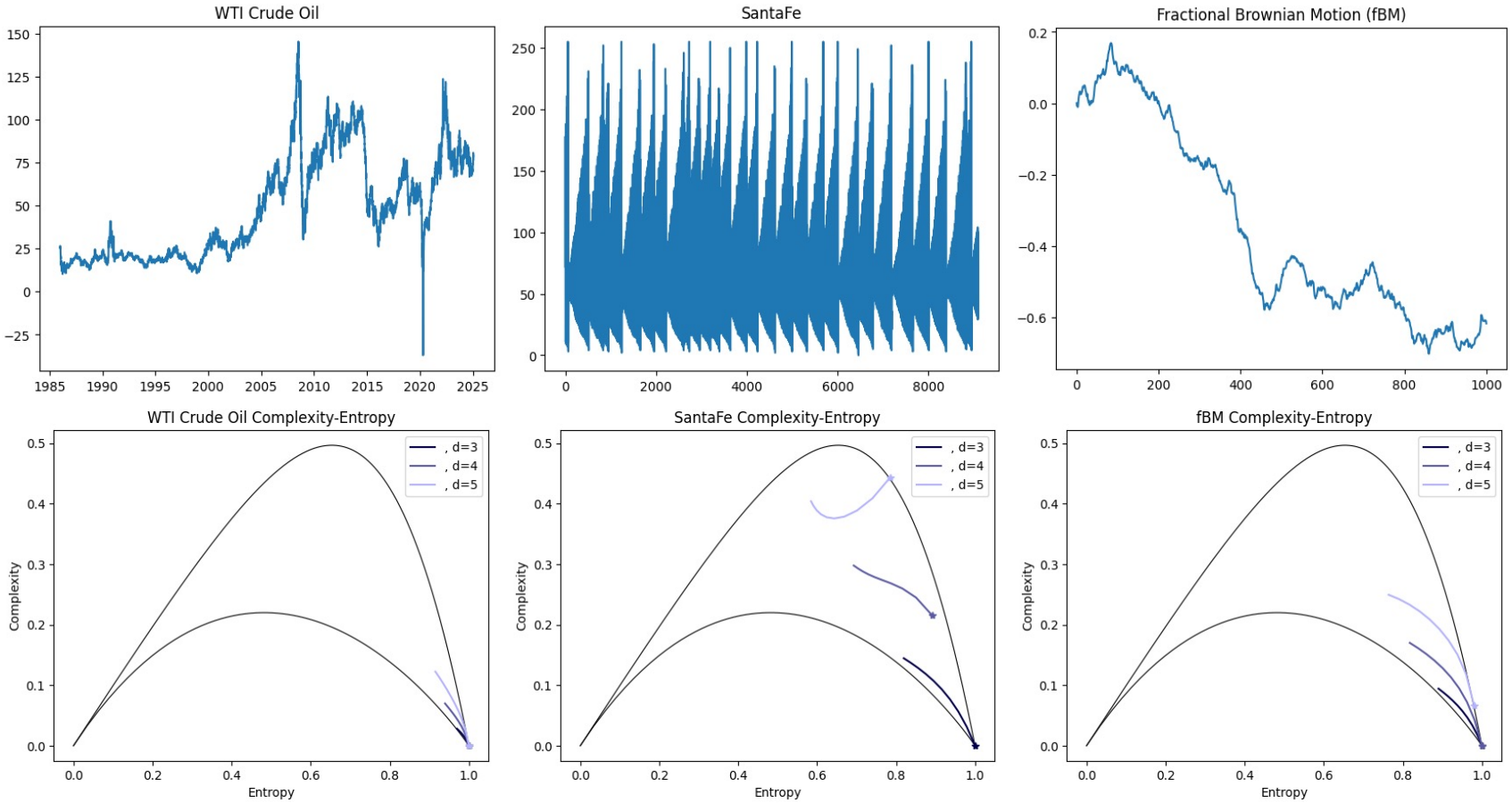
Этап I:

- Собрать разные примеры рядов
- Посчитать характеристику
- Обучить кластеризацию

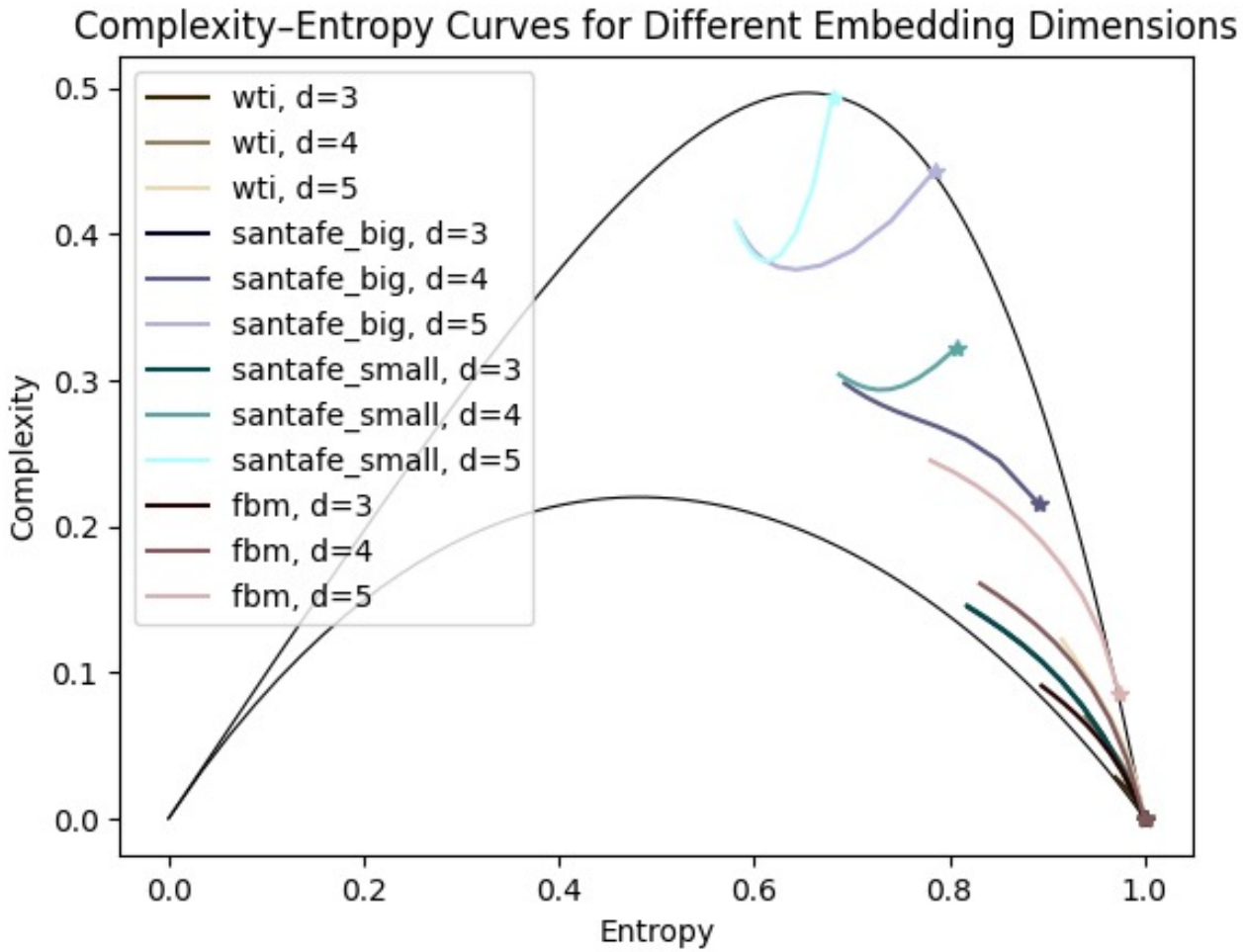
Этап II:

- Найти изображения
- Придумать метрику
- Сделать свертку
- Классифицировать изображения

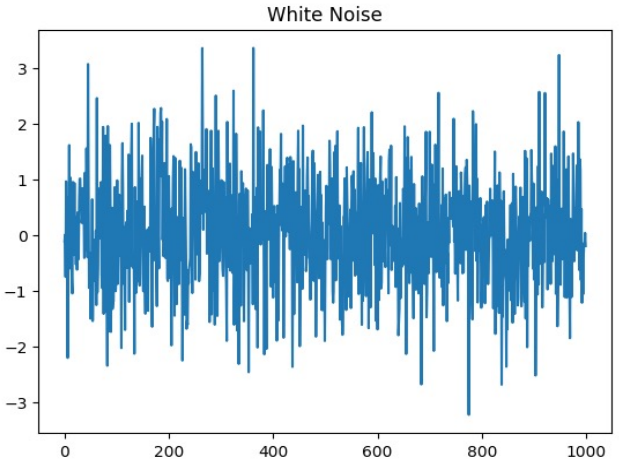
I. Одномерные данные

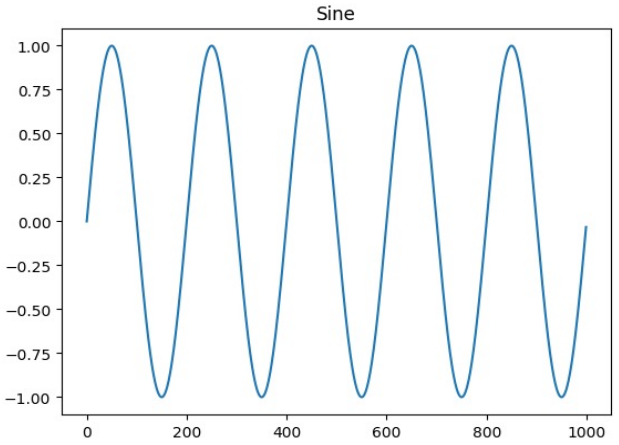


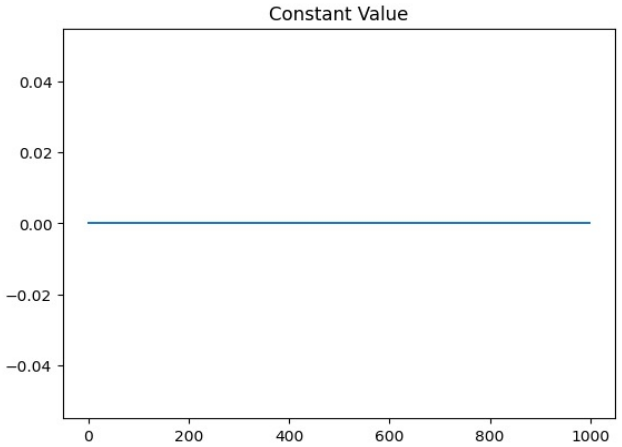
I. Комплексия–Энтропия для одномерных данных

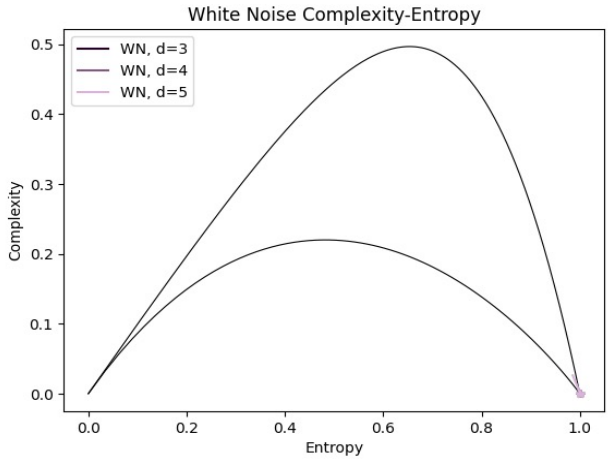


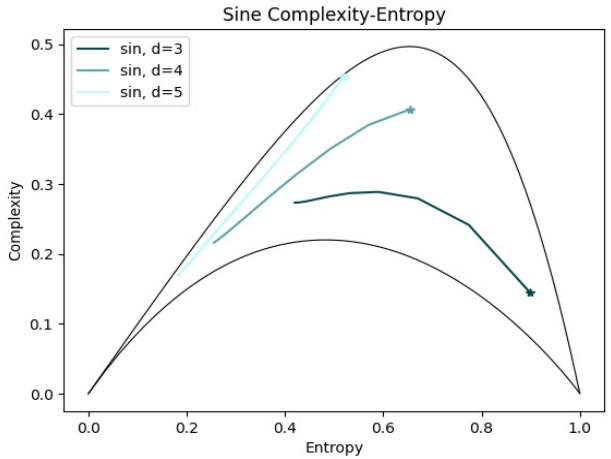
I. Ещё данные

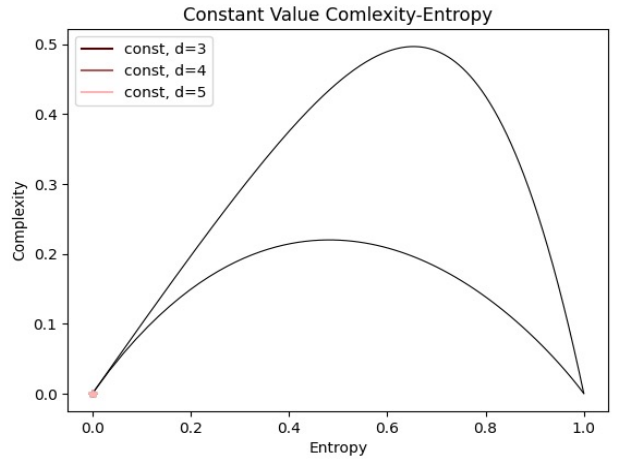

White Noise
3
2
1
0
−1
−2
−3
0
200
400
600
800
1000


Sine
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
200
400
600
800
1000

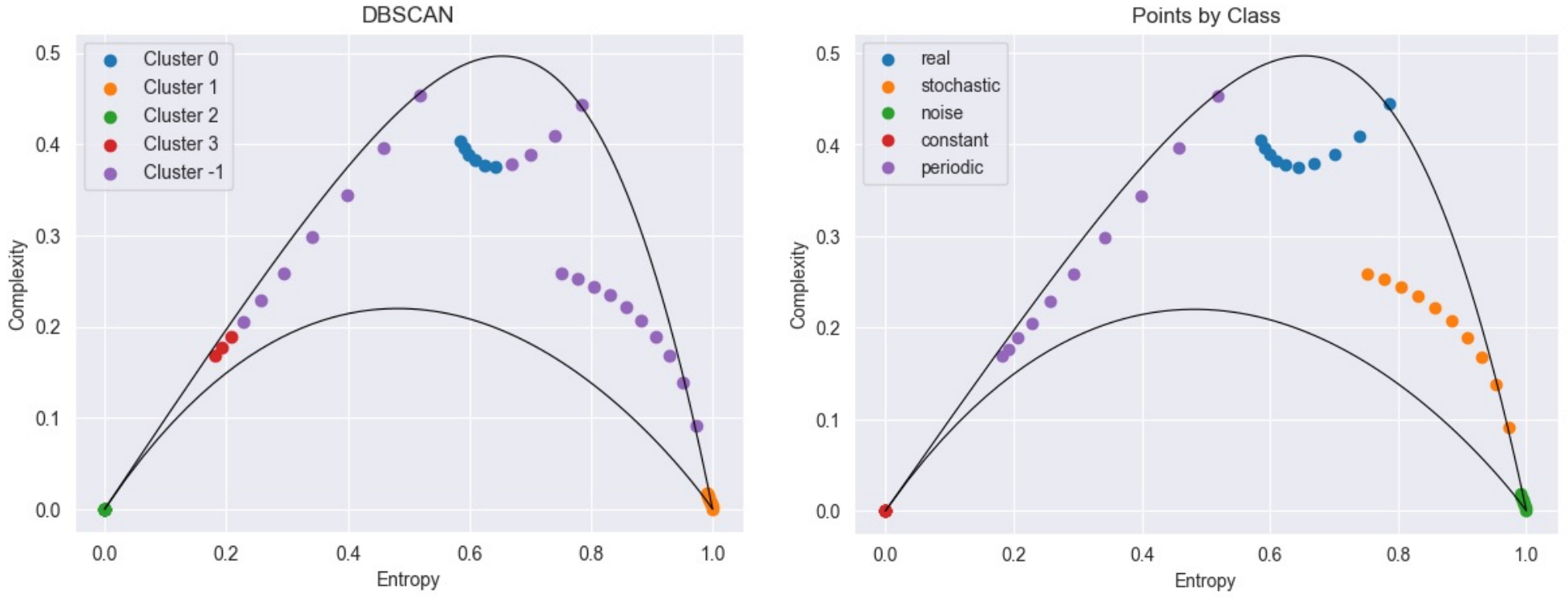

Constant Value
0.04
0.02
0.00
−0.02
−0.04
0
200
400
600
800
1000


White Noise Complexity-Entropy
WN, d=3
WN, d=4
WN, d=5
Complexity
Entropy
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0

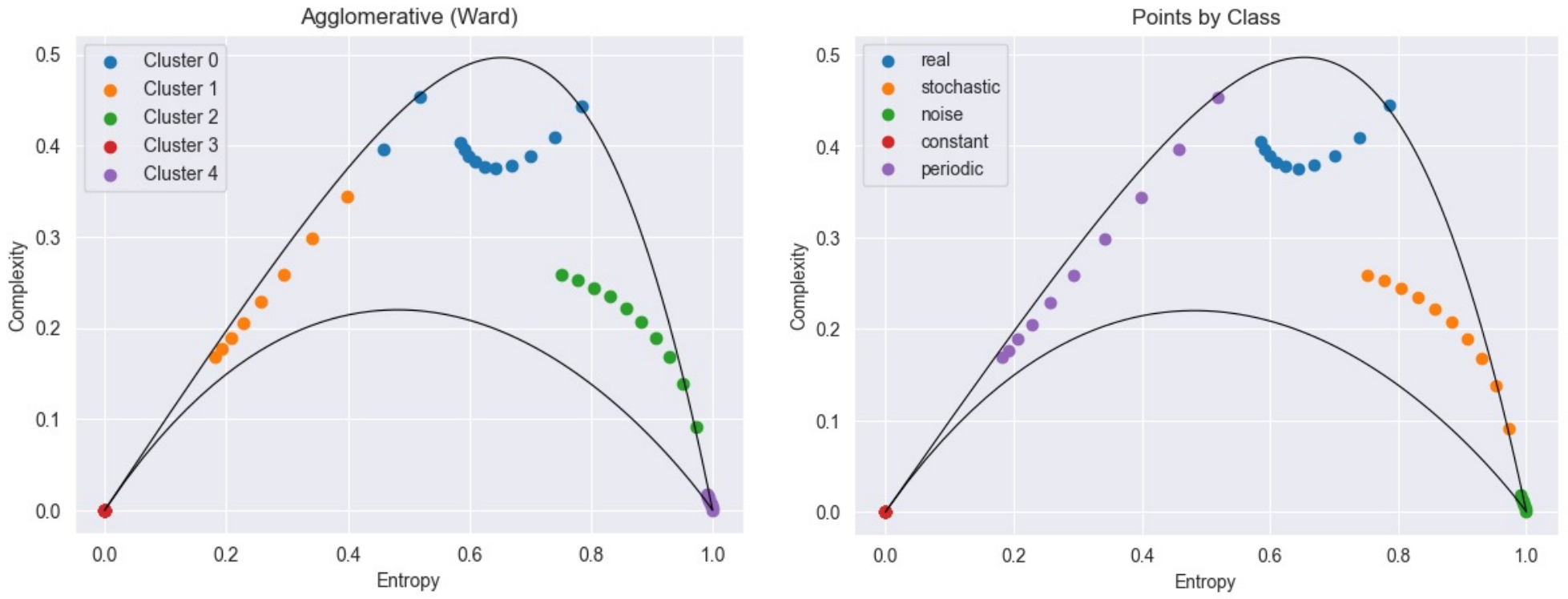

Sine Complexity-Entropy
sin, d=3
sin, d=4
sin, d=5
Complexity
Entropy
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0


Constant Value Comlexity-Entropy
const, d=3
const, d=4
const, d=5
Complexity
Entropy
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0

I. Кластеризация: DBSCAN



I. Кластеризация: Agglomerative (Ward)



II. Изображения

https://cocodataset.org/

https://stablediffusion.fr/gallery/seascape

https://huggingface.co/datasets/nlphuji/whoops

II. Метрика

Ищем разницу между изображениями с галлюцинациями и без

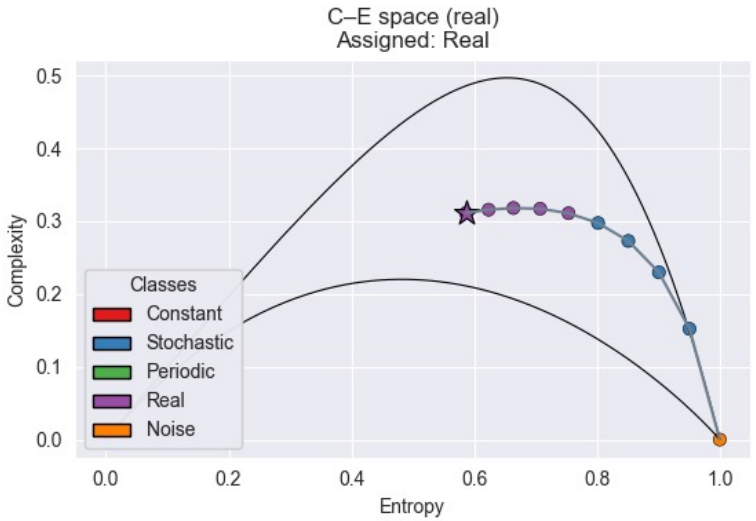
$$\text{score}(\Theta) = \underbrace{\mathbb{E}\big[p_{\text{hall}} \mid \text{bad}, \Theta\big]}_{\text{want high}} - \underbrace{\mathbb{E}\big[p_{\text{hall}} \mid \text{real}, \text{good}, \Theta\big]}_{\text{want low}},$$

II. Свертки

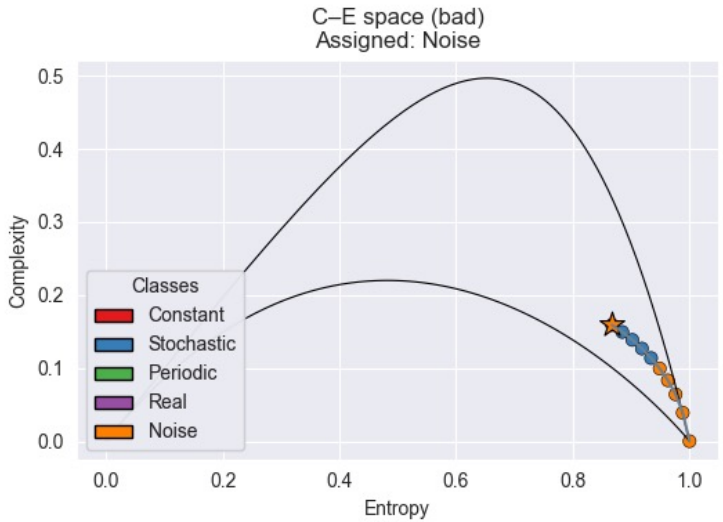
kernel_size	stride	padding	out_channels	Score
5	1	0	2	0.0046
5	2	0	1	–0.0011
5	2	1	4	–0.0013
5	1	1	2	–0.0022
3	2	1	2	–0.0057

II. Результаты: отличия есть

Real image

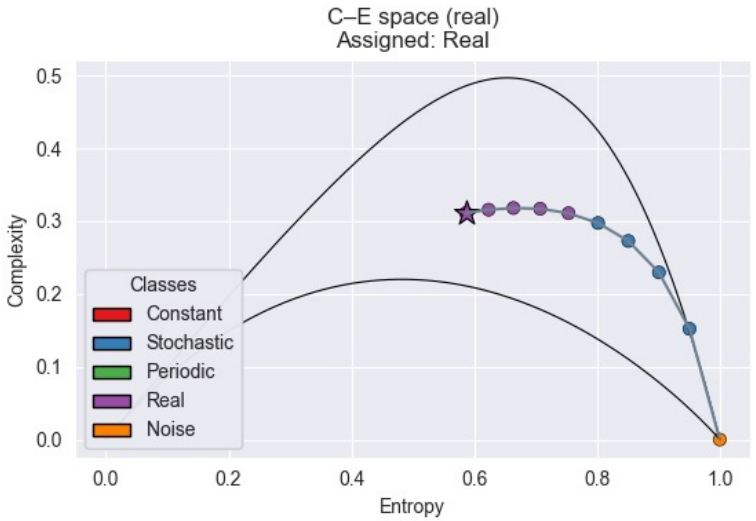


Bad image

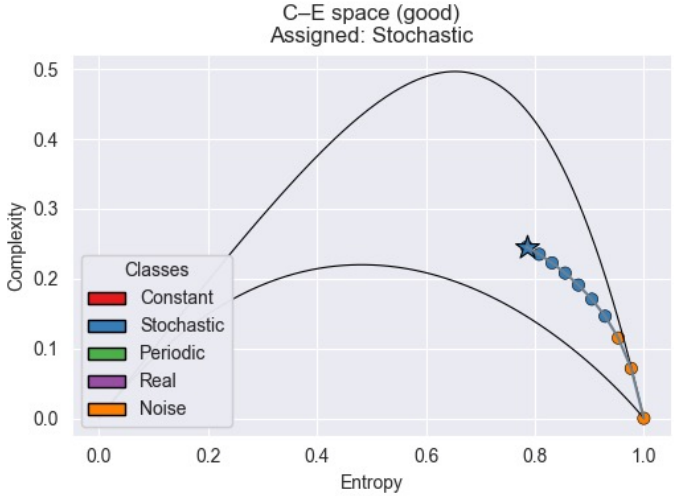


II. Результаты: бонус

Real image



Good image



Будущие направления

- Посчитать метрики бинарной классификации
- Улучшить свертки
- Изучить влияние параметра d
- Применить на видео, тексте

Заключение

- Борьба с галлюцинациями – актуальная задача
- Методы для рядов работают с изображениями
- Необязательно использовать огромные классификаторы

Лаборатория

https://cs.hse.ru/iai/aimf/

Семинары

https://cs.hse.ru/iai/aimf/polls/1031825101.html

Докладчик

https://t.me/teddybeerpkh


HSE UNIVERSITY
EST 1992

7.5 Clustering Methods Compared

- **KMeans** ($k = 5$, $n_{init} = 10$);
- **DBSCAN** ($\varepsilon = 0.02$, $min\_samples = 2$);
- **Agglomerative** (Ward linkage, $k = 5$).

Method	Silhouette score	Comment
KMeans	**0.714**	highest within-cluster cohesion
Agglomerative (Ward)	0.707	comparable to KMeans
DBSCAN	0.367	density-based, no k required

Table 1: Mean silhouette scores in the 2-D Rényi CE space.

Object and subject matter of the research

Object of the Research:

- Diffusion Models and Their Outputs: The research examines diffusion probabilistic models, specifically focusing on the outputs they generate.

Subject Matter of the Research:

- Hallucination Phenomena: It investigates the occurrence of hallucinations —i.e., inaccuracies or deviations in the generated outputs.
- Evaluation Methods: It develops and evaluates detection and classification methods to assess these hallucinations, particularly in novel data generation scenarios.

Primary Studies and Their Methods

Understanding Hallucinations in Diffusion Models through Mode Interpolation
Method: Navigates the latent space by interpolating between modes. This helps pinpoint regions where generated outputs deviate—i.e., hallucinate—from expected patterns.
Key Idea: The smooth transitions between latent modes expose inconsistencies and non-realistic outputs.

Evaluating Hallucination in Text-to-Image Diffusion Models with Scene-Graph based Question-Answering Agent
Method: Employs a scene-graph based QA agent to interrogate generated images. By asking structural and semantic questions, the method assesses whether the content accurately reflects the input text, flagging hallucinated details.
Key Idea: Structured querying helps quantify the degree of coherence between the textual prompt and the visual output.

Tackling Structural Hallucination in Image Translation with Local Diffusion (ECCV'24 Oral)
Method: Focuses on ensuring local consistency in image translation. Local diffusion processes are analyzed to guarantee that spatial features and structures adhere to realistic transformations, thereby reducing structural hallucinations.
Key Idea: Concentrating on localized features maintains structural integrity even when the global output might be complex.

Characterizing Time Series via Complexity-Entropy Curves (and Related Works such as Characterization of Time Series via Renyi Complexity-Entropy Curves & Entropy Test for Complexity in Chaotic Time Series)
Method: These studies apply complexity-entropy measures (including Renyi entropy) to time series data. By plotting these curves, the methods distinguish between randomness (entropy) and structural complexity, providing a framework to detect anomalies or chaotic behavior.
Key Idea: Mapping time series onto a complexity-entropy plane offers insights into the balance between deterministic dynamics and randomness.

Methodological Workflow

1. **Data Extraction:**

   Transform the diffusion model outputs into 1D time series data.

2. **Metric Computation:**

   Calculate entropy and complexity for each time series.

3. **Complexity-Entropy Analysis:**

   Generate curves to visualize the interplay between randomness and structural complexity.

4. **Clustering:**

   Apply clustering techniques to categorize the outputs based on their complexity-entropy characteristics, isolating potential hallucination clusters.

5. **Evaluation:**

   Assess the clusters to determine the quality and authenticity of the diffusion model outputs, thereby detecting and classifying hallucinations.

Entropy and Complexity calculation

Entropy Calculation:

- **Shannon Entropy:**
  Compute $H = -\sum_{(i)} p_i \ln p_i$
  where p_i is the probability of the i-th ordinal pattern.
- **Rényi Entropy:**
  For parameter α ($\alpha \neq 1$), calculate
  $H_a = 1/(1-\alpha) \ln(\sum_{(i)} p_i^{\alpha})$
  as detailed in “Characterization of Timeseries via Renyi Complexity-Entropy Curves.”

Complexity Measure:

- Use a disequilibrium or divergence metric (e.g., Jensen–Shannon divergence) between the observed distribution and the uniform distribution.
- Combine this with the normalized entropy to compute complexity (e.g., LMC complexity), often expressed as:
  $C = Q \cdot H\_norm$
  where Q represents the degree of deviation (or disequilibrium) from uniformity.
- This approach is outlined in “Characterizing Time Series via Complexity-Entropy Curves” and “Entropy test for complexity in chaotic time series.”

Entropy and Complexity calculation

$$H_\alpha(P) = \frac{1}{1-\alpha} \log\Big[\sum_{i=1}^{d!} p_i^\alpha\Big],$$

$$C_\alpha(P) = \Big[\frac{D_{\mathrm{JS},\alpha}(P, P_u)}{D_{\mathrm{JS},\alpha}^{\max}}\Big] \frac{H_\alpha(P)}{H_\alpha^{\max}},$$