На правах рукописи

Кузнецова Маргарита Валерьевна

Вариационное моделирование правдоподобия с триплетными ограничениями в задачах информационного поиска

05.13.17 — Теоретические основы информатики

Автореферат
диссертации на соискание учёной степени кандидата технических наук

Москва — 2021

Работа прошла апробацию на кафедре Интеллектуальных систем Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)».

Научный руководитель: **Чехович Юрий Викторович**
кандидат физико-математических наук, Зав. отделом Вычислительного центра, Федеральный исследовательский центр «Информатика и управление» Российской академии наук.

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт системного программирования им. В.П. Иванникова Российской академии наук (ИСП РАН).

Защита состоится 22 ноября 2021 г. в 10:00 на заседании диссертационного совета ФРКТ.05.13.17.003 по адресу: 141701, Московская область, г. Долгопрудный, Институтский переулок, д.9.

С диссертацией можно ознакомиться в библиотеке и на сайте Московского физико-технического института (национального исследовательского университета) https://mipt.ru/education/post-graduate/soiskateli-tekhnicheskie-nauki.php

Работа представлена 31 августа 2021 г. в Аттестационную комиссию федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» для рассмотрения советом по защите диссертаций на соискание ученой степени кандидата наук, доктора наук в соответствии с п.3.1 ст. 4 Федерального закона «О науке и государственной научно-технической политике».

Общая характеристика работы

Актуальность темы. Задача синтеза алгоритмов отображения данных, пришедших из разных источников (доменов), в одно общее скрытое пространство (пространство оценок) (Рудаков К.В.: 1992), крайне актуальна для многих областей машинного обучения, таких как информационный поиск (Li:2019), перевод между доменами (Pham:2019) и генерация объектов (Suzuki:2016).

В настоящей работе рассматриваются задачи, которые характеризуются наличием (возможно потенциальным) данных из различных источников, описывающих некоторое множество объектов. Или, говоря иначе, представляющим разные модальности объектов, составляющим это множество. Примером задачи с данными одинаковой структуры может являться машинный перевод (Koehn:2005) — дана выборка соответствующих предложений на разных языках, нужно построить отображение, переводящее тексты на одном языке в тексты на другом языке.

Также, по одной имеющейся модальности объекта можно осуществлять кросс-доменный поиск и формировать поисковую выдачу, состоящую из соответствующих модальностей другого домена. Примерами такого поиска могут являться тематический кросс-языковой поиск (Wei:2017), или поиск фотографий одной и той же местности с разных ракурсов (Choi:2017).

Настоящая работа посвящена синтезу обучаемых алгоритмов отображения объектов разных доменов в одно общее пространство оценок. В этом пространстве, с помощью полученных скрытых представлений этих объектов, отражающих все модальности, с ними можно будет работать напрямую, решая поставленные задачи и не прибегая к промежуточным этапам, например таким как использование отдельной системы для перевода между доменами. Промежуточные этапы обладают еще одним недостатком — при сведении данных разных доменов к одному, теряются важные характеристики объектов доменов. Другой актуальной задачей является генерация данных (Su:2018), в случае работы с двумя доменами — генерация условно-похожих пар. Таким образом, предлагаемая модель должна быть *генеративной* — исследователь должен иметь возможность генерировать пары из общего пространства, например, для задач пополнения выборок. Одной из наиболее известных генеративных моделей является автокодировщик (Olshausen:1997). В статье (Alain:2014) также было доказано, что автокодировщики с некоторыми видами регуляризации позволяют оценить распределение данных $p(\mathbf{X})$. В работе (Kamyshanska:2013) проводились эксперименты с полученными оценками, в том числе для разных функций активации. Также в качестве генеративных моделей рассматриваются машины Больцмана (Ackley:1985),

глубокие сети доверия (Hinton:2006), генеративные состязательные сети (Goodfellow:2014). Более удобным инструментом для оценки $p(\mathfrak{D})$ и обучения внутренних представлений является вариационный автокодировщик, представленный в (Kingma:2013), с помощью которого можно проще решить все вышеперечисленные задачи, используя вариационные методы (Jordan:1999).

Задача построения совместного отображения также носит название совместного обучения (joint learning) (Suzuki:2016), дистрибутивного обучения (distributed representation learning) (Wei:2017). В ряде работ проводились исследования, позволяющие моделировать совместную плотность вероятности, используя вариационные методы. Основное различие работ заключается в используемых вспомогательных вариационных распределениях. В работе (Weiran:2017) вводятся две модели: VCCA (variational canonical correlation analysis) и ее усовершенствованный вариант — VCCA-private. В работе (Suzuki:2016) вводится JMVAE — Joint Multimodal Variational Autoencoder. В работе (Vedantam:2017) используется TELBO (triple ELBO) — сумма трех нижних вариационных оценок с различными коэффициентами перед слагаемыми ошибки реконструкции и дивергенции Кульбака-Лейблера. В ряде работ для решения требуемых задач используются генеративные состязательные сети (Liu:17). Случай двух модальностей рассмотрен в статье (Wu:2018) авторы использовали вариационный вывод и представили вспомогательное вариационное распределение с помощью техники Product of experts (PoE) (Hinton:2002).

С другой стороны, при решении различных практических задач анализа данных часто приходится сталкиваться с выборками, содержащими ошибки разметки (Angluin:1988). Очень небольшое возмущение в выборке может легко обмануть модель. Ограниченное число работ (Li:2019) рассматривает проблему устойчивости модели в задаче совместного обучения скрытых представлений нескольких доменов. На практике, собрать выборку пар высокого качества сложно и дорого с точки зрения затрат человеческого труда и времени. Из-за непрекращающегося роста данных для таких областей, как обработка естественного языка, компьютерное зрение, речь и т.д. эта задача становится практически невозможной. Таким образом, предлагаемый алгоритм должен быть в состоянии использовать эти объекты для обучения. Более того, алгоритм должен уметь выявлять такие объекты в процессе обучения и использовать полученные знания. Устойчивость использования вариационных методов исследуется в работе (Futami:2018) — авторы рассматривают разные формы дивергенции Кульбака-Лейблера и их эффект на функцию влияния (Hampel:2011).

В данной работе, для создания алгоритма, устойчивого к небольшому количеству выбросов в выборке, предлагается использовать подход из метрического обучения (Bellet:201), а именно так называемых “триплетных

ограничений" (Karaletsos:2015). Использование относительных ограничений как информации о том, что один объект более близок к другому, чем некоторый третий (также распространено название "ложный сосед"), позволяет форсировать обучение модели в верном направлении, при этом однако не накладывая жестких правил на соответствие объектов друг другу. В данной работе идея модифицируется для случая двух доменов — предполагается, что объекты в паре, составленной из разных доменов, близки друг другу, однако на каждой итерации обучения выбирается "ложный сосед" из других объектов доменов, не входящих в пару. Таким образом, формируется триплет. Моделируя правдоподобие такой тройки объектов (объекты в паре и выбранный "ложный сосед") и используя его в основной модели правдоподобия, как компоненту штрафа, можно научить модель разносить объекты в некорректно сопоставленной паре. Если выбранный "ложный сосед" оказывается ближе к объекту в паре, чем назначенный, накладываемый на модель штраф разносит их, не позволяя выучить некорректное соответствие. Комбинация триплетных ограничений с вероятностными моделями значительно повысила качество решения задачи генерации в работе (Karaletsos:2015). Модификация триплетных ограничений для случая двух доменов позволит решить рассматриваемые в данной работе задачи. Так как модель содержит штраф, основанный на триплетных ограничениях, это позволяет снизить влияние ошибок в обучающем наборе данных на итоговое качество решения задач.

Цели работы.

1. Исследовать методы вариационной оценки правдоподобия для случая двух доменов.
2. Предложить и исследовать метод оценки правдоподобия для двух доменов, позволяющий снизить влияние ошибок в обучающем наборе данных на итоговое качество решения задач.
3. Разработать программный комплекс и исследовать с его помощью практическую значимость предложенного метода на примерах модельных и реальных задач информационного поиска.

Методы исследования. Для достижения поставленных целей используется комбинация вероятностных генеративных моделей (Kingma:2013, Suzuki:2016) и метрического обучения (Bellet:2013, Karaletsos:2015). Для получения вариационных нижних оценок логарифма правдоподобия используется вариационный Байесовский вывод (Bishop:2006). Для исследования устойчивости модели к выбросам используются методы робастной статистики (Huber:2004), а именно исследование функции влияния (Hampel:2011) применительно к вариационному выводу (Futami:2018).

Основные положения, выносимые на защиту.

1. Предложена новая модель вариационного автокодировщика, моделирующего совместное правдоподобие данных разных доменов.

2. Предложена модель, основанная на триплетных ограничениях, позволяющая снизить влияние ошибок в обучающем наборе данных на итоговое качество решения задач.
3. Получены оценки логарифмов правдоподобия предлагаемых моделей вариационного автокодировщика.
4. Доказана устойчивость предлагаемого алгоритма.
5. Разработан программный комплекс для решения задач информационного поиска и генерации объектов.
6. Проведены эксперименты на модельных и реальных данных.
7. Результаты работы внедрены в промышленный инструмент поиска кросс-языковых и перефразированных заимствований.

Научная новизна. Разработан подход оценки совместного правдоподобия данных разных доменов. Предложена модель, основанная на триплетных ограничениях, позволяющая повысить качество отображения доменных данных в скрытое пространство и снизить влияние ошибок в обучающем наборе данных на итоговое качество решения задач. Доказана устойчивость предлагаемого алгоритма.

Теоретическая значимость. В данной работе предложенная ранее модель Вариационного автокодировщика обобщается на случай двух доменов. Выводится оценка правдоподобия предложенной модели. Предлагается расширенная модель вариационного автокодировщика для двух доменов, включающая в себя триплетные ограничения. Предлагается метод сэмплирования триплетных ограничений между доменами, что повышает качество решения многих задач. Доказывается устойчивость данной модели по отношению к ошибкам в выборке.

Практическая значимость. Предложенные в работе методы предназначены для решения задач кросс-доменного информационного поиска и генерации объектов. Разработанный программный комплекс предназначен для использования при решении таких задач машинного обучения как: информационный текстовый поиск, информационный поиск по изображениям, междоменный перевод, генерация условно-реальных данных для пополнения обучающих выборок.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных задачах выбора моделей глубокого обучения; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях.

1. “Локальное прогнозирование временных рядов с использованием инвариантных преобразований”, Всероссийская конференция «57-я научная конференция МФТИ», 2014.
2. “A monolingual approach to detection of text reuse in Russian-English collection”, Международная конференция «Artificial Intelligence and Natural Language Conference», 2015.
3. “Machine-Translated Text Detection in a Collection of Russian Scientific Papers”, Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог-21», 2016.
4. “Methods for Intrinsic Plagiarism Detection and Author Diarization”, Международная конференция «Conference and Labs of the Evaluation Forum», 2016.
5. “Детектирование переводных заимствований в текстах научных статей из журналов, входящих в РИНЦ”, Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
6. “Automatic generation of verbatim and paraphrased plagiarism corpus”, Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог-22», 2017.
7. “Evaluation Tracks on Plagiarism Detection Algorithms for the Russian Language”, Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог-22», 2017.
8. “Style Breach Detection with Neural Sentence Embeddings.”, Международная конференция «Conference and Labs of the Evaluation Forum», 2017.
9. “ParaPlagDet: The system of paraphrased plagiarism detection”, Международная конференция «Big Scholar at conference on knowledge discovery and data mining», 2018.
10. “Variational learning across domains with triplet information”, Международная конференция «Bayesian Deep Learning workshop, Conference on Neural Information Processing Systems», 2018.
11. “CrossLang: the system of cross-lingual plagiarism detection”, Международная конференция «Document Intelligence workshop, Conference on Neural Information Processing Systems», 2019.
12. “Вариационное моделирование правдоподобия с триплетными ограничениями в задачах информационного поиска”, Всероссийская конференция «Интеллектуализация обработки информации, ИОИ», 2020.

Публикации по теме диссертации. Основные результаты по теме диссертации изложены в 8 печатных изданиях, 3 из которых изданы в журналах, рекомендованных ВАК.

Личный вклад. Все приведенные результаты, кроме отдельно оговоренных случаев, получены диссертантом лично при научном руководстве к.ф.-м.н. Ю. В. Чеховича.

Структура и объем работы. Диссертация состоит из оглавления, введения, четырех разделов, заключения и списка литературы из 150 наименований. Основной текст занимает 103 страниц.

Основное содержание работы

В **введении** обоснована актуальность диссертационной работы, сформулированы цели и методы исследования, поставлены основные задачи, обоснована научная новизна, теоретическая и практическая значимость полученных результатов.

В **главе 1** вводится формальная постановка задачи, определяются задачи информационного поиска на двух доменах. Вводится понятие генеративной модели и задача информационного поиска формулируется как задача максимизация логарифма правдоподобия выборки.

Пусть заданы множества (домены) $\mathbf{X} = \{\mathbf{x}_a\}_{a=1}^{P} \subset \mathbf{X}_{all}$ и $\mathbf{Y} = \{\mathbf{y}_b\}_{b=1}^{Q} \subset \mathbf{Y}_{all}$ всех доступных объектов одного и того же типа, где $\mathbf{X}_{all}, \mathbf{Y}_{all}$ — все объекты рассматриваемого типа. Пусть также задана выборка $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}, \mathbf{y})_i\}_{i=1}^{N}$, состоящая из N пар объектов (элементов доменов). Под парой здесь понимается биективное отображение: каждому элементу домена $\mathbf{X}$ соответствует один элемент домена $\mathbf{Y}$. Предполагается, что каждая пара $(\mathbf{x}, \mathbf{y})_i$ отражает разные модальности одного и того же объекта.

Будем считать, что на декартовом произведении $\mathbf{X}_{all}$, $\mathbf{Y}_{all}$ определена функция sim_{all}:

$$sim_{all} : (\mathbf{x}_{c_{all}}, \mathbf{y}_{d_{all}})_{\mathbf{x}_{c_{all}} \in \mathbf{X}_{all}, \mathbf{y}_{d_{all}} \in \mathbf{Y}_{all}} \to \{0, 1\}, \tag{1}$$

которая паре $(\mathbf{x}_{c_{all}}, \mathbf{y}_{d_{all}})$ ставит в соответствие 0 или 1 по следующему правилу:

$$sim_{all}(\mathbf{x}_{c_{all}}, \mathbf{y}_{d_{all}}) = \begin{cases} 1, & \text{когда } \mathbf{x}_{c_{all}} \text{ и } \mathbf{y}_{d_{all}} \text{ являются разными модальностями} \\ & \text{одного и того же объекта;} \\ 0, & \text{иначе.} \end{cases} \tag{2}$$

На данной наблюдаемой выборке $(\mathbf{X}, \mathbf{Y})$ задана функция sim с частично известными значениями:

$$sim(\mathbf{x}_c, \mathbf{y}_d) = \begin{cases} 1, & \text{когда } \mathbf{x}_c \text{ и } \mathbf{y}_d \text{ являются разными модальностями} \\ & \text{одного и того же объекта;} \\ 0, & \text{нет информации о соответствии объектов в паре.} \end{cases} \tag{3}$$

Известно, что на некоторой небольшой части пар $\epsilon << N$ sim выдает ошибочное значение, т.е некоторые пары из $(\mathbf{X}, \mathbf{Y})$ сопоставлены некорректно.

Формализуем понятие модели:

Определение 1. *Моделью* $\mathbf{f}(\mathbf{W}, (\mathbf{X}, \mathbf{Y}))$ *назовем функцию вида:*

$$\mathbf{f} : \mathbf{W} \times (\mathbf{X}, \mathbf{Y}) \to sim, \tag{4}$$

где $\mathbf{W}$ *— пространство параметров.*

Можно запросить значения функции sim_{all} на ограниченной выборке пар $\mathfrak{D}_{test} = (\mathbf{X}_{test}, \mathbf{Y}_{test})$, где $|\mathfrak{D}_{test}| << N$.

Требуется найти отображение $\hat{f}$ по выборке $(\mathbf{X}, \mathbf{Y})$, такое, что на $\mathcal{D}_{test}$ достигается минимум некоторой функции потерь l:

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}_{(\mathbf{x}_{c_{test}}, \mathbf{y}_{d_{test}}) \in \mathcal{D}_{test}} l(f(\mathbf{x}_{c_{test}}, \mathbf{y}_{d_{test}}), r), \tag{5}$$

$r \in \{0, 1\}$, $\mathcal{F}$ —заданное семейство моделей.

Связанные подзадачи

Требуется построить алгоритм для:

1. информационного поиска: по заданному $\mathbf{x}$ найти $\mathbf{y}$, удовлетворяющий некоторому критерию, и наоборот (прямой и обратный инф. поиск).

$$f_{search\ direct} : \mathbf{X}_{all} \to \mathbf{Y}, f_{search\ reverse} : \mathbf{Y}_{all} \to \mathbf{X}$$

$$\forall \mathbf{x} \in \mathbf{X}_{all}, \forall \mathbf{y} \in f_{search\ direct}(\mathbf{X}_{all}) : sim_{all}(\mathbf{x}, \mathbf{y}) = 1$$

$$\forall \mathbf{y} \in \mathbf{Y}_{all}, \forall \mathbf{x} \in f_{search\ reverse}(\mathbf{Y}_{all}) : sim_{all}(\mathbf{x}, \mathbf{y}) = 1$$

2. генерации объекта $\hat{\mathbf{y}}$ из по заданному $\mathbf{x}$ и наоборот.

$$f_{gen} : \mathbf{X} \to \hat{\mathbf{Y}} \subseteq \mathbf{Y}_{all} : \forall\ \mathbf{x}\ \in\ \mathbf{X}\ \exists\ \hat{\mathbf{y}}\ \in\ \hat{\mathbf{Y}} : sim_{all}(\mathbf{x}, \hat{\mathbf{y}}) = 1$$

3. двунаправленной генерации условно-реальных пар $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$.

$$f_{bigen} : \mathbf{X}, \mathbf{Y} \to \hat{\mathbf{X}}, \hat{\mathbf{Y}} \subseteq \mathbf{X}_{all}, \mathbf{Y}_{all} : \forall \hat{\mathbf{x}} \in \hat{\mathbf{X}}, \forall \hat{\mathbf{y}} \in \hat{\mathbf{Y}}\ sim_{all}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = 1$$

Одной из постановок задачи информационного поиска при работе с данными из разных доменов является моделирование совместного распределения $p(\mathbf{X}, \mathbf{Y})$. Задачу предлагается решать не в стандартных подходах решения задач классификации, так как метки пар не дают полной информации, тем более, он могут быть ошибочными. Нужно учитывать свойства самих объектов, формирующих пару. Оценка совместного распределения объектов доменов позволяет это сделать, как и решить задачу генерации объектов. Для этого используются *генеративные модели*, введем определение:

Определение 2. *Генеративная модель — модель, восстанавливающая плотность выборки* $(\mathbf{X}, \mathbf{Y})$ *—* $p_\theta(\mathbf{X}, \mathbf{Y})$*, описанной с помощью некоторой параметрической модели с вектором параметром* θ*.*

Задача оценки совместного правдоподобия заключается в нахождении таких параметров $\hat{\theta}$, которые доставляют максимум логарифму правдоподобия выборки:

$$\hat{\theta} = \underset{m\theta \in \mathbf{\Theta}}{\operatorname{argmax}} \log p_\theta(\mathbf{X}, \mathbf{Y}). \tag{6}$$

Для оценки логарифма правдоподобия вводится промежуточное пространство (пространство оценок) $\mathbf{Z} \in \mathbb{R}^n$ для отображения туда объектов из выборки $(\mathbf{X}, \mathbf{Y})$. Задача заключается в подборе таких параметров.

В **главе 2** описываются основные подходы для решения задачи информационного поиска. Описываются подходы к обучению на выборках, содержащих выбросы. Описываются методы оценки правдоподобия. Определяется совместное правдоподобие двух доменов и приводятся его различные оценки. Описывается подход метрического обучения, и чем он может быть полезен в случае задач информационного поиска и наличия выбросов в выборке. Описываются прикладные задачи, возникающие в области информационного поиска.

В **главе 3** рассказывается о построении пространства оценок. Для оценки совместного правдоподобия $p(\mathbf{X}, \mathbf{Y})$ и решения задач информационного поиска и генерации, предлагается использовать вариационные методы, а именно — расширение вариационного автокодировщика, вероятностной генеративной модели, предложенной в (Kingma:2013auto). Сначала приведем принцип его работы для оценки правдоподобия данных одного домена $p(\mathbf{X})$. Задана выборка $\mathbf{X} = \{\mathbf{x}\}_{i=1}^N$, состоящая из N независимых и одинаково распределенных объектов. Вводится скрытое пространство (пространство оценок) $\mathbf{Z} = \mathbb{R}^n$. Предполагается следующий порождающий процесс:

– объекты выборки $\mathbf{X}$ сгенерированы при условии скрытой случайной величины $\mathbf{z} \in \mathbf{Z}$;
– скрытая переменная $\mathbf{z}$ генерируется из многомерного нормального распределения $p_{\theta^*}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$;
– объект выборки $\mathbf{x}$ генерируется из условного распределения $p_{\theta^*}(\mathbf{x}|\mathbf{z})$.

Стоит заметить, что p_{θ^*} и $p_{\theta^*}(\mathbf{x}|\mathbf{z})$ принадлежат параметрическому семейству распределений p_θ и $p_\theta(\mathbf{x}|\mathbf{z})$ и дифференцируемы почти всюду относительно θ и $\mathbf{z}$, однако истинное значение параметров θ^*, как и $\mathbf{z}$ неизвестны. При данных предположениях, распределение данных, задаваемое вариационным автокодировщиком имеет вид (для упрощения нотаций рассматривается один объект выборки $\mathbf{x} \in \mathbf{X}$):

$$p_\theta(\mathbf{x}) = \int_{\mathbf{Z}} p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int_{\mathbf{Z}} p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \tag{7}$$

Логарифм правдоподобия $\log p_\theta(\mathbf{X})$ такой модели является трудно вычислимым, поэтому задача максимизации по параметрам θ не может быть проведена явно. Для взятия данного интеграла применяются вариационные методы. Для получения вариационной оценки на логарифм правдоподобия вводится вспомогательное распределение $q_\phi(\mathbf{z}|\mathbf{x})$ (вариационное распределение).

Определение 3. *Вариационное распределение $q_\phi(\mathbf{z}|\mathbf{x})$ — параметрическое распределение, приближающее истинное апостериорное распределение $p_\theta(\mathbf{z}|\mathbf{x})$.*

Вводится определение Нижней вариационной оценки логарифма правдоподобия:

Определение 4. *Нижняя вариационная оценка логарифма правдоподобия (Evidence Lower Bound, ELBO) — это следующая величина:*

$$\mathcal{L} = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \log \frac{p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \tag{8}$$

где KL — расстояние Кульбака-Лейблера (Kullback:1951) между двумя распределениями. Таким образом, задача взятия интеграла свелась к задаче оптимизации по параметрам θ, ϕ:

$$\mathcal{L} \to \max_{\theta, \phi} \tag{9}$$

Теперь формализуем определение вариационного автокодировщика:

Определение 5. *Вариационный автокодировщик (variational autoencoder, VAE) — генеративная модель, максимизирующая нижнюю вариационную оценку $\mathcal{L}$. Распределения $q_\phi(\mathbf{z}|\mathbf{x})$ и $p_\theta(\mathbf{x}|\mathbf{z})$ моделируются нейронными*

сетями, на выходном слое которых предсказываются параметры этих распределений. Чаще всего в качестве таких распределений используются нормальные распределения:

$$q_\phi(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}^2_\phi) \text{ — } \textit{энкодер},$$

$$p_\theta(\mathbf{x}|\mathbf{z}) \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}^2_\theta) \text{ — } \textit{декодер}.$$

$\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}^2_\phi, \boldsymbol{\mu}_\theta, \boldsymbol{\sigma}^2_\theta$ *— выходы нейронных сетей.*

Дальше в тексте будем использовать обозначение $\mathcal{L}_{VAE}$ для нижней вариационной оценки вариационного автокодировщика.

Далее, перейдем к описанию моделирования и оценки совместного правдоподобия выборки $(\mathbf{X}, \mathbf{Y})$. В большинстве работ, например в (Wei:2017), для совместного правдоподобия модели предполагается, что пары $(\mathbf{x}, \mathbf{y})$ порождены при условии одной скрытой переменной $\mathbf{z}$, причем $p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})$. Правдоподобие представимо в виде:

$$p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \tag{10}$$

Модель носит название *BiVAE* или *JointVAE*, задача оптимизации ставится аналогично (9):

$$\mathcal{L}_{BiVAE} \to \max_{\theta,\phi} \tag{11}$$

Лемма 1. *(Wei:2017) Нижняя вариационная оценка логарифма правдоподобия модели* (10) *представима в виде:*

$$\mathcal{L}_{BiVAE} = -KL(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}) \parallel p(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\big[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log p_\theta(\mathbf{y}|\mathbf{z})\big]. \tag{12}$$

Основная идея введения вспомогательного вариационного распределения вида $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ в том, что одна скрытая переменная $\mathbf{z}$ несет информацию о паре $(\mathbf{x}, \mathbf{y})$, однако данный подход сильно ограничивает количество решаемых задач. Еще одно ограничение заключается в том, что использование одной латентной переменной $\mathbf{z}$ для представления нескольких модальностей зачастую недостаточно, так как из за этого может происходить потеря информации о важных характеристиках отдельных модальностей. Для снятия ограничений, описанных выше, рассмотрим различные виды порождающих процессов и определим правдоподобие модели наиболее удобным образом. Полученную модель назовем Variational Bidomain Autoencoder (VBA).

Сумма оценок по двум вспомогательным вариационным распределениям

Рассмотрим порождающий процесс с двумя вспомогательными вариационными распределениями $q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})$ и $q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})$. Процесс порождения данных выглядит следующим образом:

– $\mathbf{z}_x$ и $\mathbf{z}_y$ генерируем из априорного распределения $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$,
– $\mathbf{x}$ генерируем из апостериорного распределения $p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}_x)$,
– $\mathbf{y}$ генерируем из апостериорного распределения $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}_y)$.

Правдоподобие можно определить следующим образом:

$$p(\mathbf{x}, \mathbf{y}) = \int_{\mathbf{z}_x} \int_{\mathbf{z}_y} p_{\theta_x}(\mathbf{x}|\mathbf{z}_x) p_{\theta_y}(\mathbf{y}|\mathbf{z}_y) d\mathbf{z}_x d\mathbf{z}_y, \tag{13}$$

где $\mathbf{z}_x$, $\mathbf{z}_y$ — скрытые представления пары $(\mathbf{x}, \mathbf{y})$. Ставится задача максимизации логарифма правдоподобия:

$$\mathcal{L}_{VBA}(\mathbf{x}, \mathbf{y}) \to \max_{\theta_{\mathbf{x}}, \theta_{\mathbf{y}}, \phi_{\mathbf{x}}, \phi_{\mathbf{y}}}, \tag{14}$$

Так как вводятся два вспомогательных вариационных распределения, то итоговая оценка будет являться суммой двух оценок — по вариационным распределениям $q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})$ и $q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})$.

Теорема 1.

$$\mathcal{L}_{VBA} = \mathcal{L}_{VBA_1} + \mathcal{L}_{VBA_2} = \mathbb{E}_{q_{\phi_x}(\mathbf{z}_x|\mathbf{x})} \log \frac{p_\theta(\mathbf{x}, \mathbf{y})}{q_{\phi_x}(\mathbf{z}_x|\mathbf{x})} + \mathbb{E}_{q_{\phi_y}(\mathbf{z}_y|\mathbf{y})} \log \frac{p_\theta(\mathbf{x}, \mathbf{y})}{q_{\phi_y}(\mathbf{z}_y|\mathbf{y})} =$$

$$= \underbrace{\Big[-KL\big(q_{\phi_{\mathbf{x}}(\mathbf{z}_x|\mathbf{x})}(\mathbf{z}_x|\mathbf{x}) \parallel p(\mathbf{z}_x)\big) - KL\big(q_{\phi_{\mathbf{y}}(\mathbf{z}_y|\mathbf{y})}(\mathbf{z}_y|\mathbf{y}) \parallel p(\mathbf{z}_y)\big) \Big]}_{\textit{Штраф}} +$$

$$+ \underbrace{\Big[\mathbb{E}_{q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})} \big[\log p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}_x)\big] + \mathbb{E}_{q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})} \big[\log p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}_y)\big] \Big]}_{\textit{Ошибка реконструкции}} +$$

$$+ \underbrace{\Big[\mathbb{E}_{q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})} \big[\log p_{\theta_{\mathbf{x}}}(\mathbf{y}|\mathbf{z}_x)\big] + \mathbb{E}_{q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})} \big[\log p_{\theta_{\mathbf{y}}}(\mathbf{x}|\mathbf{z}_y)\big] \Big]}_{\textit{Ошибка взаимного соответствия}}$$

Другой спрсоб — использование факторизованного вариационного распределения. Введем вариационное распределение $q_\phi(\mathbf{z}_{x,y}|\mathbf{x}, \mathbf{y})$ и его факторизацию:

$$q_\phi(\mathbf{z}_{x,y}|\mathbf{x}, \mathbf{y}) = q_{\phi_{\mathbf{z}_x}}(\mathbf{z}_x|\mathbf{x}) q_{\phi_{\mathbf{z}_y}}(\mathbf{z}_y|\mathbf{y}), \tag{15}$$

где $\mathbf{z}_{x,y}$ обозначает пару объектов $\{\mathbf{z}_x, \mathbf{z}_y\}$. Правдоподобие определяется следующим образом:

$$p(\mathbf{x}, \mathbf{y}) = \int_{\mathbf{z}_x} \int_{\mathbf{z}_y} p_{\theta_x}(\mathbf{x}|\mathbf{z}_x) p_{\theta_y}(\mathbf{y}|\mathbf{z}_y) d\mathbf{z}_x d\mathbf{z}_y, \tag{16}$$

Теорема 2. *При факторизации* (15) *нижняя вариационная оценка выглядит следующим образом:*

$$\mathcal{L}^*_{VBA} = \mathbb{E}_{q_\phi(\mathbf{z}_{x,y}|\mathbf{x},\mathbf{y})} \log \frac{p_\theta(\mathbf{x},\mathbf{y})}{q_\phi(\mathbf{z}_x,\mathbf{z}_y|\mathbf{x},\mathbf{y})} =$$

$$= \underbrace{-KL\big(q_\phi(\mathbf{z}_{x,y}|\mathbf{x},\mathbf{y}) \parallel p(\mathbf{z})\big)}_{\textit{Штраф}} + \underbrace{\mathbb{E}_{q_{\phi_\mathbf{x}}(\mathbf{z}_x|\mathbf{x})}\big[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}_x)\big] + \mathbb{E}_{q_{\phi_\mathbf{y}}(\mathbf{z}_y|\mathbf{y})}\big[\log p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z}_y)\big]}_{\textit{Ошибка реконструкции}}$$

Заметим, что при оптимизации логарифма правдоподобия, введенного в (13) и в (16) возникают сложности, если выборка $(\mathbf{X},\mathbf{Y})$ содержит некоторый уровень шума, а именно — некоторого количества пар ϵ, где объекты в паре сопоставлены некорректно. Введем определение "шумной" пары, т.е пары, содержащей некорректные соответствия.

Определение 6. *Назовем "шумной" пару* $(\mathbf{x},\mathbf{y})$ *такую, что* $d_{all}(\mathbf{x},\mathbf{y}) \geq \delta$, *где* d_{all} *— некоторая функция расстояния, определенная между элементами разных доменов,* δ *— порог.*

Нужно внести в модель штраф за несоответствие пар. Причем, штрафовать в процессе обучения модель должна сама себя на основе выявленных несоответствий объектов в паре. Для работы с "шумными" парами выборки $(\mathbf{X},\mathbf{Y})$ предлагается использовать подход, основная идея которого идет из метрического обучения и триплетных ограничений. Для рассматриваемой в данной работе задачи наибольший интерес представляет множество $\mathcal{T}$ — использование относительных ограничений как информации о том, что один объект более близок к другому, чем некоторый третий (также распространено название "ложный сосед"). Использование такого рода ограничений позволяет форсировать обучение модели в верном направлении, при этом однако не накладывая жестких правил на соответствие объектов друг другу. Во многих задачах бывает сложно, а зачастую и невозможно, корректно формализовать множества $\mathcal{S}$ и $\mathcal{D}$. Получение этих множеств связано также с большими трудозатратами на разметку похожих и непохожих пар, в то время как третье множество требует только введенной функции расстояния.

Адаптируем эту идею для рассматриваемой задачи — будем предполагать, что объекты в паре $(\mathbf{x},\mathbf{y})_i$ близки друг другу, однако на каждой итерации обучения будем выбирать для этих объектов "ложного соседа" поочередно из других объектов доменов $\mathbf{X}$ и $\mathbf{Y}$, формируя таким образом тройку объектов. Моделируя правдоподобие такой тройки объектов и используя его в основной модели, как компоненту штрафа, можно научить модель разносить объекты в некорректно сопоставленной паре. Если выбранный "ложный сосед" оказывается ближе к объекту в паре, чем назначенный, накладываемый на модель штраф разносит их, не позволяя выучить некорректное соответствие. Также, выбор третьего объекта

позволяет решить другую озвученную выше особенность в данных — соответствие объектов как многие ко многим. Таким образом, подход, предлагаемый в данной работе, позволит моделировать совместное правдоподобие данных разных доменов. Так как модель содержит штраф, основанный на триплетных ограничениях, это позволяет снизить влияние ошибок в обучающем наборе данных на итоговое качество решения задач. Выбор "ложного соседа" будет производиться в пространстве оценок $\mathbf{Z}$ в процессе обучения модели. На каждой итерации "ложный сосед" выбирается, исходя из текущих параметров модели.

Основываясь на работе (Karaletsos:2015) определим верный триплет в пространстве оценок $\mathbf{Z}$ следующим образом:

Определение 7. *Верный триплет — такой триплет, в котором:*

$$\mathcal{T} = \{t_{x,y,k}\} = \{(\mathbf{z}_x,\mathbf{z}_y,\mathbf{z}_k) : d(\mathbf{z}_x, \mathbf{z}_y) < d(\mathbf{z}_x, \mathbf{z}_k)\}, \tag{17}$$

где

- d — *некоторая функция близости;*
- $\mathbf{z}_k \in \{\mathbf{Z}_X \cup \mathbf{Z}_Y\}$, $\mathbf{Z}_X$ *и* $\mathbf{Z}_Y$ — *скрытые представления всех объектов выборки* $(\mathbf{X}, \mathbf{Y})$*;*
- $\mathbf{z}_k$ — *некоторое преобразование объекта* k*;*
- $\mathbf{z}_x$, $\mathbf{z}_y$ — *скрытые представления объектов пары* $(\mathbf{x},\mathbf{y})$.

На практике, выбор $\mathbf{z}_k$ происходит с помощью процедуры сэмплирования из скрытых представлений объектов доменов $\mathbf{X}$ и $\mathbf{Y}$. На каждом шаге оптимизационного процесса поочередно делается шаг 1 или шаг 2:

1. Выбирается ближайший объект к $\mathbf{z}_x$ из скрытых представлений данных домена $\mathbf{Y}$:
   $\mathbf{z}_k = \operatorname{argmin}_{\mathbf{z}_{y'} \in \mathbf{Z}'_\mathbf{Y} \setminus \mathbf{z}_y} d(\mathbf{z}_x, \mathbf{z}_{y'})$,
2. Выбирается ближайший объект к $\mathbf{z}_y$ из скрытых представлений данных домена $\mathbf{X}$:
   $\mathbf{z}_k = \operatorname{argmin}_{\mathbf{z}_{x'} \in \mathbf{Z}'_\mathbf{X} \setminus \mathbf{z}_x} d(\mathbf{z}_y, \mathbf{z}_{x'})$,

где $\mathbf{Z}'_\mathbf{X} \subset \mathbf{Z}_\mathbf{X}, \mathbf{Z}'_\mathbf{Y} \subset \mathbf{Z}_\mathbf{Y}$ содержат данные из $\mathbf{X}, \mathbf{Y}$ (минибатчи), d — функция расстояния, в данном случае l_2 норма. Выбор третьего объекта особенно важен при работе с выборками, содержащими шум. Выбор ложного соседа в этом случае позволяет разделить объекты в паре, где они сопоставлены некорректно, с помощью моделирования триплетного ограничения. Определим его, основываясь на работе (Karaletsos:2015), однако важно заметить, что мы расширим это определение на случай двух доменов:

$$p(t_{x,y,k}) = \int_\mathbf{z} p(\mathbf{t}_{x,y,k}|\mathbf{z}_x,\mathbf{z}_y,\mathbf{z}_k)p(\mathbf{z}_x)p(\mathbf{z}_y)p(\mathbf{z}_k)d\mathbf{z}_x d\mathbf{z}_y d\mathbf{z}_k, \tag{18}$$

Функция правдоподобия моделируется распределением Бернулли, параметризованным функцией softmax:

$$p(t_{x,y,k}) = Ber(t_{x,y,k}) = \frac{e^{-d(\mathbf{z}_x,\mathbf{z}_y)}}{e^{-d(\mathbf{z}_x,\mathbf{z}_y)} + e^{-d(\mathbf{z}_x,\mathbf{z}_k)}}. \tag{19}$$

$$p(t_{x,y,k}) = \begin{cases} 1, & \text{если } Ber(t_{x,y,k} \in \mathcal{T}|x,y,k) \geq \frac{1}{2}; \\ Ber(t_{x,y,k} \in \mathcal{T}|x,y,k), & \text{если } Ber(t_{x,y,k} \in \mathcal{T}|x,y,k) < \frac{1}{2}. \end{cases} \tag{20}$$

При появлении некорректной пары в выборке, условие, определенное в (20) наложит штраф на модель — $Ber(t_{x,y,k} \in \mathcal{T}|x,y,k)$ (в дальнейшем будем использовать обозначение). Еще один способ определить триплетное правдоподобие представлен в работе (Van:2012), где сравнивается правдоподобие вида (19) и правдоподобие, представленное с помощью распределения Стьюдента с α степенями свободы:

$$p(t_{x,y,k}) = \frac{\left(1 + \frac{\|\mathbf{z}_x - \mathbf{z}_y\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{\|\mathbf{z}_x - \mathbf{z}_y\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{\|\mathbf{z}_x - \mathbf{z}_k\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}} \tag{21}$$

Рассмотрим вариационную оценку для модели VBTA с условием триплетного правдоподобия. Для выборки $\{(\mathbf{X}, \mathbf{Y})\}$ можно определить правдоподобие модели с триплетными ограничениями:

$$p(\mathbf{x},\mathbf{y},\mathbf{t}) = \int_{\mathbf{z}} p(\mathbf{x},\mathbf{y},\mathbf{t},\mathbf{z}_x,\mathbf{z}_y,\mathbf{z}_k) = p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}_x) p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}_y) p(t_{x,y,k}|\mathbf{z}_x,\mathbf{z}_y,\mathbf{z}_k) p(\mathbf{z}) d\mathbf{z}. \tag{22}$$

Представленную модель назовем *VBTA* — Вариационный автокодировщик с триплетными ограничениями (Variational Bi-domain Autoencoder). В дальнейшем, для упрощения нотаций будем использовать обозначение $\mathbf{z}_{x,y,k}$ для тройки объектов $\{\mathbf{z}_x, \mathbf{z}_y, \mathbf{z}_k\}$. Напомним, что объект $\mathbf{z}_k$ выбирается из всех скрытых представлений данных домена $\mathbf{Y}$ или $\mathbf{X}$. Верна следующая теорема:

Теорема 3. *Нижняя вариационная оценка логарифма правдоподобия Вариационного автокодировщика с триплетными ограничениями (VBTA) с*

использованием вариационного распределения вида (15) *представим в виде:*

$$\mathcal{L}_{VBTA} = \underbrace{-KL\big(q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x}) \parallel p(\mathbf{z}_x)\big) - KL\big(q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y}) \parallel p(\mathbf{z}_y)\big)}_{Штраф} + \\ + \underbrace{\mathbb{E}_{q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})}\big[\log p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}_x)\big] + \mathbb{E}_{q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})}\big[\log p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z}_y)\big]}_{Ошибка\ реконструкции} + \\ \underbrace{\mathbb{E}_{\mathbf{z}_x,\mathbf{z}_k\sim q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})}\mathbb{E}_{\mathbf{z}_y\sim q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})}\log p(t_{x,y,k}|\mathbf{z}_{x,y,k}))}_{Ошибка\ триплетов} + \\ + \underbrace{\mathbb{E}_{\mathbf{z}_x\sim q_{\phi_{\mathbf{x}}}(\mathbf{z}_x|\mathbf{x})}\mathbb{E}_{\mathbf{z}_y,\mathbf{z}_k\sim q_{\phi_{\mathbf{y}}}(\mathbf{z}_y|\mathbf{y})}\log p(t_{x,y,k}|\mathbf{z}_{x,y,k}))}_{Ошибка\ триплетов} \quad (23)$$

Далее приводится теоретический анализ предложенного метода с помощью функции влияния (Hampel:2011). В статье (Futami:2018) данный подход был применен для исследования устойчивости вариационного автокодировщика в задачах классификации и регрессии. Введем основные определение функции влияния: она отражает влияние уровня шума ϵ в выборке $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ на некоторую статистику T. Для этого вводится эмпирическая (выборочная) функция распределения выборки $\mathbf{X}$: $G(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\{\mathbf{x}_i \leqslant \mathbf{x}\}}$, где $\mathbf{1}_{\{\mathbf{x}_i \leqslant \mathbf{x}\}}$ — индикаторная функция. При условии уровня шума ϵ, можно также ввести "зашумленную" версию эмпирической функции распределения в точке $\mathbf{x}^o$; в дальнейшем для удобства обозначений вместо $\mathbf{1}_{\{\mathbf{x}_o \leqslant \mathbf{x}\}}$ введем обозначение $\triangle_{\mathbf{x}^o}\{\mathbf{x}\}$ (единичная масса события):

$$G_{\epsilon,\mathbf{x}^o} = (1-\epsilon)G(\mathbf{x}) + \epsilon \triangle_{\mathbf{x}^o} \mathbf{x}, \quad (24)$$

Тогда, для некоторой статистики T и эмпирического распределения G функция влияния определяется как:

$$IF(\mathbf{x}_o, T, G) = \frac{\partial}{\partial\epsilon} T(G_{\epsilon,\mathbf{x}_o}(\mathbf{x}))\Big|_{\epsilon=0} = \lim_{\epsilon\to 0}\frac{T(G_{\epsilon,\mathbf{x}_o}(\mathbf{x})) - T(G(\mathbf{x}))}{\epsilon} \quad (25)$$

На практике, анализируют $sup_{\mathbf{x}_o}|IF(\mathbf{x}_o, T, G)|$, если он ограничен, модель не чувствительна к небольшому шуму в выборке. Выведем функцию влияния для VBTA-модели по аналогии с выводом (Futami:2018):

Теорема 4. *Пусть "зашумленная" версия эмпирической функции распределения:* $G_\epsilon(\mathbf{x},\mathbf{y}) = (1-\epsilon)G(\mathbf{x},\mathbf{y}) + \epsilon \triangle_{(\mathbf{x},\mathbf{y})^o}(\mathbf{x},\mathbf{y})$*, тогда функция влияния VBTA-модели определяется как:*

$$\Big(\frac{\partial^2 \mathcal{L}_{VBTA}}{\partial\phi^2}\Big)^{-1}\nabla_\phi \mathbb{E}_{q_\phi}\Big[KL\Big(q_\phi \parallel p(\mathbf{z})\Big) + N\log(\mathbf{x}^o, \mathbf{y}^o, \mathbf{t}^o|\mathbf{z})\Big],$$

где $\mathbf{t}^o$ *— "шумный" триплет, тройка объектов, содержащих "шумную" пару, введенную в определении* (6)*.*

Для анализа устойчивости проводят анализ $\log(\mathbf{x}^o, \mathbf{y}^o, \mathbf{t}^o|\mathbf{z})$, т.к. именно это слагаемое связано с выбросами. Рассмотрим поведение $sup_{(\mathbf{x},\mathbf{y})^o}|IF((\mathbf{x},\mathbf{y})^o, T, G)|$, где $(\mathbf{x},\mathbf{y})^o$ — "шумная" пара, в которой один из объектов пары, или оба, являются выбросами. Как видно из вывода функции влияния, основное слагаемое для анализа — $\log(\mathbf{x}^o, \mathbf{y}^o, \mathbf{t}^o|\mathbf{z})$, которое распадается на две ошибки реконструкции и триплетное правдоподобие. Существует ряд задач, где нужно классифицировать объект(ы) с помощью обученных внутренних представлений, например кросс-языковая классификация документов, поиск переводных заимствований. В таких задачах ошибка реконструкции играет несущественную роль в качестве классификации и ее оптимизация не представляет интереса. Слагаемое, отвечающее за триплетное правдоподобие, играет основную роль в таких задачах, в том числе в основном оно влияет на устойчивость возможным выбросам, как в обучающей, так и в тестовой выборке. Поэтому наиболее важно провести анализ для $\log p(\mathbf{t}_{x,y,k})$. Анализ проводится для правдоподобия, определенного в (21) — авторами (Van:2012) было отмечено, что использование правдоподобия вида (19) значения вероятности в случае неверного триплета слишком быстро становится очень малым и сильные нарушения не приводят к значительным штрафам, в то время как правдоподобие вида (21) не подвержено такому недостатку. Сформулируем следующую теорему:

Теорема 5. *Если триплетное правдоподобие задано в виде:*

$$p(t_{x,y,k}) = \frac{\left(1 + \frac{\|\mathbf{z}_x - \mathbf{z}_y\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{\|\mathbf{z}_x - \mathbf{z}_y\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{\|\mathbf{z}_x - \mathbf{z}_k\|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}, \tag{26}$$

то $sup_{(\mathbf{x},\mathbf{y})^o}|IF((\mathbf{x},\mathbf{y})^o, T, G)|$ для VBTA-модели ограничен при условии того, что оптимизация ошибки реконструкции не проводится. Следовательно, VBTA-модель является устойчивой к шуму в выборке.

Что доказывает теоретическую устойчивость предлагаемого метода.

В **главе 4** на базе предложенных алгоритмов описывается разработанный программный комплекс, на базе которого решаются задачи информационного поиска, описанные в главе 1. Результаты сравниваются с результатами известных алгоритмов. Также приводятся результаты работы промышленного модуля поиска переводных и перефразированных заимствований.

В заключении представлены основные результаты диссертационной работы.

1. Предложена новая модель вариационного автокодировщика, моделирующего совместное правдоподобие данных разных доменов.

2. Предложена модель, основанная на триплетных ограничениях, позволяющая снизить влияние ошибок в обучающем наборе данных на итоговое качество решения задач.
3. Получены оценки логарифмов правдоподобия предлагаемых моделей вариационного автокодировщика.
4. Доказана устойчивость предлагаемого алгоритма.
5. Разработан программный комплекс для решения задач информационного поиска и генерации объектов.
6. Проведены эксперименты на модельных и реальных данных.
7. Результаты работы внедрены в промышленный инструмент поиска кросс-языковых и перефразированных заимствований.

Публикации соискателя по теме диссертации

Публикации в журналах из списка ВАК и SCOPUS.

1. Bakhteev, O., Kuznetsova, R., Romanov, A. and Khritankov, A., 2015, November. A monolingual approach to detection of text reuse in Russian-English collection. In 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT) (pp. 3-10). IEEE.
2. Romanov, A., Kuznetsova, R., Bakhteev, O. and Khritankov, A., 2016. Machine-Translated Text Detection in a Collection of Russian Scientific Papers. Computational Linguistics and Intellectual Technologies.
3. Kuznetsov, M. P., Motrenko, A., Kuznetsova, R., Strijov, V. V., 2016. Methods for Intrinsic Plagiarism Detection and Author Diarization. In CLEF (Working Notes) (pp. 912-919).
4. Кузнецова М. В., Стрижов В. В. Локальное прогнозирование временных рядов с использованием инвариантных преобразований // Информационные технологии. 2016. Т. 22. №. 6. С. 457.
5. Safin, K., Kuznetsova, R., 2017. Style Breach Detection with Neural Sentence Embeddings. In CLEF (Working Notes).
6. Smirnov, I., Kuznetsova, R., Kopotev, M., Khazov, A., Lyashevskaya, O., Ivanova, L., Kutuzov, A., 2017. Evaluation tracks on plagiarism detection algorithms for the russian language. Computational Linguistics and Intellectual Technologies.
7. Сафин К. Ф., Кузнецов М. П., Кузнецова М. В. Определение заимствований в тексте без указания источника //Информатика и её применения, 2017, Т. 11, №. 3, С. 73-79.
8. Р.В. Кузнецова, О.Ю. Бахтеев, Ю.В. Чехович, Детектирование переводных заимствований в больших массивах научных документов, //Информатика и её применения, 2021, Т. 15, №. 1, С. 30-42.