

Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра анализа данных

Направление подготовки / специальность: 01.03.02 Прикладная математика и информатика

Направленность (профиль) подготовки: Прикладная математика и компьютерные науки

ПРИЧИННЫЙ ВЫВОД В ВАРИАЦИОННЫХ АВТОЭНКОДЕРАХ

(бакалаврская работа)

Студент:
Троешестова Лидия Сергеевна

(подпись студента)

Научный руководитель:
Волков Никита Алексеевич,
канд. физ.-мат. наук

(подпись научного руководителя)

Консультант (при наличии):

(подпись консультанта)

Москва 2024

Аннотация

Причинный вывод в вариационных автоэнкодерах

Троещестова Лидия Сергеевна

В данной работе рассматривается задача генерации изображений по нескольким меткам с учетом лежащих в основе системы причинно-следственных связей. Подробно рассматриваются проблемы, которые могут быть решены с помощью причинного вывода в области компьютерного зрения, такие как проблема ложных корреляций и доменная генерализация. Разработан новый причинно-следственный пайплайн, который последовательно выполняет численное причинное обнаружение и позволяет производить причинный вывод для генерации изображений без мешающих влияний конфаундеров. Рассмотрены базовые модели CVAE и CVAE/GAN, к латентному пространству которых применен предложенный пайплайн. Эксперименты на датасетах Colored MNIST и Pendulum показывают, что данный способ улучшает качество генерации изображений как по обучающим, так и по контрафактивным наборам меток.

СОДЕРЖАНИЕ

1 Введение	4
1.1 Причинный вывод в современном машинном обучении	4
1.2 Основы причинного вывода	5
1.3 Проблема ложных корреляций	7
1.4 Ложные корреляции в генеративных моделях	9
2 Постановка задачи	11
3 Обзор существующих решений	13
3.1 Причинная генеративно-состязательная сеть (CausalGAN) [1]	13
3.2 Причинный вариационный автоэнкодер (CausalVAE) [2]	13
3.3 Причинная диффузионная модель (Diff-SCM) [3]	14
3.4 Модели для причинного обнаружения	16
4 Исследование и построение решения задачи	17
4.1 Устройство CVAE	17
4.2 Каузальное обучение на основе разбиения данных	18
4.3 Пайплайн причинного обнаружения и вывода	18
4.4 Применение CVAE-GAN	21
5 Описание практической части	23
5.1 Датасет Colored MNIST	23
5.1.1 Обучение базовых моделей	23
5.1.2 Обучение моделей причинного обнаружения	24
5.1.3 Результаты на датасете Colored MNIST	27
5.2 Датасет Pendulum	29
6 Заключение	32

1 ВВЕДЕНИЕ

1.1 ПРИЧИННЫЙ ВЫВОД В СОВРЕМЕННОМ МАШИННОМ ОБУЧЕНИИ

Появление искусственного интеллекта привело к трансформационным изменениям в различных областях, продемонстрировав его огромный потенциал в реальных приложениях. Среди различных аспектов глубинное обучение, разновидность искусственного интеллекта, добилось значительных успехов, особенно в сфере компьютерного зрения. Этот прогресс очевиден в его ключевой роли в совершенствовании таких технологий, как автономные транспортные средства, дроны и робототехника. Эти достижения поддерживаются инновационными стратегиями обучения, включая механизмы внимания и методы предобучения.

Большинство существующих методов в глубинном обучении основаны на статистических методах, которые обладают ограничениями в частых явлениях реального мира, таких как причинно-следственные связи между признаками и существенные сдвиги между распределениями. Парадокс Симпсона [4] демонстрирует один из таких недостатков. Например, анализ эффективности лекарства по некоторой группе людей может показать, что в целом оно увеличивает шанс выздоровления, но при этом если разделить группу на две (мужчин и женщин), то оказывается, что это лекарство уменьшает шанс выздоровления в каждой группе. Этот парадокс подчеркивает зависимость статистических результатов от уровней интерпретации при одном и том же наборе данных. Следовательно, статистические корреляции как на индивидуальном, так и на популяционном уровне не могут полностью отразить тонкую взаимосвязь между потреблением данного лекарства и выздоровлением. Поэтому применяются методы причинного вывода, включающие предварительные знания о причинных структурах. Эти методы позволяют выявить точные причинно-следственные связи на определенных уровнях интерпретации. В результате подходы, основанные на каузальности, предлагают более логические и качественные результаты по сравнению с их аналогами, основанными на статистической корреляции.

В задачах классификации компьютерного зрения на основе глубинного обучения классическая цель заключается в эффективной обработке изображений, обозначенных как X . Основная цель — обучить нейронную сеть, способную точно предсказывать соответствующую метку Y . Для достижения этой цели используется статистическая модель, адаптированная к подходящей целевой функции, которая, в свою очередь, помогает опе-

нить условное распределение вероятностей $P(Y|X)$ [5]. Однако эта оценка справедлива только при условии независимого одинакового распределения данных, то есть, чтобы изученное условное распределение вероятностей $P(Y|X)$ оставалось применимым не только в пределах набора обучающих данных, но и плавно распространялось на набор тестовых данных. Это условие основано на ожидании того, что новые выборки прогнозов будут соответствовать характеристикам распределения исходного обучающего набора.

Такие методы, как адаптация предметной области и генерализация, появились для устранения сдвигов в распределениях. Хотя современные методы отдают приоритет аппроксимации функций, более аналитический подход может оптимизировать обобщаемость и интерпретируемость. Благодаря своим сильным сторонам, причинность в последнее время привлекла к себе значительное внимание, найдя применение в различных областях, включая статистику, экономику, эпидемиологию и информатику. По своей сути причинные методологии можно разделить на два основных аспекта: *причинное обнаружение* и *причинный вывод*. При наличии оснований, установленных причинно-следственным обнаружением, причинно-следственный вывод использует эти отношения для более глубокого анализа.

1.2 Основы причинного вывода

Причинный вывод — это процесс определения независимого воздействия конкретного явления, которое является частью более крупной сущности. Главное различие между причинно-следственным выводом и статистическим выводом — в том, что причинно-следственный вывод исследует реакцию переменной результата при изменении переменной причины. В этой части обзора введем некоторые концепции из причинности. Будем использовать нотацию Перла [6], которая описывает структурные причинные модели с помощью уравнений и ориентированных ациклических графов для описания причинных связей между случайными величинами.

При предположении причинной достаточности случайная величина X *влияет* на случайную величину $Y \iff$ существует функция f и некоторая ненаблюдаемая случайная величина $E : E \perp\!\!\!\perp X; Y = f(X, E)$. Такие ненаблюдаемые переменные также называют *экзогенными*. Этому отношению соответствует граф $X \rightarrow Y$. В общем случае причинный граф — ориентированный ациклический граф, подразумеваемый структурными уравнениями: родители вершины в причинном графе представляют причины этой

переменной. Причинно-следственный граф можно построить на основе структурных уравнений следующим образом: родителями вершины являются те вершины, которые присутствуют в уравнении, задающем значение этой переменной.

Формально структурная каузальная модель – кортеж $\mathcal{M} = (\mathcal{V}, \mathcal{E}, \mathcal{F}, \mathcal{P}_E)$, содержащий набор функций $\mathcal{F} = \{f_1, \dots, f_n\}$, набор случайных величин $\mathcal{V} = \{X_1, \dots, X_n\}$, набор экзогенных случайных величин $\mathcal{E} = \{E_1, \dots, E_n\}$ и вероятностное распределение над экзогенными переменными \mathcal{P}_E . Набор наблюдаемых переменных \mathcal{V} имеет совместное распределение, определяемое распределениями \mathcal{E} и функциональными зависимостями \mathcal{F} . Причинный граф D – ориентированный ациклический граф на вершинах \mathcal{V} , такой, что вершина X_j – родитель вершины X_i тогда и только тогда, когда $X_i = f_i(X_j, S, E_i)$ для некоторого $S \subset \mathcal{V}$. Набор родителей переменной X_i обозначается Pa_i . Тогда D можно считать байесовской сетью для совместной вероятности над наблюдаемыми переменными \mathcal{V} .

Интервенция – операция, изменяющая исходный каузальный механизм, и, как следствие, соответствующий каузальный граф. Интервенция на переменной X_i обозначается $\text{do}(X_i = x_i)$. Она отличается от обусловленности на $X_i = x$ следующим: интервенция удаляет связи вершины X_i с ее родителями, в то время как обусловленность не изменяет каузальный граф, из которого семплируются данные. Совместное распределение исходного графа может быть факторизовано как $P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{Pa}_i)$, где вершинам в Pa_i присваиваются соответствующие значения из $\{x_i\}_{i=1}^n$. После интервенции на наборе переменных $X_S := \{X_i\}_{i \in S}$, то есть операции $\text{do}(X_s = \mathbf{s})$, пост-интервенционное распределение выражается как $\prod_{i \notin S} P(x_i | \text{Pa}_i^S)$, где Pa_i^S обозначает присваивание X_j соответствующей компоненты \mathbf{s} в случае интервенции ($j \in S$). Пример интервенции представлен на рисунке 1.

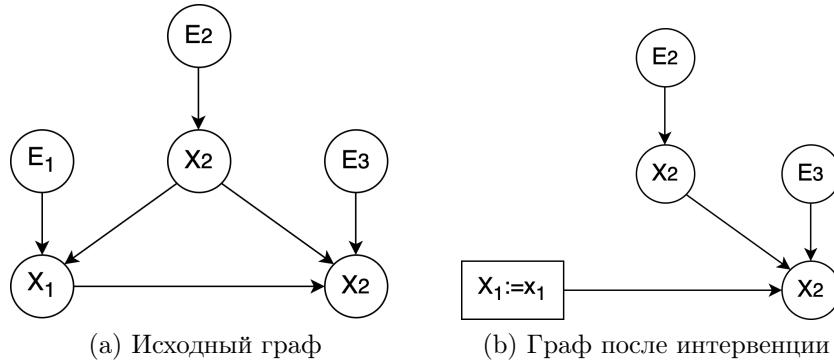


Рис. 1: Интервенция на X_1

Часто в структурных каузальных моделях встречается структура, представленная

на рисунке 2. В этом случае переменная C называется *конфаундером* переменных X, Y . Эта структура повсеместна и в задачах распознавания в компьютерном зрении, если X – изображение, Y – метка класса, а C – некоторый контекст, влекущий как визуальные признаки X , так и класс Y . Примеры будут приведены в секции 1.4. Такие контексты могут привести к ложным корреляциям в истинной причинно-следственной связи между изображениями и метками. Чтобы противодействовать этому эффекту, производится интервенция на C .

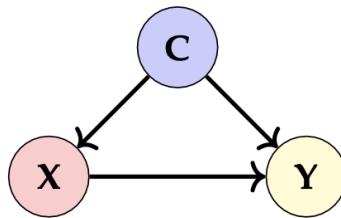


Рис. 2: Типичная каузальная структура

Наконец, обсудим понятие контрафактивности. Контрафактивное рассуждение соопределяет наблюдаемые данные с гипотетическими результатами в различных условиях. Например, изменив атрибут изображения цвета автомобиля с красного на синий, и проанализировав изменения качества классификации, можно оценить важность этого атрибута. В рамках данной работы *контрафактивными* комбинациями меток называются те, которые не доступны модели в ходе обучения.

1.3 ПРОБЛЕМА ЛОЖНЫХ КОРРЕЛЯЦИЙ

Фраза «корреляция не является причинностью» объясняет, что то, что две вещи кажутся связанными друг с другом, не означает, что одна является причиной другой. При рассмотрении большого количества данных, которые следуют одной и той же схеме, статистическое обучение может принести пользу. Визуальное распознавание в текущее время регулируется минимизацией эмпирического риска, что ограничивает ошибку обобщения, когда распределения обучающего и тестирующего датасета совпадают. Когда обучающие наборы охватывают все изменяющиеся факторы, такие как фоновый контекст или точки зрения камеры, дискриминативные модели учатся инвариантам и прогнозируют метки категорий объектов по нужным, правильным причинам. Однако визуальный мир огромен и собрать представительный, сбалансированный набор данных сложно, а в некоторых случаях невозможно, потому что он может непредсказуемо измениться после обучения.

Оптимизация эмпирического риска напрямую склонна к обучению ложных корреляций, которые не учитывают нижележащую (скрытую, лежащую в основе системы) причинную структуру. В фотографиях часто объект интереса и контекст сцены имеют мешающие факторы, создавая ложные корреляции. Например, при распознавании изображений модель может угадать класс «птица», когда видит на изображении «небо», просто потому, что птицы и небо часто появляются в данных вместе. В этом случае контекст, в котором находится объект, является конфаундером, или мешающим фактором для извлечения истинного эффекта. Несколько исследований выявили эту проблему, продемонстрировав существенное ухудшение производительности, когда искажающее смещение исключается из датасета тестирования. Например, в датасете ObjectNet, в тестовой части которого удалено несколько частых ложных корреляций, точность современных моделей ухудшается на 40% по сравнению с валидационной частью ImageNet.

Эта проблема называется *проблемой ложных корреляций*, и существует не только в задаче классификации изображений, но и в генеративных и мультимодальных задачах машинного обучения. Следствие проблемы ложных корреляций в случае классификации – плохое качество на более сложных датасетах, таких как ObjectNet и ImageNetV2, а в общем случае – плохая генерализационная способность модели.

Многообещающим направлением для улучшения визуального распознавания является изучение причинных представлений. Если представления способны идентифицировать причинно-следственный механизм между признаками изображения и метками категорий, становится возможным надежное обобщение. Причинное обучение моделей отличается от классического тем, что оно направлено на поиск причинно-следственных связей, помимо просто корреляций в данных.

Базовым решением проблемы ложных корреляций в задаче классификации является применение аугментаций к обучающему датасету: зашумление, изменение цветовой палитры, повороты, отражения, обрезания изображений. Однако большинство аугментаций не могут выполнять высокоуровневые преобразования изображений, таких как изменение точки зрения или фона. Мао, Чай [7] разработали метод аугментации GenInt на основе генеративных интервенций с использованием генеративно-состязательных сетей (GAN). Предложенный метод устраняет ложные корреляции в большей степени, чем обычные аугментации. Интервенции в изображения проводятся с помощью изменения латентных компонент низкоразмерного представления изображений по основным направлениям, полученным с помощью РСА. Латентное представление в генеративных

моделях призвано умещать в низкоразмерном пространстве всю важную информацию об изображении, и показано, что компоненты латентных векторов зачастую отвечают конкретным интерпретируемым концептам: цвету фона, точке зрения, длину и ширину объекта. Поэтому идея интервенции в изображения посредством изменения латентных кодов в последнее время набирает популярность и в других задачах, таких как генерация контрафактивных изображений.

1.4 Ложные корреляции в генеративных моделях

В настоящее время область машинного обучения рассматривает две основные разновидности моделей. Пусть x – признаковое описание объекта, y – соответствующая метка. Дискриминативные модели оценивают распределение $p_x(y)$, в то время как генеративные модели оценивают распределение $p(x, y)$ или же просто $p(x)$ в случае отсутствия меток.

Проблема ложных корреляций повсеместна в традиционных методах машинного обучения на данных сложной природы, таких как изображения и текст. Она проявляется не только в дискриминативных моделях (задачи классификации изображений, оценки тональности текста), но и в генеративных (задачи генерации текста по изображению, генерации изображения по тексту или меткам).

В мультимодальной задаче генерации подписи к изображению, где часто для извлечения визуальных признаков используют хорошо обученный детектор, проблема ложных корреляций проявляется как в модальности изображений, так и в модальности текста. Например, визуальные признаки объекта «вилка», извлеченные с помощью детектора, имеют тенденцию быть окружающими ее элементами, похожими на торт, поскольку вилки и торт очень часто встречаются в датасете. Как следствие, на представление класса «вилка» сильно влияют визуальные признаки торта, которые в данном случае являются конфаундером истинного влияния $X \rightarrow Y$. Следовательно, очень важно распутать визуальные особенности на этапе визуального представления, чтобы уменьшить ложную корреляцию между областью торта и классом «вилка». Аналогичная проблема в модальности генерации текста вызвана тем, что в датасете готовых подписей к изображениям некоторые сочетания слов встречаются чаще, чем другие. Это приводит к предвзятости генератора текста по полу, действию и объектному контексту.

В задаче генерации изображений проблема ложных корреляций приобретает но-

вый смысл. Рассмотрим истинный каузальный граф для датасета лиц людей: $\text{gender} \rightarrow \text{mustache}$. Когда метки выбираются независимо друг от друга, изображения, созданные с меткой $\text{mustache} = 1$, должны содержать как мужчин, так и женщин, что явно отличается от обусловливания на $\text{mustache} = 1$. Ключом к пониманию и объединению этих двух понятий, обусловливания и возможности семплирования из контрафактивного распределения, является использование каузальной модели. Обусловливаясь на $\text{gender} = \text{male}$, мы ожидаем видеть сгенерированные лица мужчин и с усами, и без, в зависимости от их соотношения в популяции. Когда мы обуславливаемся на $\text{gender} = \text{female}$, в генерации должны присутствовать только лица с $\text{mustache} = 0$ в силу каузальной связи. Кроме того, после интервенции распределения значений в вершинах-предках не должны меняться, так как переменные-последствия не влияют на переменные-причины. Например, если есть каузальный граф для датасета фотографий птиц $\text{species} \rightarrow \text{color}$, то при интервенции $\text{do}(\text{color} = \text{blue})$ мы сможем увидеть сгенерированные изображения птиц в цветах, которые не встречаются у данного вида птиц.

В данной работе исследована задача генерации изображений с учетом каузальности между метками изображения и скрытыми факторами, которые можно считать ненаблюдаемыми конфаундерами. Подробнее о борьбе с проблемой ложных корреляций при генерации изображений рассказано в разделе 3. Исследуется генерация изображений с помощью вариационного автоэнкодера (VAE), который широко используется для извлечения низкоразмерных независимых факторов из наблюдаемых изображений. Чтобы иметь возможность генерации изображений с заданными метками, используется Conditional VAE, в энкодер и декодер которого дополнительно подается нужная метка. Утверждается, что при обусловленности на метки латентные признаки Z перестают быть независимыми. Мы предлагаем способ учета причинно-следственных связей между метками и латентными признаками для улучшения качества генерации.

2 ПОСТАНОВКА ЗАДАЧИ

В данной задаче рассматриваются изображения с несколькими метками, они могут быть независимыми, быть статистически зависимыми (иметь ненулевую корреляцию) или иметь нижележащую причинно-следственную структуру. Метки могут быть как категориальными, так и вещественными переменными. Пусть дан датасет $\mathcal{D}_{\text{train}} = (\mathcal{X}_{\text{train}}, \mathcal{Y}_{\text{train}})$, где $\mathbf{x} \in \mathcal{X}_{\text{train}}$ – изображение, $\mathbf{y} \in \mathcal{Y}_{\text{train}} \subset \mathbb{R}^k$ – вектор меток. Также дан набор *контрафактивных* меток \mathcal{Y}_{CF} , то есть таких, что $\mathcal{Y}_{\text{train}} \cap \mathcal{Y}_{\text{CF}} = \emptyset$. Задача – обучить генеративную модель G на датасете \mathcal{D} , способную генерировать правдоподобные контрафактивные изображения по меткам из \mathcal{Y}_{CF} . При этом модель G также должна уметь генерировать изображения и по уже виденным в $\mathcal{D}_{\text{train}}$ меткам.

Критерием качества сгенерированного по меткам \mathbf{y}_{CF} изображения $\hat{\mathbf{x}}_{\text{CF}} = G(\mathbf{y}_{\text{CF}})$ в данной работе является соответствие $\hat{\mathbf{x}}_{\text{CF}}$ каждой из меток \mathbf{y}_{CF} . Для получения численной оценки этого соответствия обучается модель M на совместном датасете обучающих и контрафактивных изображений $\mathcal{D} = (\mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{CF}}, \mathcal{Y}_{\text{train}} \cup \mathcal{Y}_{\text{CF}})$. В случае категориальных меток M – multi-label классификатор, выдающий логиты для каждой из меток. Пусть $M_{\mathbf{y}}(\mathbf{x}) \in \mathbb{R}^k$ – вектор логитов, соответствующих меткам $\mathbf{y} \in \mathbb{R}^k$, тогда метрика качества:

$$\sum_{\mathbf{y} \in \mathcal{Y}_{\text{train}} \cup \mathcal{Y}_{\text{CF}}} M_{\mathbf{y}}(G(\mathbf{y})) \rightarrow \max_G \quad (1)$$

В случае вещественных меток модель M – регрессор, возвращающая по изображению численные предсказания для каждой из меток. Здесь качество сгенерированного изображения можно оценить с помощью среднеквадратичной ошибки между истинной и предсказанной регрессором метками:

$$\sum_{\mathbf{y} \in \mathcal{Y}_{\text{train}} \cup \mathcal{Y}_{\text{CF}}} \text{MSE}(\mathbf{y}, M(G(\mathbf{y}))) \rightarrow \min_G \quad (2)$$

Данная задача является разновидностью задачи доменной генерализации, однако в отличие от стандартной постановки в нашем случае требуется обучить не дискриминативную модель, а генеративную, что значительно сложнее. Для решения задач доменной генерализации широко применяются техники причинного вывода [8, 9, 10], позволяющие отделить каузальные факторы от конфаундеров и использовать эти знания для классификации на новом домене.

Мы расширяем эту идею на задачу генерации изображений и вводим следующие постановки задач причинного обнаружения и вывода:

1. Причинное обнаружение. Имея метки $\mathbf{y} \in \mathbb{R}^k$ и соответствующие изображениям \mathbf{x} факторы $\mathbf{z} \in \mathbb{R}^d$, обучить структурно-каузальную модель в виде ациклического ориентированного графа, задающего связи между $k + d$ компонентами.
2. Причинный вывод. По найденному графу произвести интервенцию, установив желаемое значение $\text{do}(Y = \mathbf{y})$, и вывести соответствующие значения факторов $\mathbf{z} \in \mathbb{R}^d$. Генератор G принимает \mathbf{z}, \mathbf{y} и возвращает сгенерированное изображение \mathbf{x} , качество которого оценивается с помощью [1] или [2].

Здесь в качестве факторов \mathbf{z} могут выступать любые латентные переменные, которые улавливают скрытую структуру данных. Эти переменные могут включать в себя такие характеристики, как стиль изображения, освещение, положение объектов, а также любые другие аспекты, неявно присутствующие в данных, но влияющие на визуальное представление. В последующем, в контексте вариационных автоэнкодеров (VAE), \mathbf{z} будут интерпретироваться как латентные векторы, обученные для захвата значимой вариации в данных, что позволяет генератору G создавать новые изображения, придерживаясь заданных меток \mathbf{y} .

3 ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

3.1 ПРИЧИННАЯ ГЕНЕРАТИВНО-СОСТАЗАТЕЛЬНАЯ СЕТЬ (CAUSALGAN) [1]

В этой работе предлагается процедура состязательного обучения для обучения неявной генеративной модели для заданного причинного графа на метках изображений. По данному графу строится специальная архитектура генератора: вычисление признаков происходит строго в порядке от родителей к детям, таким образом, каузальные связи напрямую учитываются в процессе генерации. Такая архитектура позволяет легко проводить интервенции в значения меток – достаточно передать желаемые значения меток на вход в соответствующие блоки генератора. Авторы рассматривают случай *бинарных* меток, таких как пол, наличие улыбки, наличие усов и волос в датасете фотографий лиц. Утверждается, что обучить модель совместному распределению изображений x и набору бинарных меток y – крайне сложная задача, поэтому авторы предлагают способ двухстадийного обучения модели. Сначала обучить модель на определенном подмножестве меток, а потом дообучить ее, обуславливаясь на первом обученном множестве меток. Авторы предлагают новую функцию потерь, аналогичную классической функции потерь GAN, и добиваются неплохого качества генерации изображений с контрафактивными комбинациями.

Ключевыми недостатками модели CausalGAN являются:

1. нестабильность обучения, как и любой другой сети с состязательным режимом обучения;
2. ограничение на бинарность меток;
3. необходимость задания причинно-следственного графа для датасета.

3.2 ПРИЧИННЫЙ ВАРИАЦИОННЫЙ АВТОЭНКОДЕР (CAUSALVAE) [2]

Данная работа направлена на учет причинно-следственных связей между метками в процессе генерации изображений с помощью вариационного автоэнкодера (VAE). Ориентированный ациклический граф на *концептах* из изображения представляется в виде верхнетреугольной матрицы смежности A и обучается в процессе обучения всего автоэнкодера. Эта матрица используется для преобразования независимых экзогенных факторов ε_i в причинные представления z_i , и именно они подаются в генератор VAE для получения изображений. Для подсчета z_i предлагается линейная структурная кау-

зальная модель (SCM) в виде $z_i = g_i(\mathbf{A}_i \otimes \mathbf{z}; \eta_i) + \varepsilon_i$, где η_i – обучаемый параметр, g_i – нелинейная обратимая функция. Таким образом, $\mathbf{A}_i \otimes \mathbf{z}$ – вектор, содержащий информацию только о предках z_i в каузальном графе. Этот слой причинного маскирования позволяет проводить интервенции на латентном причинном представлении z_i – для этого нужно зафиксировать желаемое значение z_i и на входе, и на выходе каузального слоя SCM. В результате маскирования эффект интервенции передается всем потомкам вершины z_i , не затрагивая ее предков. В итоге, интервенция производится в некоторое изображение, которое проходит через энкодер, далее устанавливаются некоторые новые значения латентных признаков и генерируется новое измененное изображение.

Модель CausalVAE обладает существенными преимуществами по сравнению с CausalGAN, так как не требует двухстадийного обучения, имеет способность работать и с небинарными метками и обучается более стабильно, сразу на несколько задач: обучение каузального графа в виде числовой матрицы (с ограничением на DAG-ness) и максимизация ELBO (Evidence Lower BOund) в рамках вариационного байесовского вывода. Однако CausalVAE имеет ряд ограничений:

1. для интерпретируемости интервенций размерность латентного пространства должна быть в точности равно количеству меток, что сильно ограничивает гибкость VAE для настройки под датасеты разной сложности;
2. интервенция производится установкой вещественных значений в латентных причинных представлениях z_i , нежели установкой значений категориальных меток y_i напрямую, что уменьшает интерпретируемость интервенций.

3.3 Причинная диффузионная модель (DIFF-SCM) [3]

Семейство генеративных моделей, основанных на процессах диффузии [11], недавно привлекло внимание, достигнув state-of-the-art качества генерации изображений [12]. Санчез, Цафтарис [3] предложили алгоритм контрафактивной генерации изображений с помощью диффузионной модели, названный Diff-SCM. По аналогии с методом генеративной аугментации GenInt и интервенции в латентные коды в CausalVAE, контрафактивные изображения генерируются в три этапа:

1. *noise abduction* – извлечение экзогенного шума, или латентного представления исходного изображения;
2. *action* – проведение интервенции в каузальном графе, т. е. отрезание ребер, ведущих

щих в вершину, на которой производится интервенции;

3. *prediction* – генерация изображения из измененного латентного вектора, в случае Diff-SCM – обратный диффузионный процесс с измененным направлением.

В статье Diff-SCM рассматривается простейший каузальный граф на двух вершинах *класс* → *изображение*, а интервенции производят на метке класса, таким образом минуя второй этап. Чтобы придать обратному процессу нужное направление, обучаются модели в анти-причинном направлении, или же классификаторы. Во время третьего этапа в формуле обратного стохастического дифференциального уравнения [13] используются градиенты классификаторов по входу вместо score-функции (оценки $\nabla_{\mathbf{x}_t} p_t(\mathbf{x}_t)$), где \mathbf{x}_t – зашумленный элемент на шаге t , $p_t(x)$ – распределение \mathbf{x}_t .

Авторы предлагают вариант использования их метода и в общем случае. Каузальный граф на концептах изображения считается известным, а вершины, по которым производится контрафактивная генерация, должны иметь возможность восстановить информацию о своих родителях (например, анти-причинный классификатор). Для каждой из этих вершин обучаются своя диффузионная модель, и семплирование производится в причинных направлениях с использованием анти-причинных градиентов. В такой постановке процесс диффузии можно воспринимать как постепенное ослабление причинных связей. Проводятся эксперименты по генерации контрафактивных изображений на датасете MNIST (изображения других классов с заданным стилем, т.е. начертанием цифры). Хотя с помощью такой модели возможна и простая генерация по желаемой метке, и изменение данного изображения, по сравнению с нашим подходом Diff-SCM отличается в следующем:

1. данная модель не подразумевает причинного обнаружения, каузальный граф считается известным;
2. в исходном варианте, реализованном в статье, SCM модель строят только для случая одной метки – класса изображения, не учитывая multi-label случай;
3. применении Diff-SCM к multi-label случаю крайне вычислительно затратно, так как необходимо обучить диффузионную модель для каждой их меток, на которой планируется проводить интервенцию.

3.4 Модели для причинного обнаружения

В причинном обнаружении часто рассматриваются обобщения направленных ациклических графов (ADMG), которые описывают причинные связи между переменными с использованием направленных и двунаправленных ребер. Традиционные методы причинного обнаружения, учитывающие скрытые помехи, такие как алгоритм FCI [14] и его расширения [15, 16, 17], опираются на идентификацию класса эквивалентности ADMG графов, обладающих одинаковыми условными независимостями. Однако без дополнительных допущений эти методы могут давать малоинформативные результаты, так как они не способны различать члены одного и того же марковского класса эквивалентности [18]. Недавно для выявления скрытых искажающих факторов были разработаны методы обнаружения причин, основанные на структурно-каузальных моделях [19, 20].

Все вышеупомянутые подходы используют поиск в дискретном пространстве причинных структур, что часто требует процедур поиска, специфичных для конкретной задачи, и налагает вычислительную нагрузку для крупномасштабных задач. Чжэн и др. [21] предложили дифференцируемое ограничение на ориентированные ациклические графы (DAG) и сформулировали проблему обучения структуры графа как задачу оптимизации с дифференцируемыми ограничениями при отсутствии скрытых конфаундеров. Далее это обобщается на случай скрытого смешивания [22] посредством дифференцируемых алгебраических ограничений, которые характеризуют пространство ADMG.

Вышеупомянутые подходы полагаются на ограничительные линейные функциональные предположения и/или дискретный поиск в дискретном пространстве причинных графов. За последний год разработано несколько подходов причинного обнаружения с помощью нейронных сетей. Санчез, Лиу и др. [23] разработали алгоритм топологической сортировки для обнаружения графа на основе диффузионных моделей при очень объемных графах. Модель CAUSICA [24] – новый градиентный подход к изучению ADMG с нелинейными функциональными зависимостями на основе наблюдений. Авторы расширяют возможности дифференцируемого обнаружения ADMG для линейных моделей [22] на нелинейные случаи с использованием нейронных причинных моделей. В его основе лежит модель Deep End-to-end Causal Inference (DECI) [25], которую мы будем использовать в режиме обнаружения несмешанного направленного ориентированного графа (DAG).

4 ИССЛЕДОВАНИЕ И ПОСТРОЕНИЕ РЕШЕНИЯ ЗАДАЧИ

4.1 Устройство CVAE

Для получения скрытых факторов \mathbf{z} по изображению \mathbf{x} , следуя примеру [2], мы используем вариационный автоэнкодер (VAE). Он часто используется для выделения независимых факторов из изображений, так как в отличие от обычного автоэнкодера имеет в лоссе компоненту регуляризации дивергенции Кульбака-Лейблера (KL) между апостериорной частью скрытых факторов и стандартным нормальным распределением. Мы используем Conditional VAE (CVAE) [26], где на вход энкодеру и декодеру также подается метка \mathbf{y} , его схема представлена на рисунке 3. В случае k категориальных меток каждая из них представляется в виде one-hot вектора и подается в сконкатенированном виде.

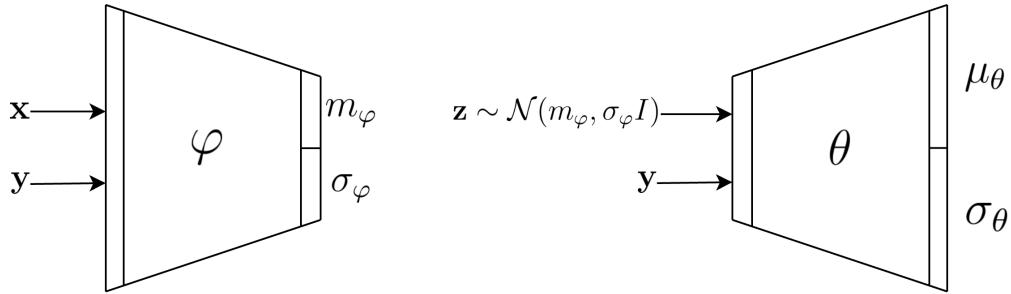


Рис. 3: Схема Conditional VAE

Пусть θ – веса декодера, который моделирует распределение $\mathbf{x}|\mathbf{z}$, φ – веса энкодера, который моделирует распределение $\mathbf{z}|\mathbf{x}$, d – размерность латентного пространства. Вводится вариационное семейство $Q = \{q_\varphi(\mathbf{z}) = \prod_{i=1}^n \prod_{j=1}^d q_{\varphi_j}(z_{ij})\}$, где q_{φ_j} – плотность $\mathcal{N}(m_{\varphi_j}(\mathbf{x}_i, \mathbf{y}_i), \sigma_{\varphi_j}^2(\mathbf{x}_i, \mathbf{y}_i))$; $m_\varphi, \sigma_\varphi \in \mathbb{R}^d$ – выход энкодера. Распределением из этого семейства будет приближаться истинное апостериорное распределение $p(\mathbf{z}|\mathbf{x})$: $q^* = \arg \min_{q \in Q} KL(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}))$. Для простоты реализации используется репараметризация:

$$\mathbf{z}_i = m_\varphi(\mathbf{x}_i, \mathbf{y}_i) + \sigma_\varphi(\mathbf{x}_i, \mathbf{y}_i) \odot \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3)$$

где \odot – поэлементное умножение. Обучение CVAE на датасете $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ происходит за счет максимизации Evidence lower bound (ELBO):

$$\text{ELBO}_{\mathcal{D}}(\theta, \varphi) = \sum_{i=1}^n \mathbb{E}_\varepsilon \log p_\theta(\mathbf{x}_i | \mathbf{z}_i = m_\varphi(\mathbf{x}_i, \mathbf{y}_i) + \sigma_\varphi(\mathbf{x}_i, \mathbf{y}_i) \odot \boldsymbol{\varepsilon}_i, \mathbf{y}_i) - \sum_{i=1}^n KL(q_\varphi(\mathbf{z}_i | \mathbf{x}_i, \mathbf{y}_i) \| p(\mathbf{z}_i)), \quad (4)$$

где $p(\mathbf{z}_i)$ – плотность $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Первая компонента – логарифм неполного правдоподобия, а вторая выступает в роли регуляризатора, который приближает распределение латентных векторов к стандартному гауссовскому распределению. Стоит отметить, что при добавлении условности на метки \mathbf{y}_i модель не имеет теоретической гарантии, что распределение латентных представлений будет независимым и стандартным гауссовским для каждой из меток. Это выполняется лишь для данных по всем меткам в совокупности. Поэтому генерация просто с использованием латентных представлений может выдавать некачественные результаты, особенно на датасетах с большим дисбалансом классов или сдвигом домена. Таким образом, при работе с латентным пространством CVAE требуются более продвинутые методы, учитывающие нижележащие зависимости между скрытыми факторами и метками.

4.2 КАУЗАЛЬНОЕ ОБУЧЕНИЕ НА ОСНОВЕ РАЗБИЕНИЯ ДАННЫХ

Одной из распространенных идей решения задачи доменной генерализации является умное разделение данных [8]. Мы попробовали разбить весь обучающий датасет $\mathcal{D}_{\text{train}}$ на две не пересекающиеся по комбинациям меток части – \mathcal{D}_{Tr} и $\mathcal{D}_{\text{TrCF}}$ и обучить CVAE, используя модификацию лосса с ELBO [4]. Пусть G – генератор, $G(\mathbf{y}) \sim \mathcal{N}(\mu_\theta(\mathbf{z}, \mathbf{y}), \sigma_\theta(\mathbf{z}, \mathbf{y}))$, $\mu_\theta, \sigma_\theta$ – выходы декодера CVAE, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, тогда модифицированный лосс:

$$\tilde{\mathcal{L}}_{\text{CVAE}}(\theta, \varphi) = -\text{ELBO}_{\mathcal{D}_{\text{Tr}}}(\theta, \varphi) + \lambda \cdot \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{TrCF}}} \ell(M(G(\mathbf{y})), \mathbf{y}), \quad (5)$$

где ℓ и M – кросс-энтропия и классификатор в случае категориальных меток, MSE и регрессор в случае вещественных; λ – вес второй компоненты. Сеть M обучается на исходном обучающем датасете $\mathcal{D}_{\text{train}}$. Таким образом, сеть обучает латентное пространство на \mathcal{D}_{Tr} и одновременно максимизирует качество сгенерированных изображений для комбинаций меток из $\mathcal{D}_{\text{TrCF}}$, что повышает генерализационную способность модели. Такой подход имеет ряд минусов, основным из которых является необходимость в большем количестве данных. Подробности обучения см. в разделе [5].

4.3 ПАЙПЛАЙН ПРИЧИННОГО ОБНАРУЖЕНИЯ И ВЫВОДА

Опишем решение, имплементирующее идею причинного обнаружения и вывода, описанного в разделе [2]. В качестве скрытых факторов, соответствующих изображениям \mathbf{x} , используются латентные представления \mathbf{z} стандартной модели CVAE.

Далее требуется восстановить структурно-каузальные отношения между метками $\mathbf{y} \in \mathbb{R}^d$ и скрытыми факторами $\mathbf{z} \in \mathbb{R}^k$. В общем случае задача стоит в выявлении причинного графа на $k+d$ вершинах, где направленное ребро между двумя вершинами означает отношение влечения. Как бейзлайн для нахождения графа будем использовать модель для причинного обнаружения DECI, лежащую в основе CAUSICA. Она обучается на наборе данных в численном виде и возвращает граф в виде бинарной матрицы смежности ориентированного графа (рисунок 4a). Мы рассмотрели несколько вариантов обучения модели CAUSICA:

1. Запрет на ребра, ведущие в компоненты меток y_i . Влечения $z_i \rightarrow y_j$ невозможны логически, так как компоненты латентного пространства вычисляются с использованием y_j . А влечения $y_i \rightarrow y_j$ возможны, но не используются в нашем пайплайне причинного вывода, где интервенция проводится в каждую из компонент вектора \mathbf{y} (рисунок 4b). Без последнего ограничения модель решает более общую задачу, так как имеет возможность восстанавливать некоторые отсутствующие метки, но и является более сложной для обучения. В разделе 5 описаны результаты экспериментов и с ограничением, и без.
2. Регуляризация на количество ребер. С ростом размерности латентного пространства d количество вершин в графе возрастает, и возникает потребность регулировать сложность модели, возможно, с небольшими потерями в точности графа. Регуляризация реализуется с помощью добавления в лосс нового слагаемого – текущего количества ребер с некоторым коэффициентом κ .

Однако для получения численных значений латентных компонент отношения влечения в графе также должны иметь численное представление. Поэтому введем новую модель Structural Causal Model (SCM), где каждая вершина имеет линейную зависимость от всех своих родителей с обучаемым свободным членом. Обучение SCM проводится на тех же данных, что и обучение CVAE. Итак, пайплайн обучения состоит из следующих шагов:

1. обучение базовой модели, например CVAE, на датасете изображений \mathbf{x} и меток \mathbf{y} ;
2. причинное обнаружение: выявление графа отношений влечения на компонентах соответствующих меткам \mathbf{y} латентных векторов \mathbf{z} , полученных из энкодера базовой модели;
3. уточненное причинное обнаружение: обучение модели SCM, которая обучает чис-

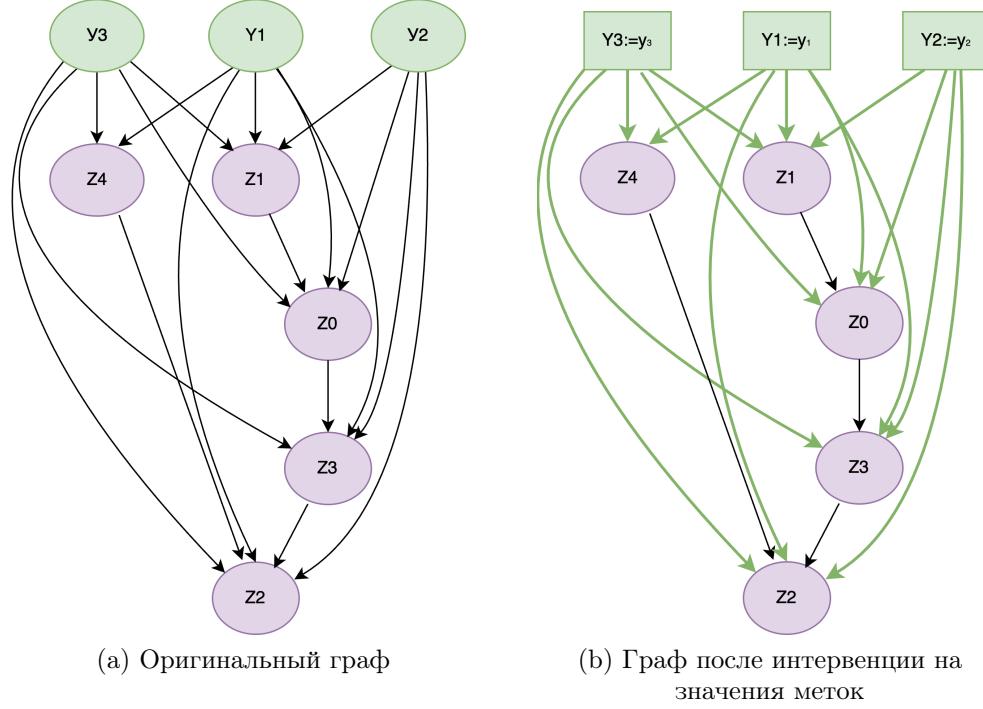


Рис. 4: Обученный график причинно-следственных связей на 3 категориальных метках и размерности латентного пространства $d = 5$

ленную линейную зависимость между вершинами и влекущими их родителями.

Общая схема обучения представлена на рисунке 5

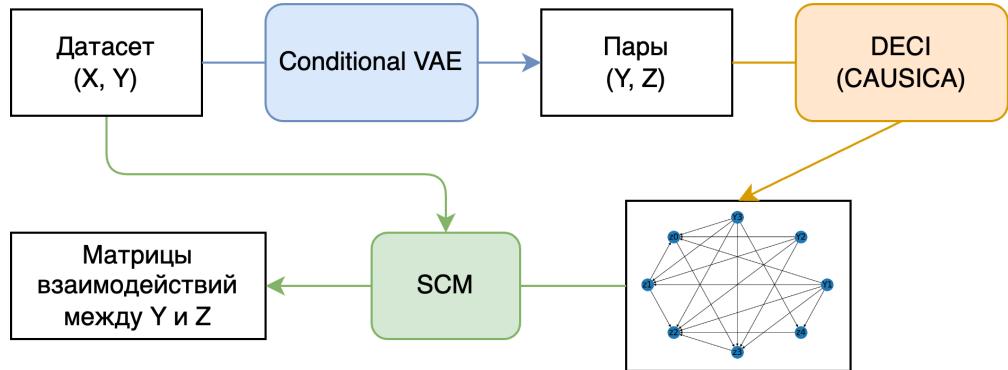


Рис. 5: Пайплайн причинного обнаружения и вывода

На этапе применения желаемые значения меток \mathbf{y} подставляются в график, с помощью обученных весов SCM вычисляются компоненты среднего μ_{z_i} , каждая из которых семплируется: $z_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i})$. Процесс происходит рекурсивно в порядке от родителей к детям. Выбор σ_{z_i} нетривиален, и несколько вариантов рассмотрено в разделе 5. Готовый латентный вектор \mathbf{z} подается в декодер CVAE для получения изображения \mathbf{x} .

4.4 ПРИМЕНЕНИЕ CVAE-GAN

В ходе проведения экспериментов выясноено, что при увеличении размерности латентного пространства d CVAE лучше реконструирует изображения, но в то же время теряет способность генерации изображений, когда на вход подается произвольная метка \mathbf{y} и $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Это может быть вызвано тем, что в задаче с несколькими метками обусловленность на \mathbf{y} становится намного сложнее. Другая возможная причина – то, что в лоссе CVAE нет компоненты, отвечающей за простую генерацию изображений по стандартному нормальному вектору из латентного пространства.

Для решения проблемы неточной генерации в случае с большой размерностью латентного пространства ($d > 10$) мы рассматриваем модель Conditional VAE/GAN [27], которая помимо энкодера E и генератора G CVAE имеет дискриминатор D и обучается одновременно на несколько задач. К стандартному лоссу VAE из формулы (4) добавляется состязательный лосс GAN: $\mathcal{L}_{\text{GAN}} = \log(D(\mathbf{x})) + \log(1 - D(G(\mathbf{z})))$. Схема модели изображена на рисунке 6. Таким образом, во время обучения распределение латентных векторов приближается к стандартному гауссовскому и энкодер с генератором обучаются восстанавливать изображения. Но благодаря GAN-лоссу генератор также учится генерировать правдоподобные изображения, получая на вход только \mathbf{z} .

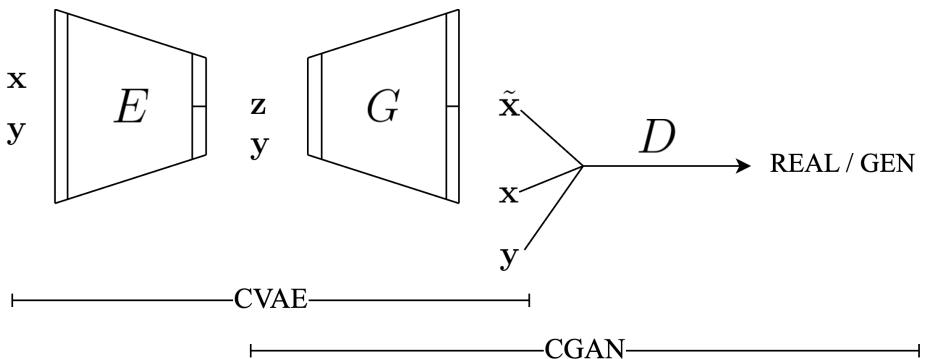


Рис. 6: Схема модели CVAE/GAN;
 E – энкодер, G – декодер / генератор, D – дискриминатор

Модель VAE/GAN обучается состязательно – несколько эпох обновления весов энкодера и генератора с замороженным дискриминатором, затем несколько эпох обновления весов дискриминатора с замороженным энкодером и генератором. Перечислим компоненты лосса модели.

1. Обучение E и G . Учитываем точность восстановления изображений ($\mathbf{x} \xrightarrow{E} \mathbf{z} \xrightarrow{G} \hat{\mathbf{x}}$), правдоподобие восстановленных изображений ($\mathbf{x} \xrightarrow{E} \mathbf{z} \xrightarrow{G} \hat{\mathbf{x}} \xrightarrow{D} 1$) и правдо-

подобие сгенерированных изображений ($\mathbf{z} \xrightarrow{G} \hat{\mathbf{x}} \xrightarrow{D} 1$).

2. Обучение D. Различаем сгенерированные изображения от истинных ($\mathbf{x} \xrightarrow{E} \mathbf{z} \xrightarrow{G} \hat{\mathbf{x}} \xrightarrow{D} 0$; $\mathbf{z} \xrightarrow{G} \hat{\mathbf{x}} \xrightarrow{D} 0$; $\mathbf{x} \xrightarrow{D} 1$).

Как и любое состязательное обучение, оно менее стабильно по сравнению с обычным CVAE. Также оно требует обучения трех сетей и больше времени для сходимости. В нашей постановке с multi-label датасетами мы используем версию VAE/GAN, обусловленную на \mathbf{y} , где по аналогии с CVAE каждой из сетей E, G, D на вход подаются еще и метки в сконкатенированном one-hot представлении.

Чтобы убедиться в эффективности каузального пайплайна (рисунок 5), мы проводим эксперименты для простой генерации из латентного пространства CVAE/GAN и для генерации с использованием обученного с помощью CAUSICA и SCM графа.

5 ОПИСАНИЕ ПРАКТИЧЕСКОЙ ЧАСТИ

Реализация генеративных моделей и оценка их качества произведены на языке программирования Python. В частности, для реализации нейронных модулей, сетей классификаторов и регрессоров использовалась библиотека PyTorch [28], а все вероятностные модели реализованы с помощью библиотеки Руго [29].

5.1 ДАТАСЕТ COLORED MNIST

Следуя примеру [30], мы сгенеририровали модификацию датасета MNIST [31] с изменением цвета фона и цифры под названием Colored MNIST. Он состоит из цветных изображений размера 28×28 , которые разделены на обучающую (60 000 изображений) и тестовую (10 000 изображений) части. Каждое изображение имеет 3 метки: цифра, цвет цифры, цвет фона, всего 10 возможных цветов; комбинации с одинаковым цветом фона и цифры пропускаются. Для разбиения данных согласно постановке задачи ($\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{CF}}$) на контрафактивную часть выделено 5 комбинаций меток, которые можно видеть на рисунке 7.

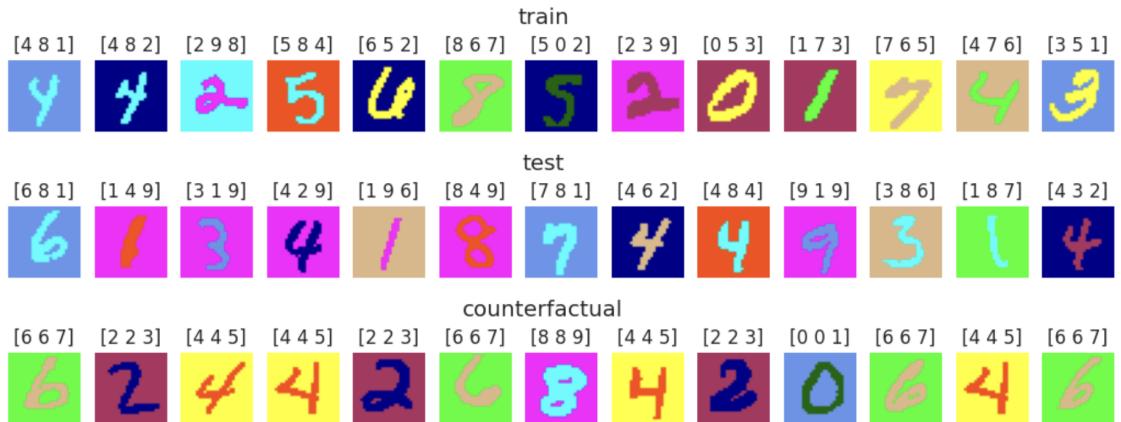


Рис. 7: Примеры изображений и меток датасета Colored MNIST, метка представлена в виде [цифра, цвет цифры, цвет фона]

Для оценки качества изображений предварительно обучен классификатор M на совместном датасете \mathcal{D} . Это небольшая сверточная сеть с 2 сверточными и 2 линейными слоями, имеющая 323.8К параметров, выдающая 30 логитов, по 10 на каждую метку. После обучения 20 эпох на CPU удалось добиться 99.21% точности на тестовом датасете.

5.1.1 ОБУЧЕНИЕ БАЗОВЫХ МОДЕЛЕЙ

Модель CVAE реализована с помощью Руго, в котором есть возможность регистрировать нейросетевые модули, например энкодер и декодер. Энкодер представляет из

себя 4-слойную сверточную нейронную сеть с батч-нормализацией, max-pooling, функцией активации Leaky ReLU и dropout с вероятностью 0.1. Метка подается вместе с признаками изображения в линейные слои для получения $m(\mathbf{x}), \sigma(\mathbf{x}) \in \mathbb{R}^d$. Декодер – аналогичная 4-слойная сверточная нейросеть с повышением размерности по методу ближайшего соседа. Сеть обучалась для $d = 3$ и $d = 5$ с learning rate 5×10^{-4} на GPU NVIDIA T4 в течении 100 эпох.

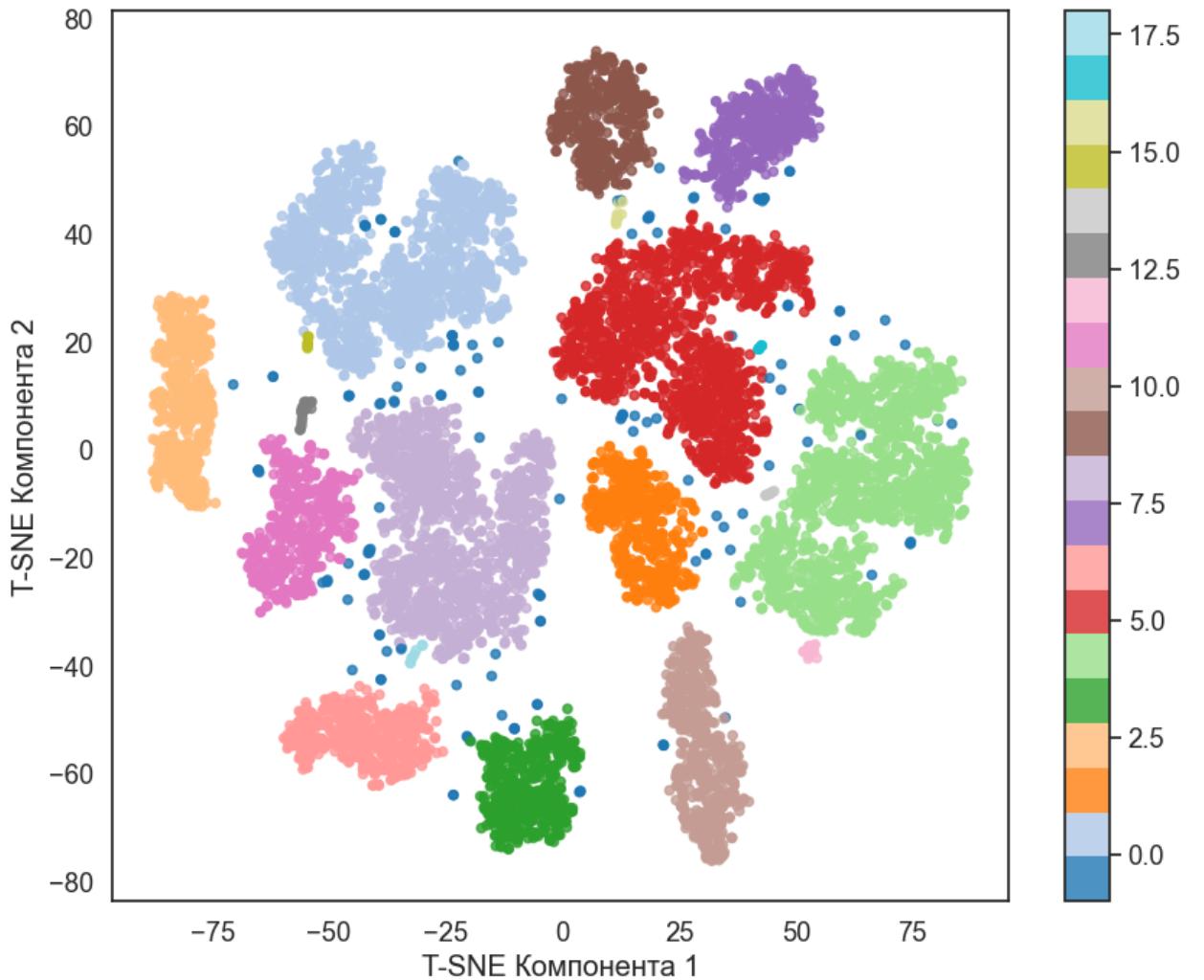
В режиме каузального обучения на основе разбиения данных (секция 4.2) часть датасета $\mathcal{D}_{\text{train}}$ разбита на \mathcal{D}_{Tr} и $\mathcal{D}_{\text{TrCF}}$ так, что $\mathcal{Y}_{\text{TrCF}} = \{(i, i, (i + 2)\%10)\}_{i=0,\dots,9}$. Архитектуры энкодера и декодера такие же, как в обычном CVAE, для честности их сравнения. Как видно из формулы 5, коэффициент λ отвечает за вес лосса классификатора по отношению к лоссу VAE, в экспериментах положено $\lambda = 0.1$. Этую модель обозначим CVAE'.

Как было замечено в пункте 4.1, при обусловленности на метки \mathbf{y} распределение латентных представлений CVAE перестает быть независимым и стандартным гауссовским для каждой из меток. Чтобы продемонстрировать это, мы собрали выходы энкодера $m_\varphi(\mathbf{x}, \mathbf{y})$ на всем датасете $\mathcal{D}_{\text{test}}$ для размерности латентного пространства $d = 30$, сделали их кластеризацию методом DBSCAN [32] с последующим понижением размерности до 2 компонент методом t-SNE [33]. Результат кластеризации представлен на рисунке 8, а анализ зависимостей кластеров от комбинаций меток – на рисунке 9. Явно видно, что данные разбиваются на кластеры по определенным значениям меток, что говорит о неполном покрытии латентного пространства. Это является основной причиной ухудшения качества генерации CVAE при больших d и при большом количестве меток путем простого семплирования $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Для $d = 10$ была обучена сеть CVAE/GAN. Архитектуры энкодера и декодера отличаются от предыдущих добавлением 1 сверточного слоя, дискриминатор – 3-слойная сверточная сеть, возвращающая одно число (логит правдоподобности). Соотношение количества эпох энкодера и декодера по отношению к дискриминатору – 12 к 10, learning rate 3×10^{-2} , 2×10^{-4} соответственно. Обучение проводилось на GPU NVIDIA T4 в течении 1900 итераций. График обучения представлен на рисунке 10.

5.1.2 Обучение моделей причинного обнаружения

Следующим шагом пайплайна является причинное обнаружение, в нашем случае применение модели CAUSICA к датасету пар меток и соответствующих им латентным

Рис. 8: Кластеризация $m_\varphi(\mathbf{x})$ для CVAE при $d = 30$

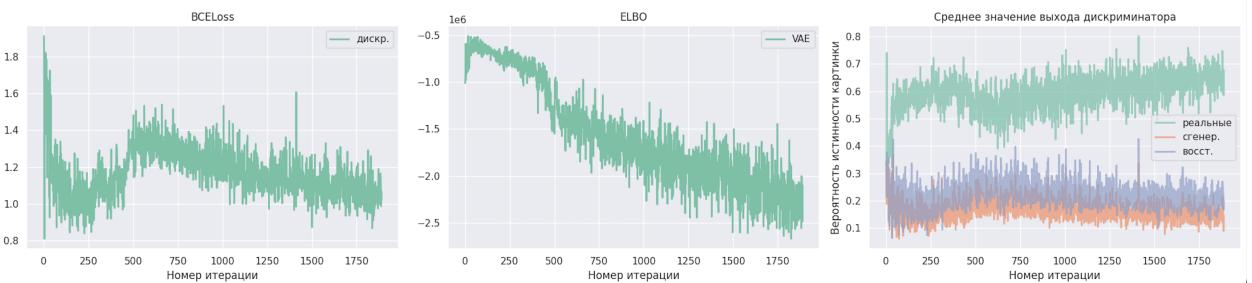
векторам. Для $d = 3$ и $d = 5$ с базовой моделью CVAE обучено два варианта модели: без запрета на ребра, входящие в Y-вершины и с ним. На самом деле, ребра $y_i \rightarrow y_j$ возможны, но не используются в нашем ходе инференса, то есть генерации изображений по полностью данной метке \mathbf{y} . Модель без запрета ребер могла бы быть полезна для неполных интервенций в метки, например, для переноса стиля изображения на частично новый класс.

Для $d = 10$ с базовой моделью CVAE/GAN модель обучена в двух вариантах: с регуляризацией по количеству ребер между z-компонентами и без. При увеличении размерности латентного пространства количество ребер квадратично возрастает, что значительно усложняет следующую в пайплайне модель SCM и может привести к переобучению. В ходе эксперимента без регуляризации CAUSICA обнаружила 42 ребра, а с регуляризацией с коэффициентом $\kappa = 0.07 - 36$ ребер.

Наконец, для каждого из рассмотренных d была обучена SCM-модель для опре-



Рис. 9: Случайные представители кластеров


 Рис. 10: График обучения CVAE/GAN для $d = 10$

деления численной зависимости в отношениях влечения в причинной модели. Модель реализована с помощью Руго и рекурсивного обхода графа, случайные величины задаются в порядке от родителей к детям, на вход на каждой итерации подаются батчи из выходов энкодера и меток: $\{m_\varphi(\mathbf{x}, \mathbf{y}), \sigma_\varphi(\mathbf{x}, \mathbf{y}), \mathbf{y}\}$. Например, если на вершину z_0 влияют вершины y_0, y_1 и z_1 , то перед обучением объявляется матрица весов a_{z_0} размера $1 \times (10 + 10 + 1 + 1)$, где последняя единица отвечает свободному члену. В процессе обучения сначала семплируются все 3 родителя, все y -компоненты представляются в виде one-hot вектора, перемножается покомпонентно с весами a_{z_0} , и получается значение

ние μ_{z_0} . Теперь можно сэмплировать $z_0 \sim \mathcal{N}(\mu_{z_0}, \sigma_{z_0})$, где σ_{z_0} – компонента $m_\varphi(\mathbf{x}, \mathbf{y})$ из энкодера CVAE для данной пары (\mathbf{x}, \mathbf{y}) .

Сама модель SCM очень легковесная по сравнению с базовой моделью для построения латентного пространства. Пример визуализации обученных весов для $d = 10$ и графа с 42 ребрами представлен на рисунке 11.

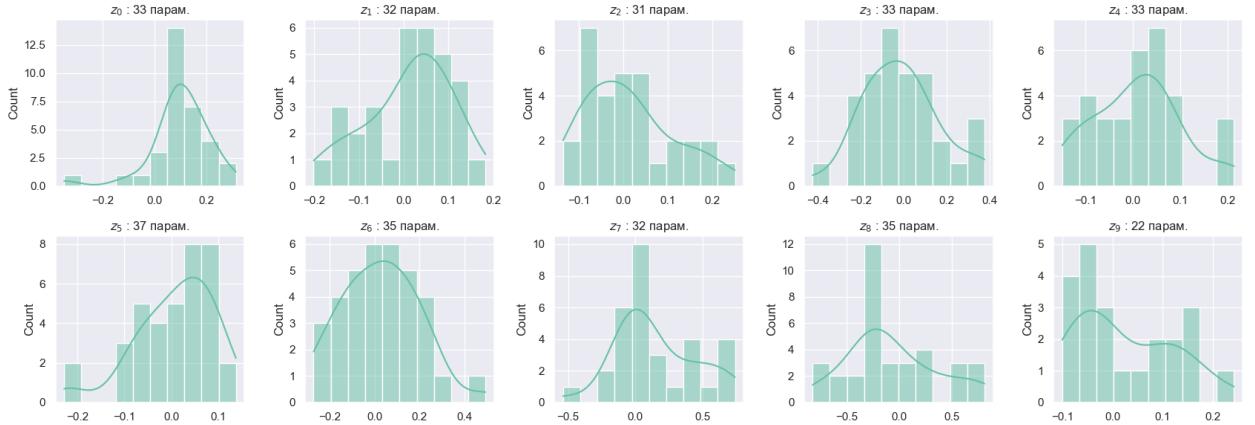


Рис. 11: Гистограммы значений весов обученной SCM для $d = 10$

Применение SCM происходит в том же порядке, все вершины без родителей семплируются со средним, равным своему параметру a , остальные – рекурсивно по методу, описанному выше. Одной из проблем является то, что на этапе инференса, согласно постановке задачи, доступны только метки \mathbf{y} , а для сэмплирования компонент z_i необходимо иметь соответствующие σ_{z_i} . Мы рассмотрели несколько вариантов выбора σ_{z_i} : $\sigma_\varphi(\mathbf{x}, \mathbf{y})$ случайного элемента датасета $\mathbf{x} \in \mathcal{X}_{\text{train}}$, его среднее, и линейно возрастающее от родителей к детям.

5.1.3 Результаты на датасете Colored MNIST

В таблице 1 представлены вероятности истинного класса, предсказанные классификатором M на сгенерированных изображениях. При оценке качества для каждой из рассмотренных комбинаций из $\mathcal{Y}_{\text{train}}, \mathcal{Y}_{\text{CF}}$ сгенерировано по 32 изображения. Введем обозначения:

- CVAE' – каузальное обучение CVAE на основе разбиения данных 4.2;
- +SCM – с добавлением причинного обнаружения и вывода;
- +SCM + constr. – с запретом на ребра, ведущие в y -компоненты;
- Y_1, Y_2, Y_3 – цифра, цвет цифры, цвет фона соответственно;

- train, CF – обучающие и контрафактивные метки.

d	Модель	$Y_{1,\text{train}}$	$Y_{2,\text{train}}$	$Y_{3,\text{train}}$	$Y_{1,\text{CF}}$	$Y_{2,\text{CF}}$	$Y_{3,\text{CF}}$
3	CVAE	0.586	0.351	0.416	0.629	0.202	0.392
3	CVAE'	0.623	0.384	0.444	0.628	0.208	0.414
3	+SCM	0.624	0.364	0.424	0.666	0.202	0.401
3	+SCM + constr.	0.703	0.388	0.431	0.830	0.204	0.397
5	CVAE	0.507	0.376	0.426	0.458	0.216	0.411
5	CVAE'	0.510	0.324	0.419	0.470	0.202	0.397
5	+SCM	0.595	0.404	0.432	0.773	0.204	0.449
5	+SCM + constr.	0.635	0.405	0.433	0.812	0.203	0.405

Таблица 1: Качество сгенерированных изображений для размерностей латентного пространства 3 и 5

Можно видеть, что применение нашего каузального пайплайна улучшает метрики, в особенности точность генерации самой формы цифры ($Y_{1,\text{train}}, Y_{1,\text{CF}}$), что демонстрируется на рисунке 12. Заметим, что изображения, сгенерированные CVAE, выглядят более блеклыми по сравнению с оригиналом. Это вызвано вероятностным характером VAE, которые предназначены для изучения вероятностного отображения латентного пространства в пространство данных. Вместо того, чтобы напрямую сопоставлять входные данные с конкретными выходными данными, они изучают распределения. Декодер реконструирует изображения путем выборки из этих распределений, что может привести к неопределенности и, как следствие, размытию генерируемых изображений.

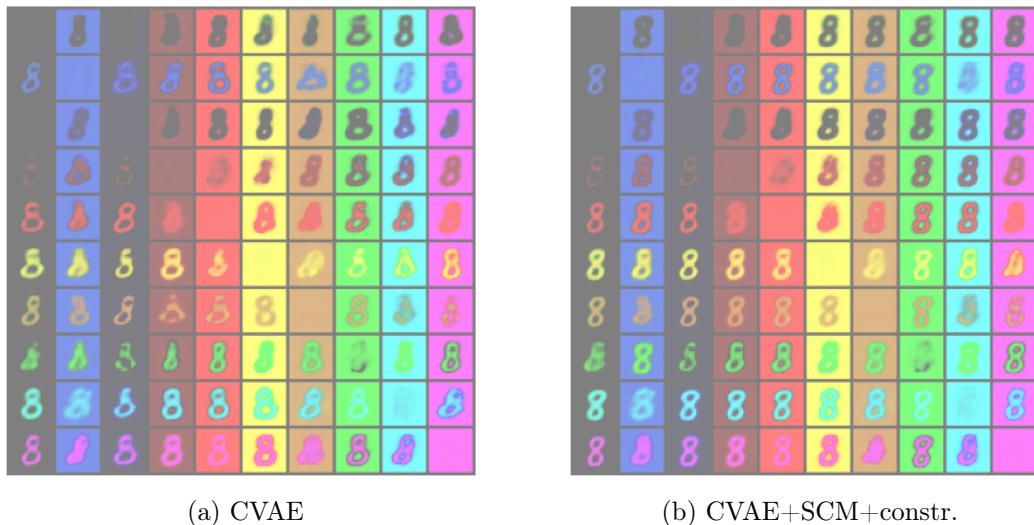


Рис. 12: Пример сгенерированных изображений цифры 8

Теперь рассмотрим эксперименты при $d = 10$, для которого была выбрана модель CVAE/GAN в качестве базовой. В силу состязательного обучения проблема размытия цвета в этой модели пропадает, что видно на рисунке 13. Причинное обнаружение с

помощью CAUSICA произведено в 2 вариантах – без регуляризации на количество ребер между z -компонентами и с ней. Также для каждого варианта рассмотрено 3 варианта семплирования на этапе SCM-вывода, а именно выбор σ_{zi} , описанный выше.

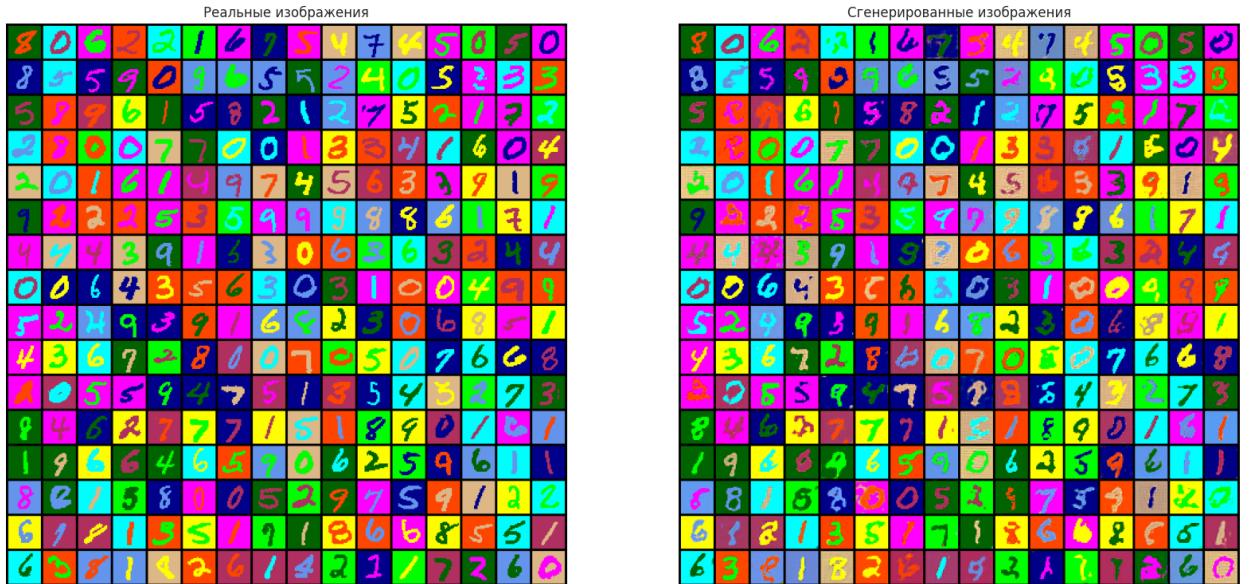


Рис. 13: Сравнение реальных изображений с сгенерированными моделью CVAE/GAN

Таблица 2 показывает, что, по сравнению с меньшими d и обычным CVAE, вероятность правильного класса сгенерированных изображений сильно возросла для меток Y_2, Y_3 . На обучающем датасете улучшения по этим метрикам не наблюдается. С другой стороны, модель хуже генерирует цифры по контрафактивным меткам. Тем не менее, применение каузального пайплайна с запретом на ребра, входящие в y -компоненты, улучшает метрики для Y_1 и $Y_{3,CF}$. Среди 3 вариантов выбора σ для генерации лучшим оказался первый, то есть соответствующий выход энкодера для случайного $\mathbf{x} \in \mathcal{D}_{\text{train}}$. Также при добавлении регуляризации качество для Y_1 немного ухудшается за счет того, что снижается сложность и вес модели SCM. Этот эксперимент является своего рода абляционным исследованием, показывающим, что использование всего множества ребер при обучении SCM важно, так как улучшает качество итоговой генерации.

5.2 ДАТАСЕТ PENDULUM

Для дальнейшего исследования модели решено проверить ее работоспособность на другом датасете, Pendulum [34]. Он состоит из цветных изображений размера 96×96 и содержит 4 метки: угол наклона маятника, положение света, длина тени и положение тени. Есть истинная каузальная зависимость на метках – первые две влекут последние две. Данные сгенерированы таким образом: $\mathcal{D}_{\text{train}}$ состоит из 10,184 изображений, где

Модель	Выбор σ	$Y_{1,\text{train}}$	$Y_{2,\text{train}}$	$Y_{3,\text{train}}$	$Y_{1,\text{CF}}$	$Y_{2,\text{CF}}$	$Y_{3,\text{CF}}$
CVAE/GAN		0.730	0.962	0.990	0.235	0.991	0.987
+SCM + constr.	$\sigma_\varphi(\mathbf{x}, \mathbf{y})$	0.746	0.962	0.991	0.307	1.000	0.992
	среднее	0.745	0.962	0.990	0.252	1.000	0.996
	лин. возр.	0.743	0.962	0.991	0.274	0.999	0.984
+SCM + constr. + reg.	$\sigma_\varphi(\mathbf{x}, \mathbf{y})$	0.742	0.963	0.990	0.257	1.000	1.000
	среднее	0.741	0.962	0.991	0.254	0.996	0.999
	лин. возр.	0.740	0.962	0.990	0.245	0.998	0.998

Таблица 2: Качество сгенерированных изображений для $d = 10$ и разного выбора σ при генерации

рассматриваются все комбинации Y_1, Y_2 , а значения Y_3, Y_4 выражаются через предыдущие с помощью формул законов физики. Датасет \mathcal{D}_{CF} содержит 1,698 изображений, в которых Y_1, Y_2, Y_3, Y_4 произвольные, соответственно законы физики не соблюдаются.

На всех данных в совокупности обучен нейросетевой регрессор M , который по изображению предсказывает все 4 метки в виде вещественных чисел. Это нейронная сеть из нескольких модулей для предсказания отдельно по маятнику, свету и тени, всего содержащая 2.2М параметров. Каузальный пайплайн был применен при следующих гиперпараметрах: $d = 3$, базовая модель – CVAE, $\kappa = 0$ (без регуляризации на количество ребер), ребра в y -компоненты запрещены, выбор σ во время генерации – $\sigma_\varphi(\mathbf{x}, \mathbf{y}), \mathbf{x} \in \mathcal{D}_{\text{train}}$. В соответствии с формулой [2], метрика качества в данном случае – MSE между истинными и предсказанными регрессором метками. Для проверки работоспособности базовой модели CVAE проведен эксперимент по реконструкции изображений из датасета. На контрафактивных комбинациях часто маятник и свет реконструируются точно, а тень – в соответствии с законами физики, как на рисунке [14]. При генерации из латентного пространства с контрафактивными комбинациями меток часто генерируется пустое изображение.

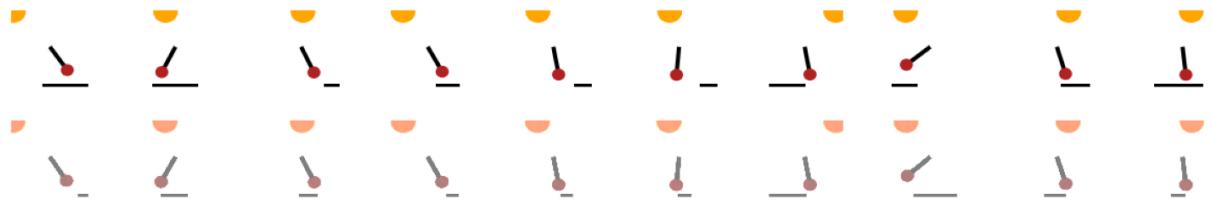


Рис. 14: Реконструкция CVAE контрафактивных комбинаций \mathbf{y}

Результаты применения каузального пайплайна к датасету Pendulum представлены в таблице [3]. Сравнения с моделью Causal VAE [2] произведено не было, так как ее реализация не выложена в открытый доступ. MSE посчитано отдельно на первой группе

меток-причин ((Y_1, Y_2)) и на группе меток-следствий ((Y_3, Y_4)). Можно видеть, что по сравнению с обычным CVAE каузальный пайплайн улучшает точность генерации тени и на $\mathcal{D}_{\text{train}}$, и на \mathcal{D}_{CF} . Не улучшилась только генерация источника света на контрафактивных метках, хотя из рисунка 15 видно, что CVAE+SCM часто генерирует более точные изображения.

Модель	$(Y_1, Y_2)_{\text{train}}$	$(Y_3, Y_4)_{\text{train}}$	$(Y_1, Y_2)_{\text{CF}}$	$(Y_3, Y_4)_{\text{CF}}$
CVAE	337.843	10.396	386.027	43.689
+SCM + constr.	330.515	10.343	391.768	42.579

Таблица 3: Ошибка MSE для сгенерированных изображений на датасете Pendulum

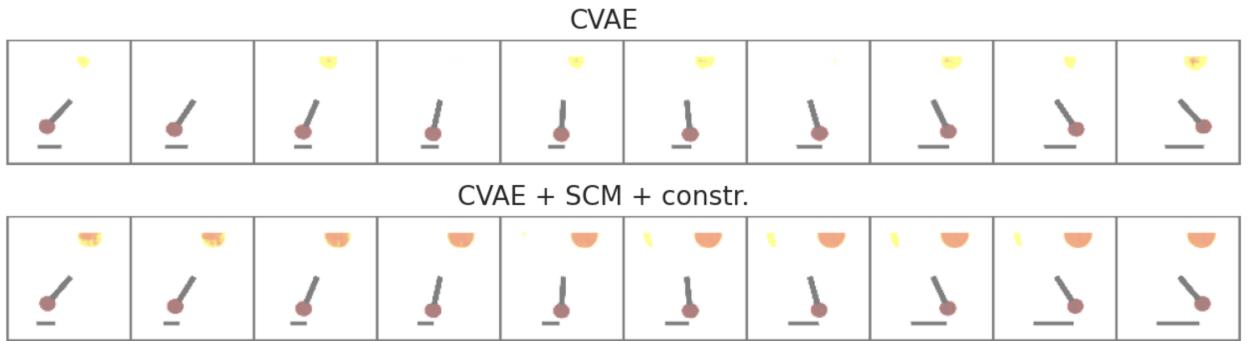


Рис. 15: Сравнение генерации двух моделей при одинаковых y

Таким образом, на датасете Pendulum также удалось увидеть улучшение в генерации при применении каузального пайплайна, особенно в метках-следствиях (параметрах тени), что подтверждает, что предложенный метод улучшения генерации имеет способность обобщения на более сложные датасеты.

6 ЗАКЛЮЧЕНИЕ

В данной работе подробно изучена тема применения причинно-следственного вывода для решения задачи генерации контрафактивных изображений. Рассматриваются датасеты с несколькими метками, и предполагается наличие нижележащей каузальной структуры. Приведен исчерпывающий обзор проблем, которые могут быть решены с помощью причинного вывода в области компьютерного зрения, среди них – проблема ложных корреляций и доменная генерализация.

Для решения задачи генерации с учетом нижележащих каузальных структур были рассмотрены существующие решения задач области причинного обнаружения и вывода. В ходе работы проведены обширные эксперименты и разработан каузальный пайплайн, решающий поставленную задачу. Исследовано множество базовых моделей для формирования скрытых факторов при различных размерностях латентного пространства d , таких как CVAE и CVAE/GAN. После обучения модели причинного обнаружения CAUSICA на наборе латентных векторов и соответствующих им меток, мы уточняем причинно-следственные связи с помощью новой модели SCM, которая обучает выявленные зависимости в виде весов. Для получения лучших значений гиперпараметров рассмотрено множество видов архитектур базовых сетей, ограничений на причинный граф и регуляризаций.

Проведено множество экспериментов на датасетах Colored MNIST и Pendulum, и измерено качество сгенерированных по меткам изображений. Они подтвердили гипотезу о том, что качество генерации действительно улучшается после применения разработанного каузального пайплайна.

В дальнейшем планируется продолжить исследование предложенной схемы для решения данной задачи, в частности, проведение следующих экспериментов:

- альтернативное обучение базовых моделей CVAE с размытием дисперсии при семplификации в латентном пространстве,
- учет кластеризации латентных векторов при построении причинного графа.

СПИСОК ЛИТЕРАТУРЫ

- [1] *Kocaoglu, Murat.* CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. — 2017.
- [2] *Yang, Mengyue.* CausalVAE: Structured Causal Disentanglement in Variational Autoencoder. — 2023.
- [3] *Sanchez, Pedro.* Diffusion Causal Models for Counterfactual Estimation. — 2022.
- [4] *Sprenger, Jan.* Simpson’s Paradox / Jan Sprenger, Naftali Weinberger // The Stanford Encyclopedia of Philosophy / Ed. by Edward N. Zalta. — Metaphysics Research Lab, Stanford University, 2021.
- [5] *Vapnik, V.N.* An overview of statistical learning theory / V.N. Vapnik // *IEEE Transactions on Neural Networks*. — 1999. — Vol. 10, no. 5. — Pp. 988–999.
- [6] *Pearl, Judea.* Causality / Judea Pearl. — 2 edition. — Cambridge University Press, 2009.
- [7] *Mao, Chengzhi.* Generative Interventions for Causal Learning. — 2021.
- [8] *Wang, Tan.* Causal Attention for Unbiased Visual Recognition. — 2021.
- [9] *Miao, Qiaowei.* Domain Generalization via Contrastive Causal Learning. — 2022.
- [10] *Lv, Fangrui.* Causality Inspired Representation Learning for Domain Generalization. — 2022.
- [11] *Ho, Jonathan.* Denoising Diffusion Probabilistic Models. — 2020.
- [12] *Dhariwal, Prafulla.* Diffusion Models Beat GANs on Image Synthesis. — 2021.
- [13] *Song, Yang.* Score-Based Generative Modeling through Stochastic Differential Equations. — 2021.
- [14] *Spirites, P.* Causation, Prediction, and Search / P. Spirtes, C. Glymour, R. Scheines. — 2nd edition. — MIT press, 2000.
- [15] Learning high-dimensional directed acyclic graphs with latent and selection variables / Diego Colombo, Marloes H. Maathuis, Markus Kalisch, Thomas S. Richardson // *The Annals of Statistics*. — 2012. — . — Vol. 40, no. 1. <http://dx.doi.org/10.1214/11-AOS940>.

- [16] *Claassen, Tom.* Learning Sparse Causal Models is not NP-hard. — 2013.
- [17] *Chen, Wenyu.* Causal Structural Learning Via Local Graphs. — 2021.
- [18] *Bellot, Alexis.* Deconfounded Score Method: Scoring DAGs with Dense Unobserved Confounding. — 2021.
- [19] *Maeda, Takashi Nicholas.* Causal additive models with unobserved variables / Takashi Nicholas Maeda, Shohei Shimizu // Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence / Ed. by Cassio de Campos, Marloes H. Maathuis. — Vol. 161 of *Proceedings of Machine Learning Research*. — PMLR, 2021. — 27–30 Jul. — Pp. 97–106. <https://proceedings.mlr.press/v161/maeda21a.html>
- [20] *Maeda, Takashi Nicholas.* Discovery of Causal Additive Models in the Presence of Unobserved Variables. — 2021.
- [21] *Zheng, Xun.* DAGs with NO TEARS: Continuous Optimization for Structure Learning. — 2018.
- [22] *Bhattacharya, Rohit.* Differentiable Causal Discovery Under Unmeasured Confounding. — 2021.
- [23] *Sanchez, Pedro.* Diffusion Models for Causal Discovery via Topological Ordering. — 2023.
- [24] *Ashman, Matthew.* Causal Reasoning in the Presence of Latent Confounders via Neural ADMG Learning. — 2023.
- [25] *Geffner, Tomas.* Deep End-to-end Causal Inference. — 2022.
- [26] *Sohn, Kihyuk.* Learning Structured Output Representation using Deep Conditional Generative Models / Kihyuk Sohn, Honglak Lee, Xinchen Yan // Advances in Neural Information Processing Systems / Ed. by C. Cortes, N. Lawrence, D. Lee et al. — Vol. 28. — Curran Associates, Inc., 2015. https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf
- [27] *Larsen, Anders Boesen Lindbo.* Autoencoding beyond pixels using a learned similarity metric. — 2016.
- [28] Automatic differentiation in PyTorch / Adam Paszke, Sam Gross, Soumith Chintala et al. — 2017.

- [29] *Bingham, Eli.* Pyro: Deep Universal Probabilistic Programming. — 2018.
- [30] *Axel Sauer, Andreas Geiger.* Counterfactual Generative Networks / Andreas Geiger Axel Sauer // International Conference on Learning Representations (ICLR). — 2021.
- [31] *Deng, Li.* The mnist database of handwritten digit images for machine learning research / Li Deng // *IEEE Signal Processing Magazine*. — 2012. — Vol. 29, no. 6. — Pp. 141–142.
- [32] *Hahsler, Michael.* dbscan: Fast Density-Based Clustering with R / Michael Hahsler, Matthew Piekenbrock, Derek Doran // *Journal of Statistical Software*. — 2019. — Vol. 91, no. 1. — Pp. 1–30.
- [33] *van der Maaten, Laurens.* Visualizing Data using t-SNE / Laurens van der Maaten, Geoffrey Hinton // *Journal of Machine Learning Research*. — 2008. — Vol. 9, no. 86. — Pp. 2579–2605. <http://jmlr.org/papers/v9/vandermaaten08a.html>
- [34] SCM-VAE: Learning Identifiable Causal Representations via Structural Knowledge / Aneesh Komanduri, Yongkai Wu, Wen Huang et al. // 2022 IEEE International Conference on Big Data. — 2022.
- [35] *Zhang, Jiji.* On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias / Jiji Zhang // *Artificial Intelligence*. — 2008. — 11. — Vol. 172.