ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК: 515.1, 004.8

Отчет об исследовательском проекте

на тему: Топологическое глубокое обучение: параметризованные фильтрации

Выполнил:
Студент группы БПМИ206 _____ Айтхаджаев Каримжан
Подпись И.О.Фамилия

19.05.2022
Дата

Принял:
Руководитель проекта Качан Олег Николаевич
Имя, Отчество, Фамилия
Стажер исследователь ФКН / МЛ АТиП
Должность, ученое звание
ФКН НИУ ВШЭ
Место работы (Компания или подразделение НИУ ВШЭ)

Дата проверки _____ 2022 _____ _____
Оценка (по 10-ти бальной шкале) Подпись

Москва 2022

Реферат

В данном проекте проводится исследование и проверка гипотез связанных с примененнием методов топологии, а конкретно теории персистентных гомологий в улучшении выполнения уже существующих методов анализа данных. Конкретно, проверялись гипотезы связанные с преимуществом параметризованных фильтраций больше чем одного направления, а также сравнение классов аггрегации. В результате также сравнивались несколько моделей в задаче классификации mnist на основе векторизованного представления персистентных диаграмм в качестве данных.
Ключевые слова: топологический анализ данных, нейронные сети, проверка гипотез

Содержание

1 Основные термины и определения

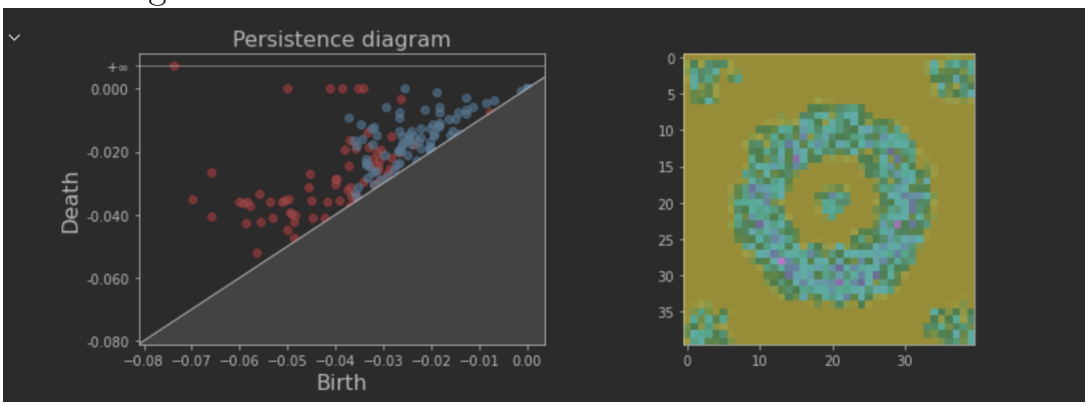
Гомология подпространства - метод формализованный под измерение присутствия n-мерных дыр в подпространстве, или измерения формы топологического объекта, также как и метод построения алгебраических объектов.

Персистентная гомология - Метод, с помощью которового можно измерять или вычеслять компактное представление топологических признаков в данных, причем нет зависимости от координат и размерностей.

Облако точек - датасет с генерированными точкам на пространстве

Параметризованные фильтрации - последовательность вложенных в друг друга подпространств, зависящих от одного или нескольких параметров, которые могут задавать структуру по которой будут создаваться эти подпространства. У каждого подпростанства может быть вычеслена его гомология, где задается понятие n-ой персистентной гомологии – последовательность вычесленных гомологий подпространств.

Персистентная диаграмма, или др словами диаграмма устойчивости - это диаграмма, описывающая временной промежуток существования топологического признака для разной размерности(циклы, воиды, дыры)
Пример диаграммы устойчивости в анализа данных crater tuto из примеров библиотеки gudhi



Симплициальный комплекс/пространство - это множество точек, прямых, треугольников, и их многомерных аналогов. Неформально говоря, это топологическое пространство построенного с помощью симплексов меньшей размерности.

Аггрегация - в данном контексте, это сборка нескольких слоев параметризованных фильтраций под одну функцию, например max, min, mean итд.

2 Введение

Топологический анализ данных является быстро развивающейся областью исследования, так как она привлекает своими результатами в методе редукции размерности датасетов и устойчивостью к шуму в данных. Также в недавнее время, TDA стало обширно использоваться в улучшении качества в задачак классификации изображений. Самый простой пример применения TDA, это построение облака точек достаточно большой размерности, и с помощью персистентной гомологии, получение информации о топологической структуре этих данных, на основе которых уже снижать размерность без существенной потери информации.
Персистентные гомологии являются краеугольным камнем TDA(Topological Data Analysis), с ее помощью можно строить на основе данных персистентные диаграммы, которые до недавнего исследования в основном использовались как для выявления, так и сокращения количества признаков в базах данных. Для разных типов данных, процесс построения персистентных диаграмм может отличаться. Например для облаков точек, основным топологическим объектом остаются они же сами, и при рассмотрении параметра ϵ можно понять, какой вид будет у структуры этих точек(Vietoris-Rips complex), когда как если применять такую же логику к изображениям, где роль точек на Евклидовом пространстве будут играть пиксели(или воксели в больших размерах), лучше всего будет эксплуатировать саму сеточную структуру изображений. Тогда рассматриваются кубические комплексы, то есть роль симплексов у облаков точек будут играть многомерные кубы. Целью проекта являлось ознакомление с областью исследования Топологического анализа данных для проверки нескольких гипотез:

- 1: Фильтрация по нескольким направлениям лучше чем фильтрация по одному направлению.
- 2: Сравнение аггрегаций.
- 3: Выбор параметров оптимизацией лучше чем выбор экспертом в зависимости от количества фильтров.

3 Обзор и сравнительный анализ источников

Основная часть исследования была проведена с использованием Python, и специальных для работы фреймворков и библиотек таких как Gudhi и Ripser. В качестве МЛ алгоритмов для сравнения использовались Random Forest, LightGBM и Multi Layer Perceptron(MLP). Также был применен Persformer, архитектура для трансформиравния персистентных диаграмм в уже готовые для работы с алгоритмами признаки. В качестве данных использовался датасет рукописных цифр MNIST и облака точек.

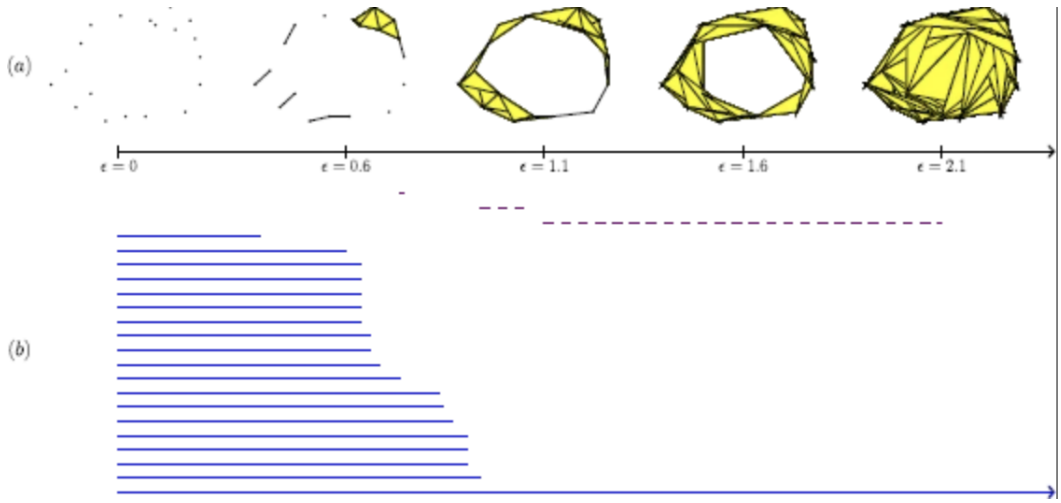
4 Теоретическая часть

Процесс реализации проекта начинается с импортирования данных таких как MNIST и генерации облаков точек. Для построения персистентный диаграмм будем строить симплициальные комплексы(а конкретно Vietoris-Rips/Delaunay complex) по облакам точек, и кубические комплексы по картинкам.
Далее, из полученного фильтрованого комплекса, нужно получить баркоды, для этого для каждого комплекса K будет ассоциирована boundary matrix - матрица которая будет содержать информацию о гранях комплекса. Для этого фильтрованных симплексов наложен порядок по вложенности, в котором каждый должен предшествовать другому по мере их построения, в плане у каждого грань симплекса будет пред самого симплекса. Формально, фильтрации будут представими ввиде следующего отображения $(X, f_\theta) \to D_{f_\theta}(X) \to \mathbb{R}^n$.

$$\mathscr{X} : X_0 \subset X_1 \subset \cdots \subset X_L.$$

$$PH_n(\mathscr{X}; \mathbb{F}) : H_n(X_0; \mathbb{F}) \to H_n(X_1; \mathbb{F}) \to \cdots \to H_n(X_L; \mathbb{F})$$

Со вложенностью комплекса связано еще и рождение и смерть топологических признаков, таких как например циклы, дыры, воиды(voids). Промежуток от рождения признака до смерти будет описываться баркодом.
Например, как на рисунке:



Пример для кубического комплекса:

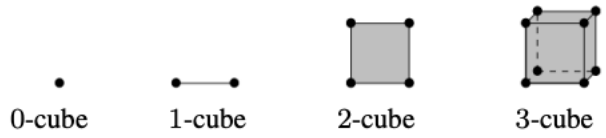
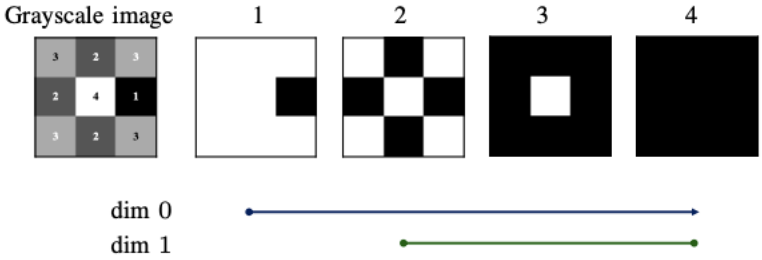


Fig. 3. Cubes of dimension 0, 1, 2 and 3.



Матрица такова, что каждый элемент ij будет индикаторным элементом на то, является симплекс $\sigma(i)$ гранью $\sigma(j)$. Нужно также редуцировать ее, например предлогается использовать метод Гаусса и Стандартный алгоритм редукции(low(i) - максимальный индекс строки где значение отличное от 0):

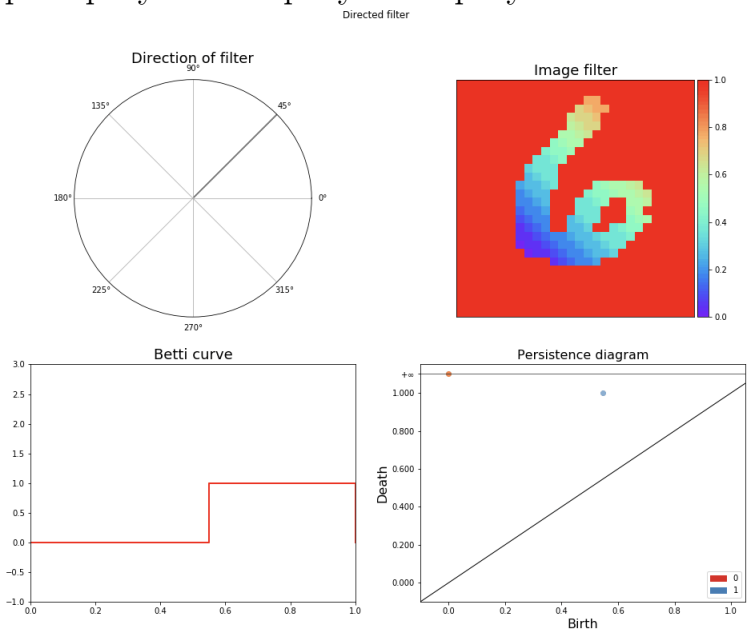
```
for j = 1 to n do
    while there exists i < j with low(i) == low(j) do
        add column i to j
    end while
end for
```

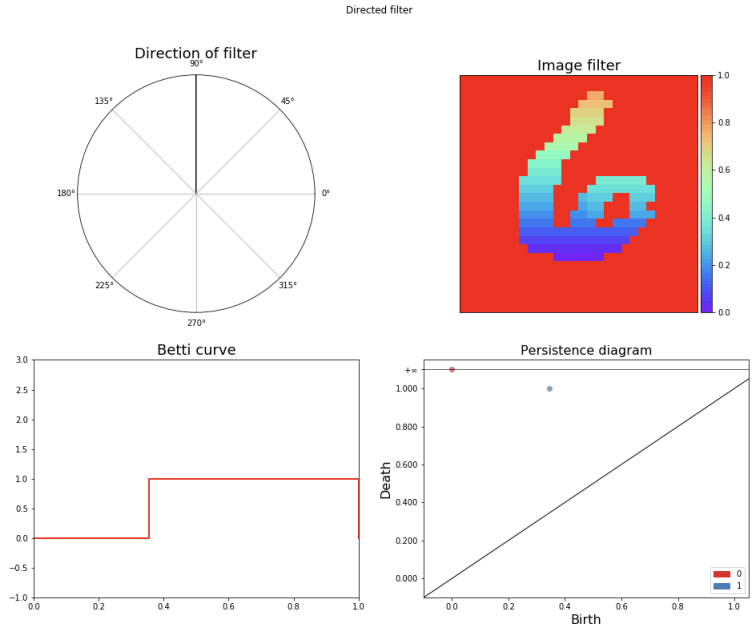
После этого строятся баркоды на интервале $[dg(i), dg(j)]$, где dg(i) - это минимальное l такое что симплекс будет принадлежать K_l. Используя библиотеку трансформеров Persformer, можно на основе облаков точек выдавать готовые персистентные диаграммы. Также есть Ripser, который наиболее предпочтителен для подсчета персистентной гомологии, и его усиленная версия, поддерживающая GPU - Ripser++. Плюс в его использовании в том, что более эффективно аллоцируется и используется память, а минус в работе на одном ядре.

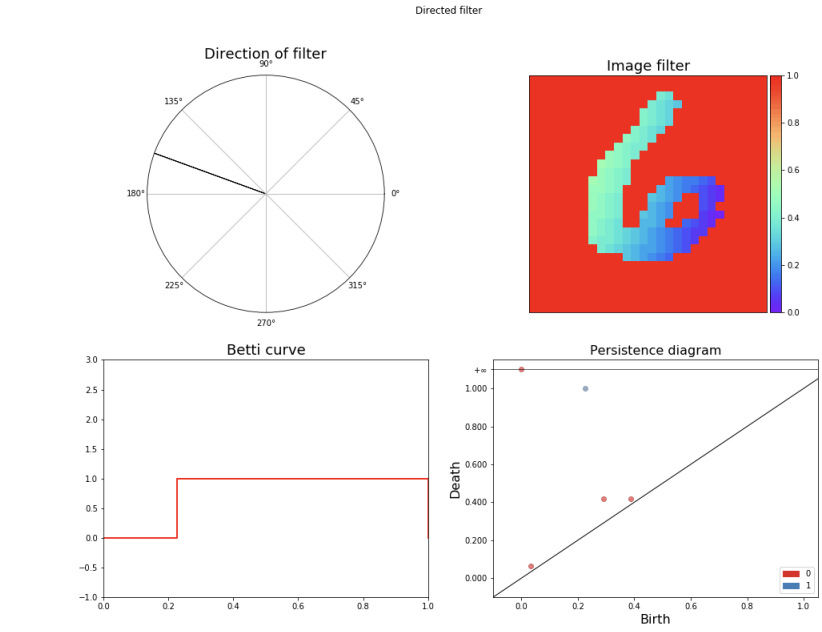
5 Описание вычислительного эксперимента

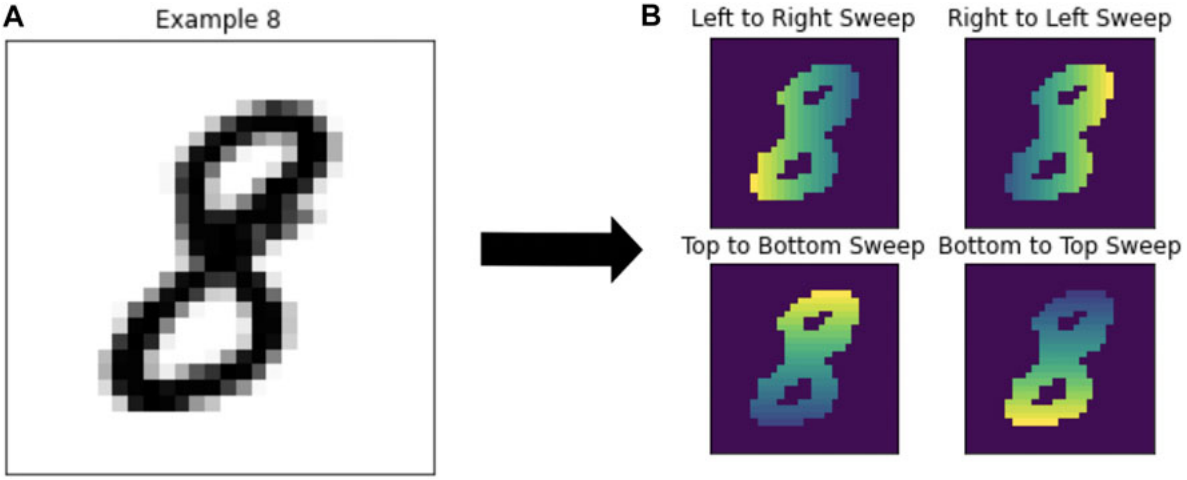
Основной частью проекта является реализация параметризованных фильтраций по разным направлениям для изображений(расстояние до касательной линии к вписанной окружности по направлению) и облаков точек(дискретизация смеси гауссовский распределений по ширине гауссовского ядра в смеси распределений) и проверка соответствующих гипотез.

Пример нужных требуемых результатов:







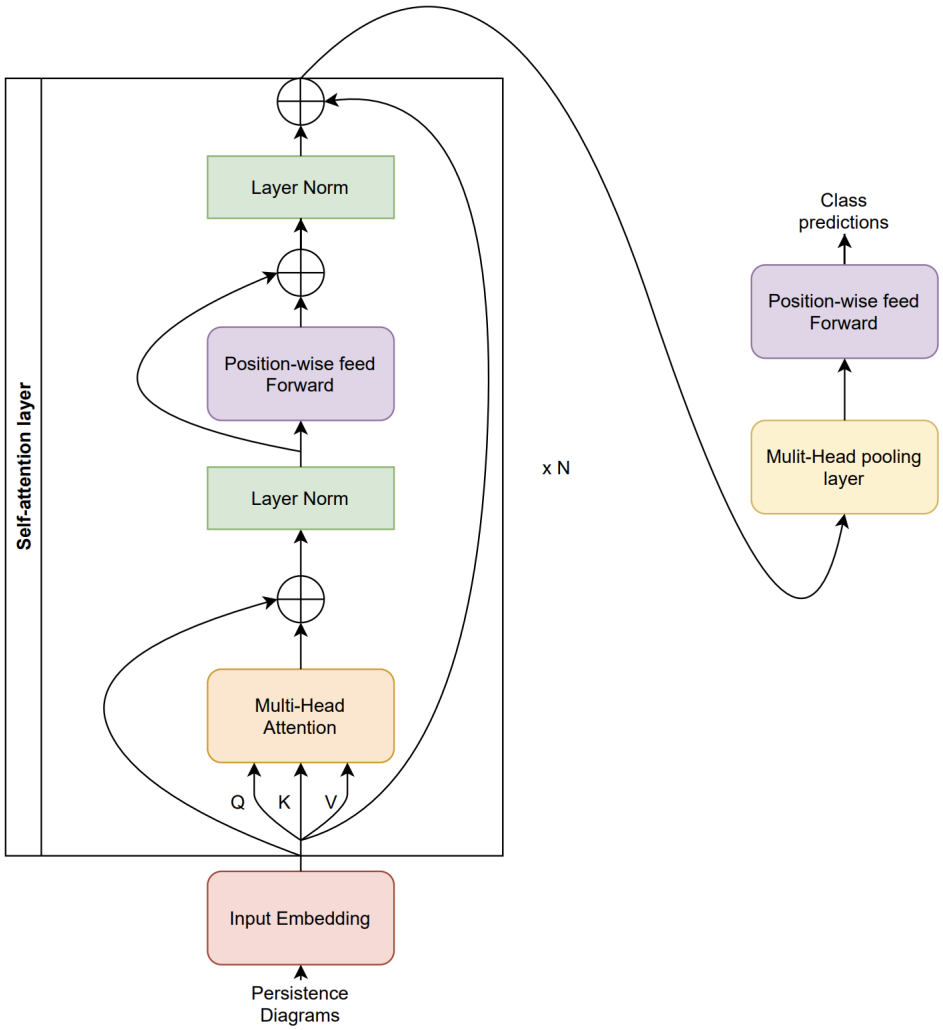


Также реализация анизотропных фильраций, то есть фильтраций которые для каждой точки/пикселя будут расширять параметр по мере наклона. Например, если для облаков точек, для каждой точки мы вводили радиус, по мере увеличения которого мы получали окружность и при пересечении соединяли точки, в результате чего изменялся/расширялся симплекс, то теперь вместо окружности рассматривался эллипсоид и его направление при котором он повернут.

Далее для перевода из диаграммы в вектор, применялась библиотека diagram2vec для получения чисел Бетти, которые в свою очередь применялись для кластеризации облаков точек.

Также была использована библиотека PersLay - слой нейронной сети для персистентных диаграмм.

Примерная архитектура работы Perslay:



Второй частью являлось написать нейросеть, которая на вход принимала несколько слоев параметризованных фильтрации, которые могут аггрегироваться как по конкатенации, так и по стандартным статистическим операциям(mean, max, min ...) и в другой реализации слои MLP.

Третьей частью уже является решение задачи классификации которая решалась для MNIST и облаков точек на 5 классах и сравнение результатов со state-of-the-art моделями.

Упрощенный вид пайплайна работы с персистентными диаграммами таков:

Данные -> Фильтрованные Комплексы -> Баркоды -> Интерпретация результатов

В качестве решения самой задачи и проверки гипотез использовались уже вышеперечисленные Ripser и Gudhi для облаков точек и MNIST соответственно. Также были применены методы из статьи Persformer и Perslay как альтер-

нативные методы векторизаций диаграмм устойчивости и представления этих диаграмм в компактный вид для преобразования их в слой нейронной сети.

В итоге средний результат для LightGBM для случайных изображений оказался 0.73 для фиксированного угла направления. Если же выбрать угол случайно, то score снижается до среднего 0.62. Линейная модель RidgeClassifier на числах Бетти дала средний score 0.66, когда как использование модели MLP со слоем параметризованных фильтраций на тестовых изображениях дало средний похожий результат 0.62 для фиксированного направления, и средний score 0.65 для нескольких слоев с разными направлениями под усредняющей функцией аггрегации.

6 Заключение

В рамках этого проекта были рассмотрены подходы для применения топологических методов для решения задач анализа данных. Также, исследовались гипотезы, из проверки которых получилось что слой в нейронной сети состоящий из аггреграции параметризованных фильтраций по нескольким направлениям топологического объекта полученного из датасета, давал довольно лучшие результаты, по сравнению со слоем фильтрации по одному направлению.
Также при аггрегировании выходов представляющих персистентные диаграммы, а также аггрегировании чисел Бетти, получились незначительно лучшие результаты, чем если брать один слой параметризованных фильтраций + слой MLP.

Список литературы

1) A Topological "Reading" Lesson: Classification of MNIST using TDA, Adélie Garin, Guillaume Tauzin

2) Smooth Summaries of Persistence Diagrams and Texture Classification

3) A Comparative Study of Machine Learning Methods for Persistence Diagrams

4) A roadmap for the computation of persistent homology

5) Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks

6) Persformer: A Transformer Architecture for Topological Machine Learning

7) Ripser++

8) A General Neural Network Architecture for Persistence Diagrams and Graph Classification

9) Efficient Computation of Persistent Homology for Cubical Data

10) Deep Sets

11) PersLay