МРНТИ 28.23.17

Ж.М. Кожирбаев, Ж.А. Есенбаев

National Laboratory Astana, Астана, Казахстан
(E-mail: zhanibek.kozhirbayev@nu.edu.kz, zhyessenbayev@nu.edu.kz)

Топологический анализ невыровненных аудио- и текстовых данных

Аннотация. Авторами выполнена предварительная работа по топологическому анализу аудио- и текстовых данных для неконтролируемой обработки речи. Работа основана на предположении, что частоты фонем и контекстуальные отношения схожи в акустической и текстовой областях одного и того же языка. Соответственно, это позволило создать отображение между этими пространствами, учитывающее их геометрическую структуру. В качестве первого шага были выбраны генеративные методы, основанные на вариационных автоэнкодерах, для отображения аудио- и текстовых данных в два скрытых векторных пространства. На следующем этапе методы персистентной гомологии используются для анализа топологической структуры двух пространств. Хотя полученные результаты подтверждают идею сходства двух пространств, необходимы дальнейшие исследования для корректного отображения акустического и текстового пространств, а также для оценки реального эффекта от включения топологической информации в процесс обучения автоэнкодера

Ключевые слова: неконтролируемая обработка речи, вариационные автокодировщики, встраивание слов, топологический анализ данных, постоянная гомология и диаграммы.

DOI: doi.org/10.32523/2616-7263-2022-141-4-116-126

Введение

Топологический анализ данных (ТДА) [1, 2] — относительно новая область интеллектуальной обработки неструктурированных данных. Он появился в результате компьютерной реализации методов алгебраической топологии [3]. Эвристически идея состоит в том, чтобы построить «форму» для «облака точек» в пространствах не слишком высокой размерности. Формы представляют собой симплициальные комплексы, т. е. множества, образованные вершинами, ребрами, гранями, тетраэдрами и т.д., полученные в результате процедуры фильтрации: покрытия облака точек шарами переменного радиуса или просто шкалы близости в соответствующей метрике. Синхронное увеличение радиусов шаров приводит к их пересечениям, которые кодируются элементами комплекса: ребрами и гранями. Комплекс позволяет вычислять топологические инварианты — так называемые числа Бетти, которые, грубо говоря, измеряют количество k-мерных дырок. Однако топология, получаемая с помощью покрытий, существенно зависит от выбранного пространственного масштаба. Первоначальное решение заключалось в одновременной оценке инвариантов по всем шкалам. Связи между инвариантами на разных масштабах кодируются их временем жизни в процессе фильтрации в виде диаграммы персистентности [4]. Грубо говоря, постоянство — это метафора времени жизни, выраженная в единицах шкалы расстояний, топологический паттерн, например, появление цикла и превращение его в грань. Были получены нетривиальные теоремы об устойчивости персистентных диаграмм по отношению к возмущениям [5], что сделало вычислительную топологию универсальным неметрическим аппаратом для анализа различных объектов [5, 6, 7, 8, 9].

Методология

Две точки пространства признаков соединены ребром, если расстояние между ними в подходящей метрике не превосходит заданного малого числа ε. Говоря формальным языком, речь идет о понятии так называемой ε-связной цепи Кантора. Так называется последовательность неэквидистантных точек, в которой количество линейно связанных компонент зависит от выбранного пространственного разрешения. Другими словами, мы объединяем две составляющие в одну, если наше зрение не разделяет их отдельно по шкале $\leq \varepsilon$.

Цепная близость приводит к понятию нерва топологического покрытия множества точек. Пусть $S = \{v_i\}_{i=1}^{N} a$ — конечное множество точек из R^2. Украсьте каждую точку диском $B(v_i, \varepsilon)$ с центром в v_i и радиусом ε. Напомним, что структура, состоящая из простейших (simplicissima) элементов - вершин, ребер и треугольных граней, называется симплициальным комплексом, если ее соседние элементы пересекаются в точке или имеют общее ребро. Одновременно будем увеличивать радиусы дисков. Пересечение полученных элементов растяжения приводит к симплициальному комплексу, который называется комплексом Чеха:

$$K(S, \varepsilon) = \bigcap_{i=1}^{N} B(v_i, \varepsilon) \quad (1)$$

Чтобы получить собственно нерв, мы соединяем соседние точки ребром, когда соответствующие им соседние диски пересекаются. Кроме того, условимся, что пересечение трех соседних дисков порождает грань, т.е. «заштрихованный» треугольник. С увеличением радиусов комплекс структурно упрощается и превращается в одну «заштрихованную» грань (Рисунок 1). Этот процесс в алгебраической топологии называется фильтрацией [10]. Напомним, что раздувание или расширение диска - одна из основных операций в математической морфологии - сложение по Минковскому, которое принято обозначать как B⊚ε [7]. Как и прежде, украсим каждую точку множества диском S радиуса ⊚. Объединение этих дисков называется параллельным телом S⊚ε для S или накрытием Минковского. Теорема о нервах утверждает, что комплекс Чеха (1) для множества S и его покрытие Минковского S⊚ε гомотопны друг другу [10]:

$$K(S, \varepsilon) \sim S \oplus \varepsilon \quad (2)$$

Другими словами, объединение кругов покрытия Минковского можно «сжать» непрерывной деформацией до симплициального комплекса Чеха: заметим, что гомотопия допускает “слипание” точек. С другой стороны, теорема позволяет сравнивать два пространства с точностью до их аппроксимации комбинаторными комплексами, нагруженными целочисленными топологическими инвариантами. Отметим, что существенными элементами комплекса являются число отдельных ε-различимых компонент, измеряемое так называемым числом Бетти β_0, и число независимых классов незаполненных дыр, т. е. циклов, образованных замкнутой цепочкой ребер. Их количество измеряется числом Бетти β_1.

С точки зрения распознавания образов, при построении фильтрации мы меняем наш допуск определения взаимосвязи между признаками от наименьшего, когда соседние точки не пересекаются со своими окрестностями - дисками, до наибольшего, когда точка родственна или подобна всем точкам в пространстве признаков, т.е. включенным в единый торец. Очевидно, что в этой процедуре времена жизни каждой точки до ее включения в ребро и каждого отверстия до ее исчезновения (затенения) различны. Продолжительность жизни или постоянство можно измерить интервалом изменения радиусов дилатации от рождения до разрушения ребра или отверстия. Такие отличия изображаются набором горизонтальных отрезков - штрих-кодов, параллельных оси изменения радиуса (Рисунок 1).

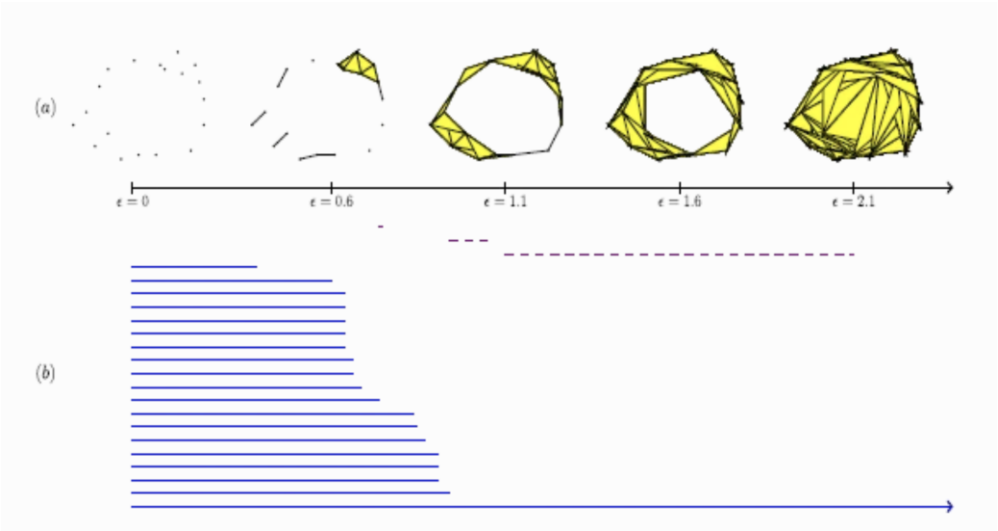

Рисунок 1. Пример (а) фильтрация (симплициальные комплексы для различных ϵ) и (б) персистентные штрих-коды, построенные на облаке точек

Штрих-коды удобнее представлять в виде персистентной диаграммы, облака точек на плоскости, координатами каждой из которых являются начало и конец штрих-кода (Рисунок 2). Все точки, естественно, лежат выше диагонали, что соответствует нулевому времени жизни [4]. Все вышеперечисленные техники относятся к вычислению персистентных гомологий методами вычислительной топологии. Так называется компьютерная версия алгебраической топологии — науки, активно развивающейся в последние годы [6, 11]. В следующем разделе мы даем более формальное определение персистентных гомологий и диаграмм.

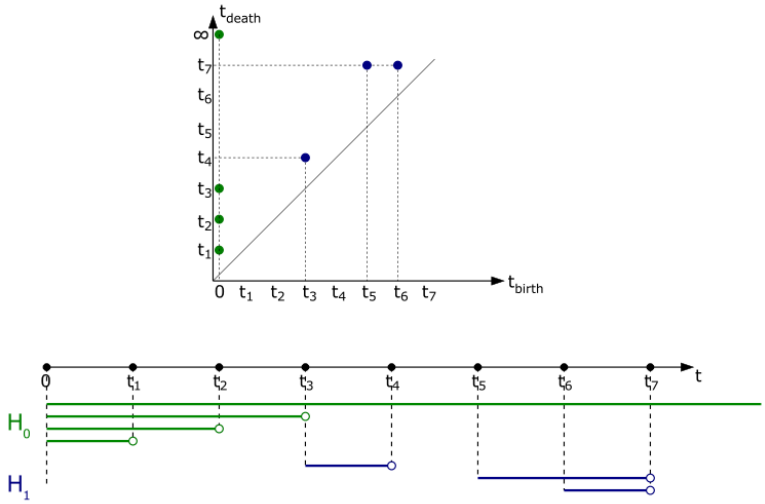

Рисунок 2. Связь между постоянными штрих-кодами и диаграммами

Рассмотрим симплициальный комплекс K и функцию $f: K \to R$ [5]. Мы хотим, чтобы f была монотонной, что означает, что она не убывает вдоль возрастающих цепочек граней, т. е. $f(\sigma) \leq f(\tau)$, если σ — грань τ. Монотонность означает, что множество подуровней $K(a) = f^{-1}(-\infty, a]$ является подкомплексом K для любого $a \in R$. Пусть m будет количеством симплексов в K, мы получим $n + 1 \leq m + 1$ различных подкомплексов, которые упорядочим в виде возрастающей вложенной последовательности комплексов: $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$.

Другими словами, если $a_0 < a_1 < \cdots < a_n$ — значения функций симплексов в K и $a_0 = -\infty$, то $K_i = K(a_i)$ для каждого i. Эту последовательность комплексов мы называем фильтрацией f и воспринимаем ее как конструкцию, добавляющую фрагменты симплексов в процессе их

агрегации в покрывающий комплекс. Мы уже видели примеры фильтрации раньше, а именно, комплекс Чеха. Нас больше интересует топологическая эволюция, выражаемая соответствующей последовательностью групп гомологий. *Для* каждого $i \leq j$ у нас есть отображение включения из основного пространства K_i в пространство K_j и, следовательно, индуцированный гомоморфизм $f_p^{i,j}: H_p(K_i) \to H_p(K_j)$ *для* каждой размерности p. Таким образом, фильтрации соответствует последовательность групп гомологий, связанных гомоморфизмами:

$$0 = H_p(K_0) \to H_p(K_1) \to \cdots \to H_p(K_n) = H_p(K) \tag{3}$$

то же самое *для* каждого измерения p. При переходе от K_{i-1} к K_i мы можем получить новые классы гомологии и можем потерять некоторые из них, когда они станут тривиальными или сольются друг с другом. Мы собираем классы, которые рождаются при или до определенного порога и умирают после другого порога в группы.

Определение. p-е персистентные группы гомологий — это образы гомоморфизмов, индуцированных включением $H_p^{i,j} = imf_p^{i,j}$ *для* $0 \leq i \leq j \leq n$. Соответствующие p-е персистентные числа Бетти являются рангами этих групп, $\beta_p^{i,j} = \operatorname{rank} H_p^{i,j}$. Точно так же мы определяем редуцированные персистентные группы гомологии и редуцированные персистентные числа Бетти. Обратите внимание, что $H_p^{i,j} = H_p(K_i)$. Постоянные группы гомологии состоят из классов гомологии K_i, которые все еще «живы» в K_j, или, более формально, $H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i))$. У нас есть такая группа *для* каждой размерности p и каждой пары индексов $i \leq j$. Мы можем быть более конкретными в отношении классов, учитываемых стойкими группами гомологии. Пусть γ — класс в $H_p(K_i)$, мы говорим, что он родился в K_i, если $\gamma \notin H_p^{i-1,i}$. *Далее*, если γ рождается в K_i, то умирает, входя в K_j, если сливается со старшим классом при переходе из K_{j-1} в K_j, то есть $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$, но $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$ (Рисунок 3). Если γ рождается в K_i и умирает, входя в K_j, то мы называем разницу в постоянстве значения функции pers(γ) = a_j - a_i. Если γ рождается в K_i, но никогда не умирает, то мы устанавливаем его постоянство на бесконечность.

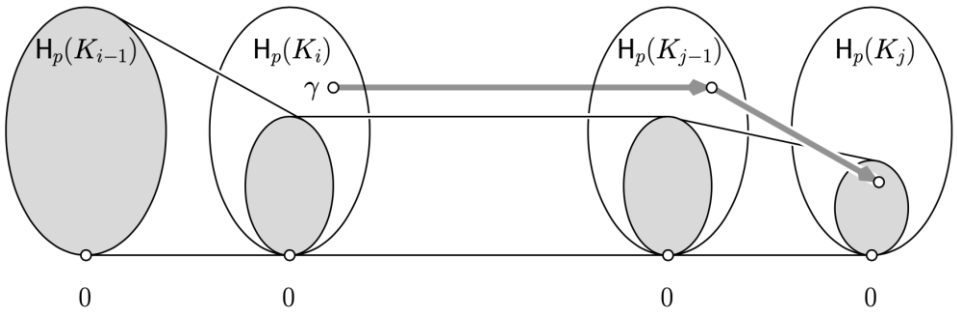

Рисунок 3. Группы гомологий и гомоморфизмы

Визуализация набора постоянных чисел Бетти осуществляется путем рисования точек в двух измерениях [5]. Некоторые из этих точек могут иметь координаты, равные бесконечности, а некоторые могут быть одинаковыми, поэтому мы действительно говорим о мультимножестве точек в расширенной вещественной плоскости, $\bar{R}^2 = (R \cup \{\pm\infty\})^2$. Обозначая $\mu_p^{i,j}$ как количество независимых p-мерных классов, которые рождаются в K_i и умирают при входе в K_j, имеем:

$$\mu_p^{i,j} = \left(\beta_p^{i,j-1} - \beta_p^{i,j}\right) - \left(\beta_p^{i-1,j-1} - \beta_p^{i-1,j}\right), \tag{4}$$

для всех $i<j$ и всех p. Действительно, первая разность в правой части подсчитывает

классы, которые родились в (или раньше) K_i и умирают при входе в K_j, а вторая разность подсчитывает классы, родившиеся в (или раньше) K_{i-1} и умирающие при входе в K_j. Проводя каждую точку (a_i, a_j) с кратностью $\mu_p^{i,j}$, мы получаем p диаграмму персистентности данной фильтрации, обозначаемую как *Dgmp(f).* Он представляет класс в виде точки, расстояние по вертикали до диагонали которой является постоянством. Так как кратности определены только при $i < j$, то все точки лежат выше диагонали. По техническим причинам мы добавляем точки на диагонали диаграммы с бесконечной кратностью. Примеры диаграмм постоянства можно увидеть на Рисунке 4. Постоянные числа Бетти легко вычислить. В частности, $\beta_p^{i,j}$ — количество точек в верхнем левом квадранте с угловой точкой (a_k, a_l). Класс, родившийся в K_i и умерший в K_j, считается тогда и только тогда, когда $a_i \le a_k$ and $a_j > a_l$. Таким образом, квадрант закрыт по вертикали справа и открыт по горизонтали снизу.

Основная лемма о стойких гомологиях. Пусть $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$ фильтрация. Для каждой пары индексов $0 \le k \le l \le n$ и каждой размерности p p-е постоянное число $\beta_p^{i,j} = \sum_{i \le k} \sum_{j > l} \mu_p^{i,j}$.

Это важное свойство. В нем говорится, что диаграмма персистентности кодирует всю информацию о персистентных гомологичных группах.

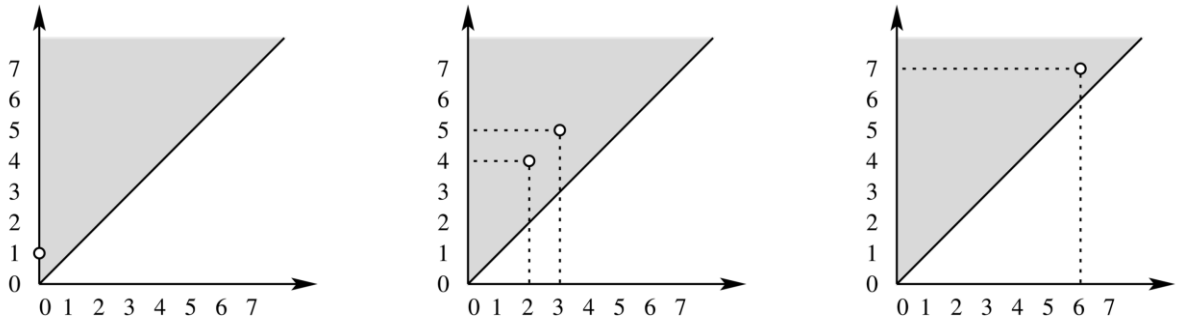

Рисунок 4. Примеры постоянных диаграмм

Вариационные автоэнкодеры (VAE) [12, 13] определяют вероятностный порождающий процесс между наблюдением x и скрытой переменной z следующим образом: $z \sim p_\theta(z)$ и $x \sim p_\theta(x|z)$, где $p_\theta(z)$ и $p_\theta(x|z)$ — функции распределения вероятностей, параметризованные по θ. В системе обучения без учителя нам дается только набор данных $X=\{x_i\}$, но истинное значение θ и скрытая переменная z для каждого наблюдения x неизвестны. Часто нас интересует предельная вероятность данных $p_\theta(x)$ или $p_\theta(z|x)$. Однако оба требуют вычисления сложного интеграла $\int p_\theta(z)\, p_\theta(x|z) dz$.

Для решения этой проблемы в VAE вводится модель распознавания $q_\phi(z|x)$, которая аппроксимирует истинную вероятность $p_\theta(z|x)$. Следовательно, мы можем переписать предельную вероятность как:

$$\log\ p_\theta(x) = D_{KL}\left(q_\phi(z|x) \middle|\middle| p_\theta(z|x)\right) + L(\theta, \phi; x) \ge L(\theta, \phi; x) = -D_{KL}\left(q_\phi(z|x) \middle|\middle| p_\theta(z)\right) + E_{q_\phi(z|x)}[\log\ p_\theta(x|z)] \quad (5)$$

где $L(\theta, \phi; x)$ — вариационная нижняя граница (ELBO), которую мы хотим оптимизировать по отношению к θ и ϕ. Оптимизация ELBO по θ и ϕ осуществляется с использованием стохастической вариации градиента Байеса (SGVB).

В рамках VAE предполагается, что модель распознавания $q_\phi(z|x)$ и порождающая модель $p_\theta(x|z)$ параметризуются с помощью диагональных распределений Гаусса, в которых вычисляются среднее значение и ковариация с помощью нейронной сети. Предполагается также, что априорная вероятность $p_\theta(z)$ является центрированным изотропным многомерным гауссианом, т. е. $p_\theta(z) = N(z, \mathbf{0}, \mathbf{I})$, не имеющим свободных параметров. На практике

математическое ожидание аппроксимируется выборкой K отсчетов из $z^k \sim q_\phi(z|x)$, а затем вычислением уравнения $E_{q_\phi(z|x)}[\log p_\theta(x|z)] \approx \frac{1}{K}\sum_{k=1}^{K} \log p_\theta(x|z^k)$. Для получения дифференцируемой сети после дискретизации используется прием репараметризации [13]. Предположим, что $z = N(z; \mu_z, \sigma_z^2\mathbf{I})$), тогда после перепараметризации имеем $z = \mu_z + \sigma_z \bullet \varepsilon$, где • обозначает поэлементное произведение, а вектор ε выбирается из N (0; I) и рассматривается как дополнительный вход.

Обсуждение

В качестве входных данных мы выбрали набор данных TIMIT [14]. Корпус прочитанной речи TIMIT предназначен для предоставления речевых данных для акустико-фонетических исследований, а также для разработки и оценки систем автоматического распознавания речи. TIMIT содержит широкополосные записи 630 носителей 8 основных диалектов американского английского, каждый из которых читает по 10 фонетически насыщенных предложений. Корпус TIMIT включает выровненные по времени орфографические, фонетические и словесные транскрипции, а также 16-битный файл формы речевого сигнала 16 кГц для каждого высказывания. Разработка корпуса была результатом совместных усилий Массачусетского технологического института (MIT), SRI International (SRI) и Texas Instruments, Inc. (TI). Речь была записана в TI, расшифрована в Массачусетском технологическом институте, проверена и подготовлена для выпуска компакт-дисков Национальным институтом стандартов и технологий (NIST).

Мы используем вариационный автоэнкодер со свёрточными слоями, которые были реализованы в Keras. Модель состоит из входного слоя, слоев кодера и декодера. Входная форма имеет размерность (256, 64, 1). Чтобы приспособиться к этому, мы изменяем наши аудио- и текстовые данные. В частности, для аудиоданных мы используем двумерные спектрограммы, тогда как для текстовых данных мы используем двумерные вложения слов. Кодер и декодер имеют по пять сверточных слоев каждый. Мы устанавливаем количество эпох равным 150, а размер пакета — 64. На Рисунке 5 показана архитектура VAE с входными и выходными параметрами. В конечном итоге нас интересуют промежуточные векторы между слоями кодера и декодера, т.е. те, которые имеют размерность 128.

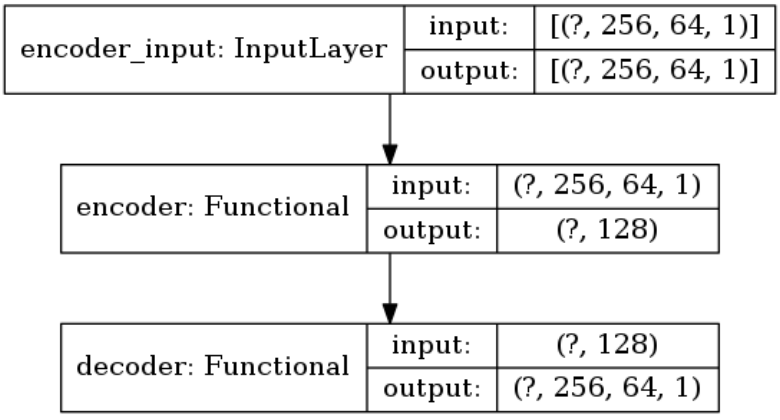

Рисунок 5. Архитектура VAE с входными и выходными параметрами

Чтобы извлечь признаков из аудиоданных, нам нужно выполнить некоторые шаги предварительной обработки. Каждый аудиофайл был разделен на части с использованием временных меток слов, указанных в аннотациях базы данных TIMIT. Затем мы выполнили заполнение сигналов и из этих сигналов извлекли векторы MFCC (спектрограммы). Для этого мы

использовали librosa [15], библиотеку для предварительной обработки музыки и аудио. В качестве последнего шага мы нормализовали векторы. Все эти шаги были выполнены для каждого из аудиофайлов в наборе данных. Эти векторы использовались для обучения нашей модели VAE и извлечения промежуточных векторов. На Рисунке 6 показаны потери при обучении модели.

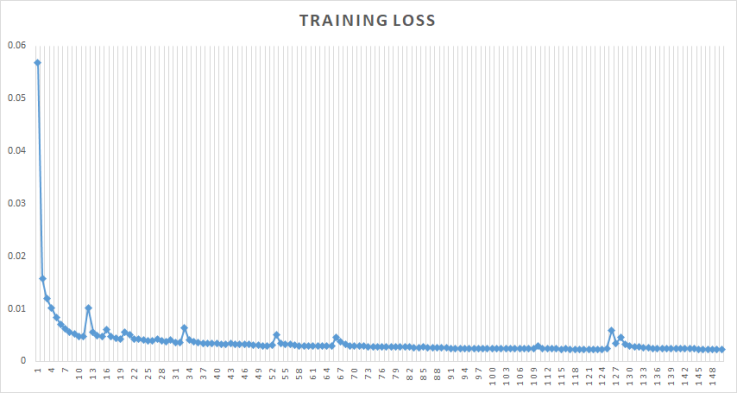

Рисунок 6. Обучение VAE для аудиоданных

Для текстовых данных мы использовали библиотеку Gensim, которая поддерживает реализацию встраивания слов Word2Vec для изучения новых векторов слов из текста. Он также предоставляет инструменты для загрузки предварительно обученных вложений слов в нескольких форматах, а также для использования и запроса загруженных вложений [16]. Как только вложения слов были получены, мы обучили модель VAE и извлекли промежуточные векторы. На Рисунке 7 показаны потери при обучении модели.

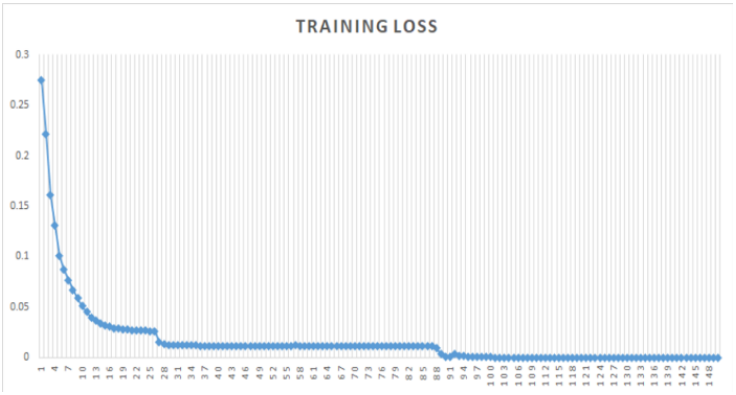

Рисунок 7 – Обучение VAE для текстовых данных

Результаты

Данные для расчета персистентной гомологии могут быть представлены по-разному в зависимости от изучаемой предметной области - взвешенные графики, изображения, облака точек. В нашем случае мы используем третий тип представления аудио- и текстовых данных — облака точек, так как данные уже заранее извлечены в виде набора векторов, соответствующих отдельным словам. Таким образом, у нас есть два 128-мерных набора векторов — для аудио и текстового доменов. Векторы записываются построчно в текстовом файле, а векторные компоненты разделяются запятыми. Здесь мы предполагаем, что векторное пространство снабжено стандартной евклидовой метрикой. В Таблице 1 приведены статистические данные о количестве векторов для обоих типов данных.

Таблица 1. Статистика данных

Тип данных	**Количество векторов**	**Размер вектора**
Аудиоданные	54378	128
Текстовые данные	6224	128

Для построения симплициальных комплексов (Виеториса–Рипса) и персистентных гомологии мы использовали пакет Ripser [17], один из лучших среди аналогов по скорости вычислений и написанный на C++. Процесс расчета занимает до 3 часов для аудиоданных и до 1,5 часов для тестовых данных. Причина в том, что объем аудиоданных больше из-за повторения отдельных слов, которые произносятся несколько раз, в то время как в текстовых данных каждое слово встречается только один раз. Наконец, мы построили персистентные диаграммы, показанные на Рисунке 8. Красные точки соответствуют числам Бетти $\beta = 0$, а синие точки – $\beta = 1$.

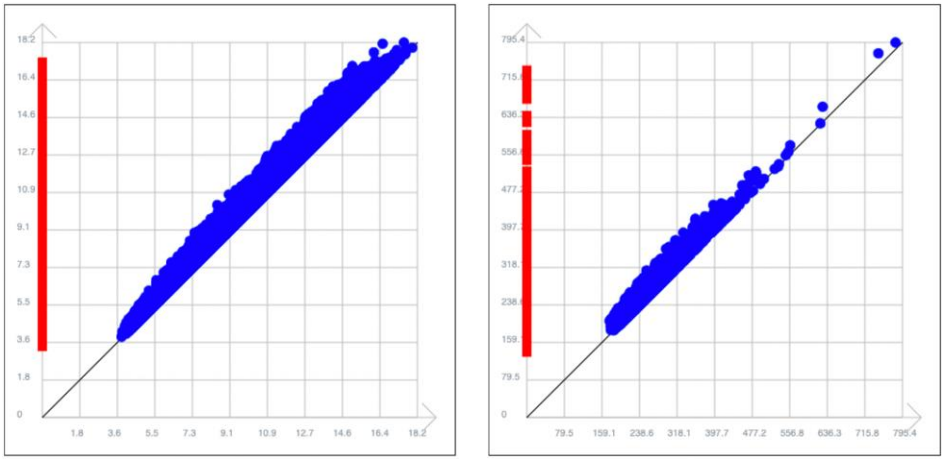

Рисунок 8 – Постоянная диаграмма аудиоданных (слева) и текстовых данных (справа)

Для анализа распределения устойчивых интервалов были построены гистограммы длин интервалов для размерности 1 (Рисунок 9). Гистограммы в обоих случаях аналогичны экспоненциальному распределению, но с другими параметрами.

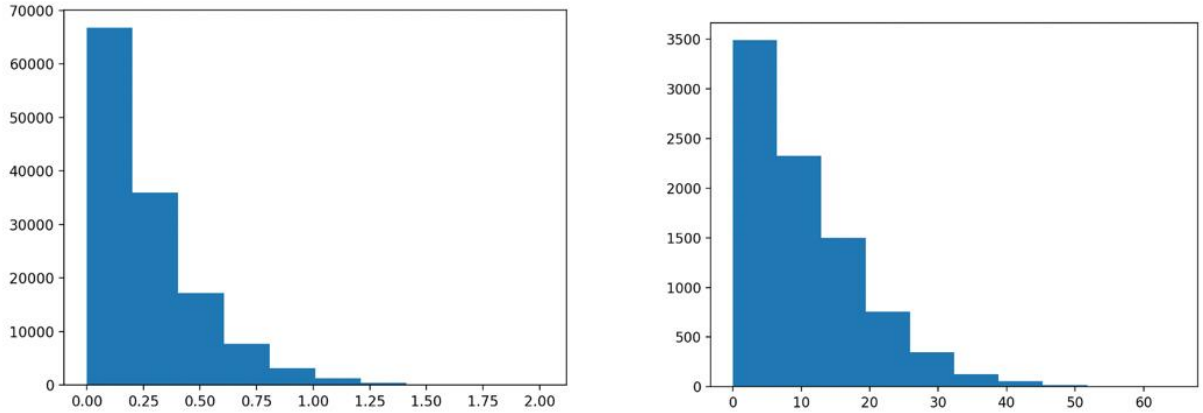

Рисунок 9. Гистограммы постоянных интервалов для размерности 1

Как видно из Рисунка 8, диаграммы аудио- и текстовых данных очень похожи, т.е. имеют схожую топологическую структуру, за исключением того, что синие точки присутствуют в меньшей степени в верхней части графика. Это может быть связано с тем, что звуковые данные для

каждого слова повторяются и, соответственно, образуют больше «отверстий» в пространстве, которые закрываются на более поздних этапах фильтрации (при больших радиусах шара).

Выводы

В данной статье нами представлена предварительная работа по анализу и визуализации аудио- и текстовых данных с использованием методов персистентной гомологии. Как показали результаты, топологическая структура обоих пространств достаточно схожа, что подтверждает гипотезу о сходстве звукового и текстового пространств, а само исследование может быть продолжено. В частности, мы планируем провести количественное сравнение диаграмм персистентности с использованием римановой метрики или метрики Вассерштейна. Кроме того, информацию о топологическом подобии обоих пространств можно использовать непосредственно при обучении вариационных автоэнкодеров и извлечении признаков.

Благодарности

Работа выполнена при поддержке грантового финансирования проектов Комитета науки Министерства науки и высшего образования Республики Казахстан (гранты No. AP13068635 и No. AP08053085).

Список литературы

1. Peikert R., Hauser H., Carr H., Fuchs R. Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications. – Springer Science & Business Media, 2012. – P. 299.
2. Carlsson G. Topology and data // Bulletin of the American Mathematical Society. – 2009. –V. 46, №2. – P. 255-308.
3. Zomorodian A. J. Topology for Computing. – Cambridge University Press, 2005. – P. 240.
4. Ghrist R. Barcodes: The Persistent Topology of Data // Bulletin of the American Mathematical Society. – 2008. –V. 45, №1. – P. 61-75.
5. Edelsbrunner H., Harer J. Computational Topology, An Introduction. – American Mathematical Society, 2010. – P. 242.
6. Kaczynski T., Mischaikow K., Mrozek M. Computational Homology. – Springer, 2004. – P. 482.
7. Carlsson E., Carlsson G., de Silva V. An algebraic topological method for feature identification // International Journal of Computational Geometry and Application. –2006. –V. 16, №4. – P. 291–314.
8. Ferri M., Frosini P., Cerri A., Di Fabio. Computational Topology in Image Context. – Springer, 2012. – P. 156.
9. De Floriani L., Spagnuolo M. Shape Analysis and Structuring. – Springer, 2008. – P. 296.
10. Najman L., Talbot H. Mathematical Morphology: From Theory to Applications. – John Wiley & Sons, Inc., 2010. – P. 507.
11. Edelsbrunner H., Morozov D. Persistent homology: theory and practice // Proceedings of the European Congress of Mathematics. – Krakow, 2012. –P. 31-50.
12. Hsu W., Zhang Y., Glass J. R. Learning latent representations for speech generation and transformation // Proceedings of Interspeech. – Stockholm, Sweden, 2017. –P. 1273-1277.
13. Kingma D. P., Welling M. An introduction to variational autoencoders // Foundations and Trends in Machine Learning. –2019. –V. 12. –P. 307-92.

14. Garofolo J. S., Lamel L. F., Fisher W. M., Fiscus J. G., Pallett D. S. DARPA TIMIT acoustic-phonetic continuous speech corpus // NIST speech disc 1-1.1. NASA STI/Recon technical report. –1993. – V. 93. –P. 27403.

15. McFee B., Raffel C., Liang D., Ellis D. P., McVicar M., Battenberg E., Nieto O. librosa: Audio and music signal analysis in python // Proceedings of the 14th python in science conference. – Austin, Texas, 2015. –V. 8. –P. 18-25.

16. Srinivasa-Desikan B. Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras / B. Srinivasa-Desikan. – Packt Publishing Ltd, 2018. – P. 286.

17. Bauer U. Ripser: efficient computation of Vietoris–Rips persistence barcodes // Journal of Applied and Computational Topology. –2021. –V. 5, №3. –P. 391-423.

Ж.М. Кожирбаев, Ж.А. Есенбаев
National Laboratory Astana, Астана, Қазақстан

Теңгерілмеген аудио және мәтіндік деректерді топологиялық талдау

Аннотация: Біз бақылаусыз сөйлеуді өңдеу үшін аудио және мәтіндік деректерді топологиялық талдау бойынша алдын ала жұмыс жасадық. Жұмыс фонема жиіліктері мен контекстік қатынастар бір тілдің акустикалық және мәтіндік салаларында ұқсас болады деген болжамға негізделген. Тиісінше, бұл олардың геометриялық құрылымын ескере отырып, осы кеңістіктер арасында кескін құруға мүмкіндік береді. Бірінші қадам ретінде аудио және мәтіндік деректерді екі жасырын векторлық кеңістікте салыстыру үшін вариациялық автокодерлерге негізделген генеративті әдістер таңдалды. Келесі кезеңде екі кеңістіктің топологиялық құрылымын талдау үшін тұрақты гомологиялық әдістер қолданылады. Алынған нәтижелер екі кеңістіктің ұқсастығы туралы идеяны қолдаса да, акустикалық және мәтіндік кеңістіктерді дұрыс кескінге түсіру үшін, сонымен қатар топологиялық ақпаратты автокодерлерді оқыту процесіне қосудың нақты әсерін бағалау үшін қосымша зерттеулер қажет.

Түйін сөздер: бақыланбайтын сөйлеуді өңдеу, вариациялық автокодерлер, сөздерді енгізу, топологиялық деректерді талдау, тұрақты гомология және диаграммалар.

Zh.M. Kozhirbayev, Zh.A. Yessenbayev
National Laboratory Astana, Astana, Kazakhstan

Topological analysis of unaligned audio and text data

Abstract: We have performed preliminary work on topological analysis of audio and text data for unsupervised speech processing. The work assumes that phoneme frequencies and contextual relationships are similar in the acoustic and text domains for the same language. Accordingly, this allowed the creation of a mapping between these spaces that considers their geometric structure. As a first step, generative methods based on variational autoencoders were chosen to map audio and text data into two latent vector spaces. In the next stage, persistent homology methods are used to analyze the topological structure of two spaces. Although the results obtained support the idea of the similarity of the two spaces, further research is needed to correctly map acoustic and text spaces, as well as to evaluate the real effect of including topological information in the autoencoder training process.

Keywords: unsupervised speech processing, variational autoencoders, word embeddings, topological data analysis, persistent homology and diagrams.

References

1. Peikert R., Hauser H., Carr H., Fuchs R. Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications (Springer Science & Business Media, 2012, 299 p.)

2. Carlsson G. Topology and data [Bulletin of the American Mathematical Society]. 2009. Vol. 46. №2. P. 255-308.

3. Zomorodian A.J. Topology for Computing (Cambridge University Press, 2005, 240 p.).

4. Ghrist R. Barcodes: The Persistent Topology of Data [Bulletin of the American Mathematical Society]. 2008. Vol. 45. №1. P. 61-75.

5. Edelsbrunner H., Harer J. Computational Topology, An Introduction (American Mathematical Society, 2010, 242 p.)

6. Kaczynski T., Mischaikow K., Mrozek M. Computational Homology (Springer, 2004, 482 p.)

7. Carlsson E., Carlsson G., de Silva V. An algebraic topological method for feature identification [International Journal of Computational Geometry and Application]. 2006. Vol. 16. №4. P. 291–314.

8. Ferri M., Frosini P., Cerri A., Di Fabio. Computational Topology in Image Context (Springer, 2012, 156 p.)

9. De Floriani L., Spagnuolo M. Shape Analysis and Structuring (Springer, 2008, 296 p.)

10. Najman L., Talbot H. Mathematical Morphology: From Theory to Applications (John Wiley & Sons, Inc., 2010, 507 p.)

11. Edelsbrunner H., Morozov D. Persistent homology: theory and practice [Proceedings of the European Congress of Mathematics], Krakow, Poland, 2012. P. 31-50.

12. Hsu W., Zhang Y., Glass J.R. Learning latent representations for speech generation and transformation [Proceedings of Interspeech], Stockholm, Sweden, 2017. P. 1273-1277.

13. Kingma D.P., Welling M. An introduction to variational autoencoders [Foundations and Trends in Machine Learning]. 2019. Vol. 12. P. 307-92.

14. Garofolo J.S., Lamel L.F., Fisher W.M., Fiscus J.G., Pallett D.S. DARPA TIMIT acoustic-phonetic continuous speech corpus [NIST speech disc 1-1.1. NASA STI/Recon technical report]. 1993. Vol. 93. P. 27403.

15. McFee B., Raffel C., Liang D., Ellis D.P., McVicar M., Battenberg E., Nieto O. librosa: Audio and music signal analysis in python [Proceedings of the 14th python in science conference], Austin, Texas, 2015. Vol. 8. P. 18-25.

16. Srinivasa-Desikan B. Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras (B. Srinivasa-Desikan. – Packt Publishing Ltd, 2018, 286 p.)

17. Bauer U. Ripser: efficient computation of Vietoris – Rips persistence barcodes [Journal of Applied and Computational Topology]. 2021. Vol. 5. №3. P. 391-423.

Сведения об авторах

Кожирбаев Ж.М. – Ph.D., старший научный сотрудник, National Laboratory Astana, пр. Кабанбай батыра, 53, Астана, Казахстан.

Есенбеав Ж.А. – Ph.D., старший научный сотрудник, National Laboratory Astana, пр. Кабанбай батыра, 53, Астана, Казахстан.

Kozhirbayev Zh. M. – Ph.D., Senior Researcher, National Laboratory Astana, 53 Kabanbay batyr ave. , Astana, Kazakhstan.

Yessenbayev Zh. A. – Ph.D., Senior Researcher, National Laboratory Astana, 53 Kabanbay batyr ave., Astana, Kazakhstan.