Методы векторного представления глубоких генеративных моделей

Мария Александровна Никитина

Московский физико-технический институт

Кафедра: Интеллектуальный анализ данных

Научный руководитель: кандидат ф.-м. наук О. Ю. Бахтеев

Научный консультант: А. Ю. Бишук

2025

Задача векторного описания генеративных моделей

Задача

Задан набор генеративных моделей, описывающий разные выборки/генеральные совокупности данных. Требуется предложить метод векторного представления этих моделей, который будет представлять их статистические свойства.

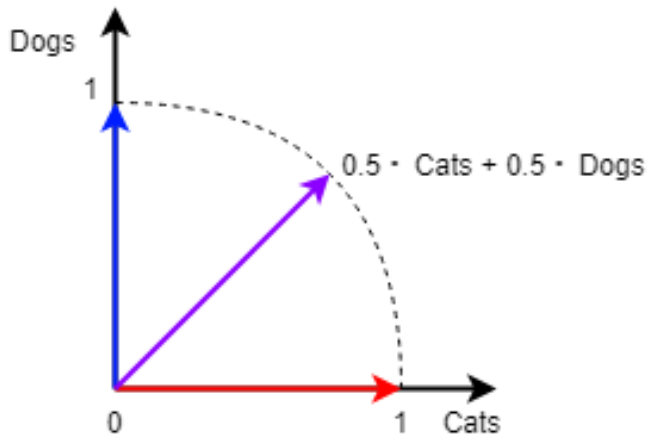
Требования

1. Расстояние между векторными представлениями моделей для близких выборок должно быть невелико (при условии, что сами модели хорошо их описывают)
2. Модели, обученные на композиции/смеси выборок должны учитывать свойства всех выборок, входящих в смесь

Свойства пространства

1. Сумма векторных представлений моделей, полученных по датасетам D_1, D_2 должна приблизительно соответствовать векторному представлению датасета $D_1 + D_2$;

2. Вместо евклидового расстояния на векторных представлениях, используется иерархия;



Бинарная классификация по датасету обучения

Мотивация

Если нужно создать вектор модели с требуемым свойством, то сначала следует проверить, насколько хорошо вектор модели описывает данные, используемые для обучения.

Метод

Обучение бинарного классификатора на определение датасета, из которого была взята выборка для обучения автоэнкодера.

Вход классификатора

Чтобы не допустить переобучения классификатора и не завязываться на размерности, энкодеры векторизуются:

1. Сингулярные числа весов
2. Гистограмма значений весов

Теоретические оценки на потерю информации

Теорема (Никитина, 2025)

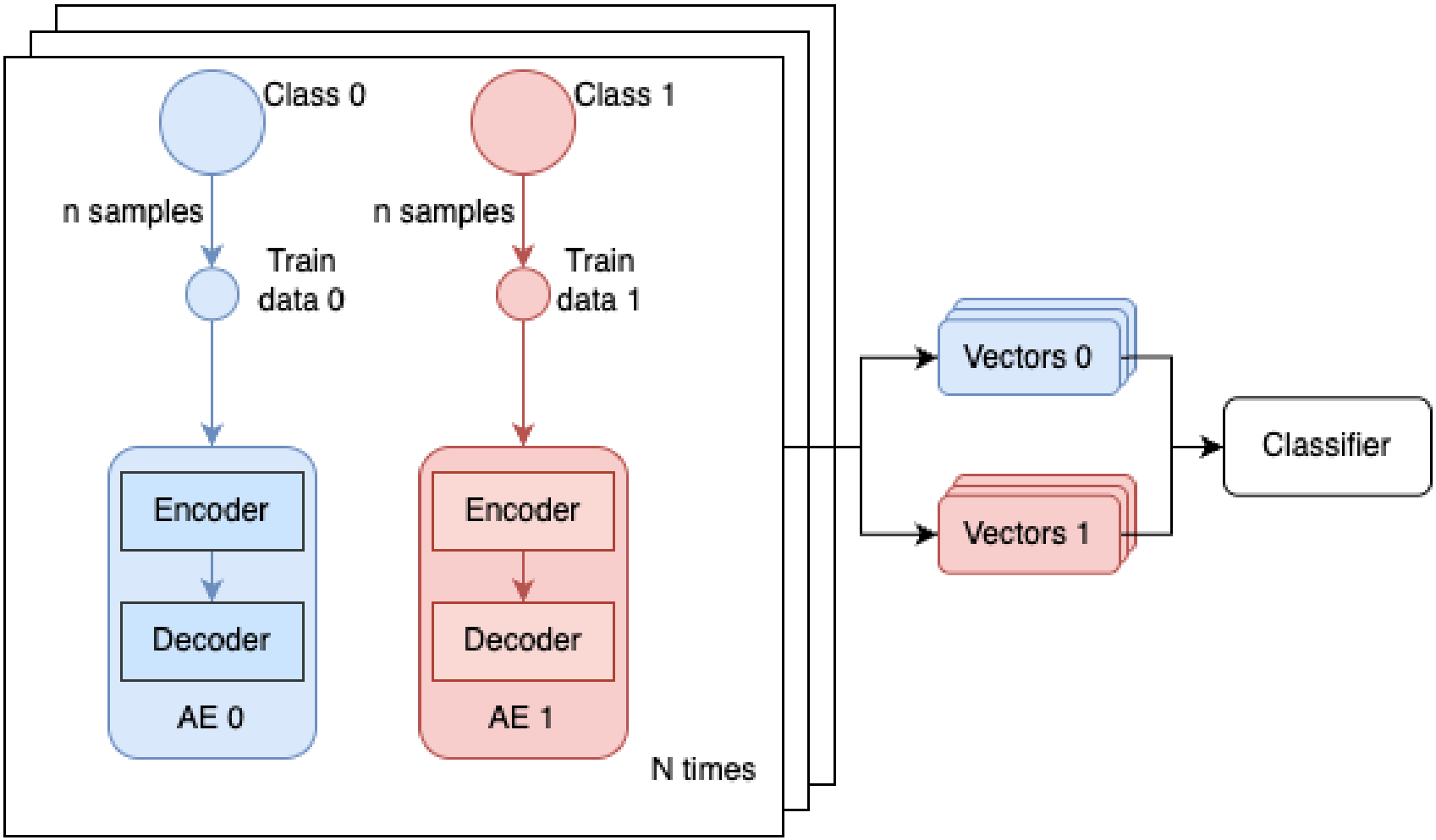
$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathbb{R}^{n \times d}$ – множество независимых векторов (пусть $\mathbf{x}_i \in \mathbb{R}^d$ и $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$).
$X_1 \in \mathbb{R}^{m \times d}$ – его подмножество, формируемое путём независимого включения каждого вектора $\mathbf{x}_i$ с вероятностью p.
AE – линейный автоэнкодер с весами $W \in \mathbb{R}^{k \times d}$. AE обучен на X_1, то есть $W = W(X_1)$. $\mathbf{s}$ – вектор сингулярных чисел AE.
Тогда имеем следующие попарные взаимные информации:

1. $$I(X; X_1) \approx np \cdot H(\mathbf{x}_i), \quad H(\mathbf{x}_i) = \frac{1}{2} \log\left((2\pi e)^d |\Sigma|\right);$$

2. $$I(X_1; W) = \frac{m}{2} \log \frac{|\boldsymbol{\Sigma}|}{|(\mathbf{I} - \mathbf{W}_2\mathbf{W}_1)\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{W}_2\mathbf{W}_1)^T|}$$

3. $$I(X_1; \mathbf{s}) \approx \frac{m}{4} \log\left((2\pi e)^k \prod_{i=1}^{k} \lambda_i^2\right).$$

Схема базовой задачи



От классификатора требуется предсказывать класс, на котором обучен автоэнкодер.

Метрики базовой задачи

Метрики качества предсказания класса, на котором обучалась модель

	Precision	Recall
1 линейный слой	0.79	1.00
1 свёрточный слой	0.55	0.78

Автоэнкодеры с линейным слоем, обученные на разных датасетах, отличить друг от друга легче, чем автоэнкодеры со свёрточными слоями. То есть, чем сложнее модель, тем больше информации теряется о ней и датасете при обучении и векторизации.

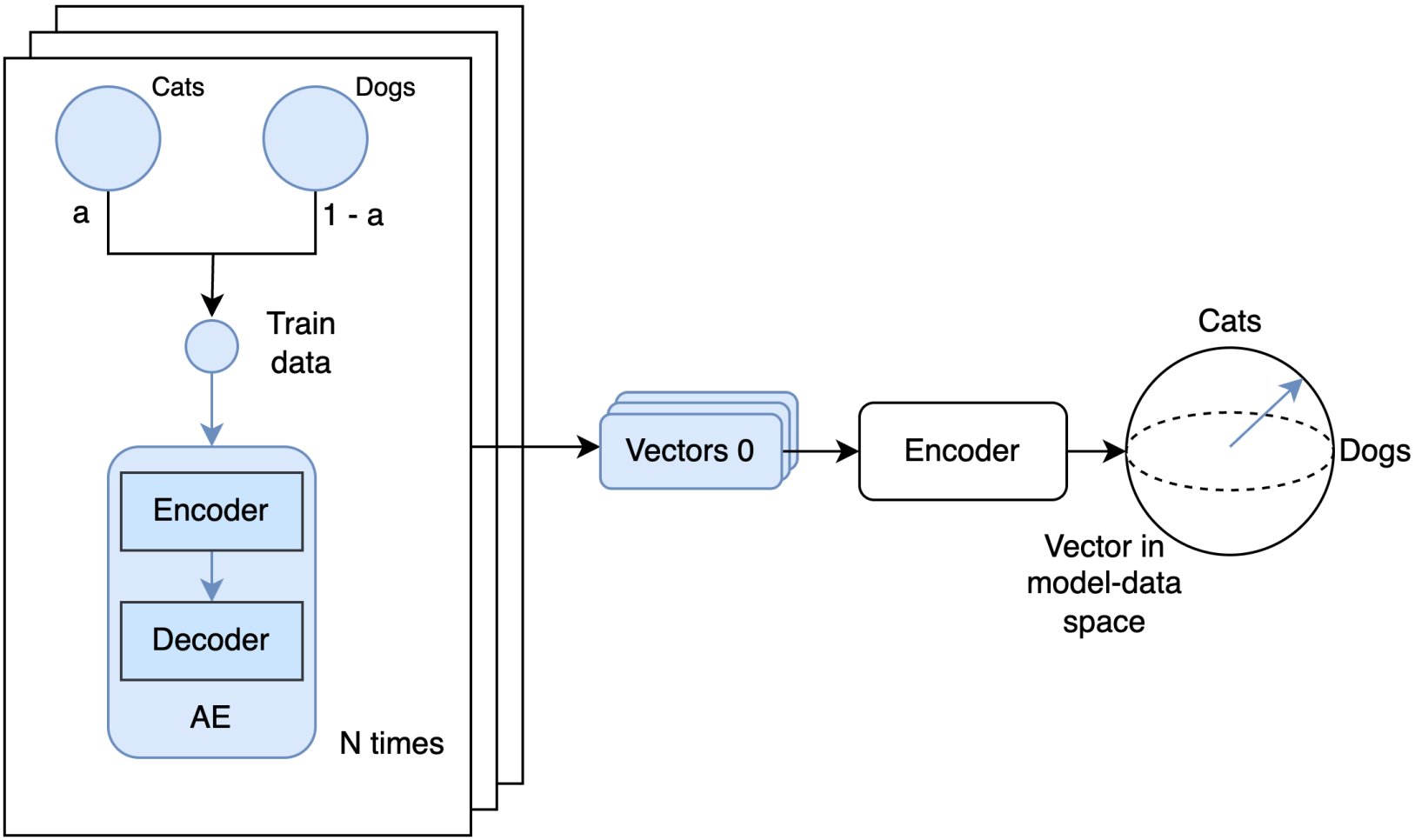
Построение векторного пространства

Берётся 3 наиболее удалённых класса из датасета CIFAR. Для поиска таких классов используется евклидово расстояние на ембеддингах, полученных из выходного слоя ResNet.

Алгоритм

1. Случайное сэмплирование долей классов в датасете для N моделей;
2. Обучение N моделей на соответствующих датасетах;
3. Получение векторов из обученных моделей;
4. Обучение энкодера на полученных моделях. Предсказание:
   - 4.1 Вектора на части единичной сферы
   - 4.2 Расстояния между двумя моделями

Схема эксперимента



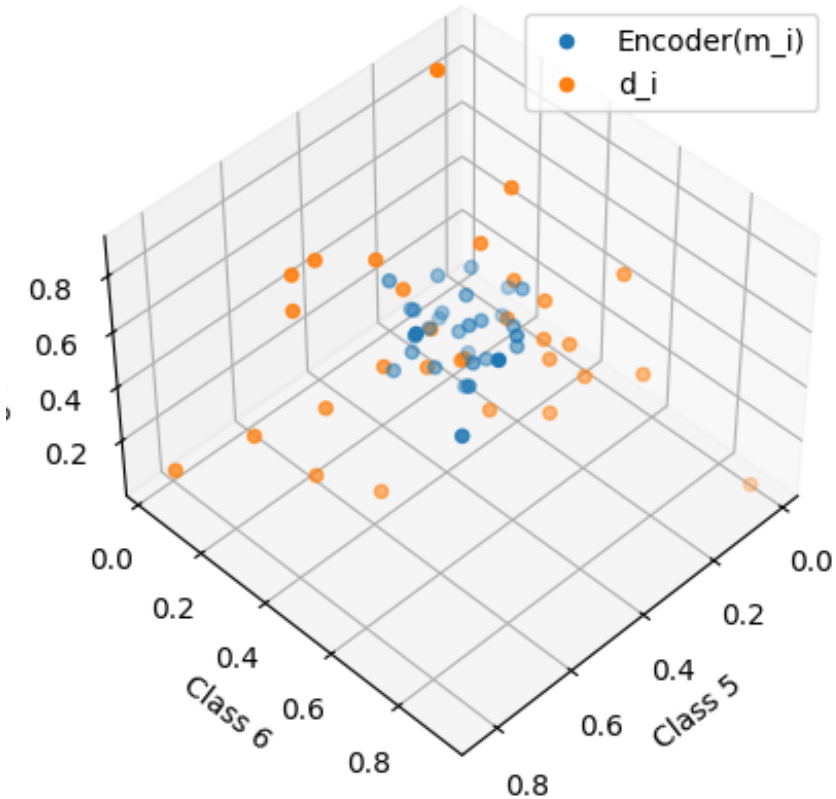
Задача энкодера: приблизить вектор обученной модели к вектору датасета

Среднее векторов скрытого пространства

Способ представления модели: среднее значение скрытого пространства автоэнкодера на тестовой выборке.

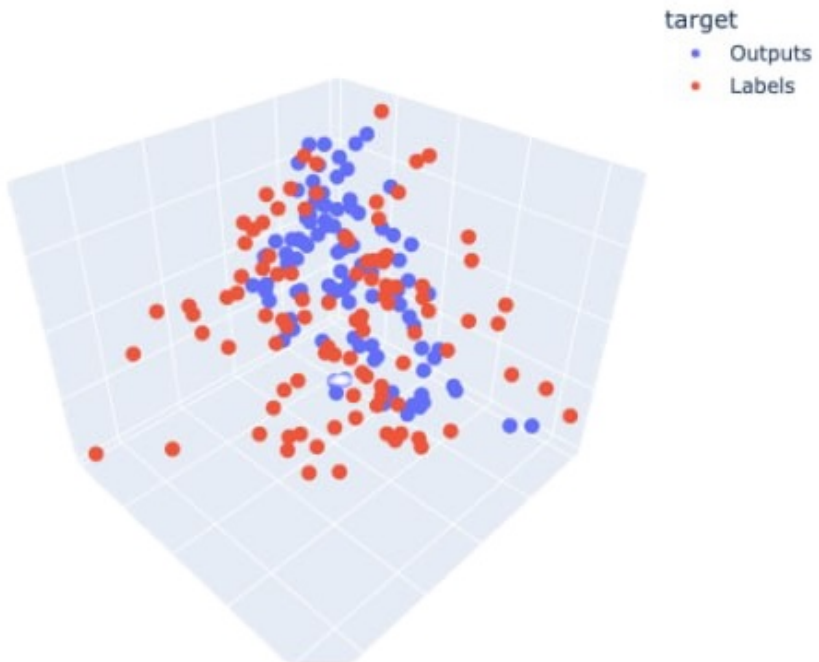
Функция потерь энкодера: MSE

Проблемы: Переобучение энкодера и неинформативность взятия среднего



Сингулярные числа весов модели

Способ представления модели: Для матриц весов автоэнкодера считаются сингулярные числа. Затем все значения вытягиваются в один вектор.



Функция потерь энкодера: MSE

Проблемы: При увеличении числа слоёв увеличивается число сингулярных чисел, а, значит, метод не масштабируется на различные архитектуры.

Выводы и дальнейшие шаги

1. Для бинарной классификации получена большая разница в качестве между экспериментами на линейных и свёрточных автоэнкодерах. Получить теоретическое обоснование. Продолжить предложенные эксперименты: построить бинарную классификацию на моделях с несколькими линейными и свёрточными слоями.
2. Выведены теоретические оценки взаимной информации между вектором модели и исходным датасетом. Получить экспериментальные оценки полученных результатов.
3. Получить теоретические оценки взаимной информации для случая кодирования моделей с помощью гистограмм.