Обучение взаимосвязанных информативных представлений в задаче генерации образов

Охотников Никита Владимирович

МФТИ

2023-2024

Введение

Цель

- Исследовать проблемы дополнения, генерации и оценки образов, состоящих из элементов заранее заданного конечного множества

Проблемы

- Высокая дисперсия оценки образа и сложность построения объективных критериев
- Нетривиальная взаимосвязь элементов образа
- Практическая невозможность решения полным перебором задач к нему сводящихся

Необходимо

- Изучить возможности аппроксимации точного решения, в случае когда оно существует, но невычислимо
- Рассмотреть способы моделирования связей между частями внутренней структуры образов

Постановка задачи

Основные понятия и обозначения

- Основная единица данных, рассматривающаяся в работе – элемент одежды, далее будем называть его *объектом* или *элементом*, множество всех рассматриваемых объектов – $\mathcal{X}$
- Каждый объект $X \in \mathcal{X}$ есть пара $X = (I, T)$ из соответственно изображения о текстового описания. $\mathcal{I}$ – множество изображений объектов.
- Для каждого элемента X определена *категория* C_x, из множества категорий $\mathcal{C}$, а множество всех элементов разбивается на подмножества $\mathcal{X}_{C_i}$элементов с общей категорией.
- Некоторое подмножество $O = \{X_i\}_{i=1}^{k} \subset \mathcal{X}$ множества всех элементов будем называть *образом*, если
  1. $O \neq \{\emptyset\}$
  2. $|O| \leqslant K$
  3. $\forall X_i, X_j \in O, i \neq j \longrightarrow C_{X_i} \neq C_{X_j}$

  где K – определяемая задачей константа. Множество всех образов $\mathcal{O}$. Из такого определения следует, в частности:

$$O \in \mathcal{O}, O' \subset O \longrightarrow O' \in \mathcal{O}$$

Постановка задачи

Основные понятия и обозначения

- Для элементов и образов будем рассматривать *функции близости*

$$S_X : \mathcal{X} \times \mathcal{X} \longrightarrow [-1, 1], \ \ \forall X \in \mathcal{X} \ S_X(X, X) = 1$$

$$S_O : \mathcal{O} \times \mathcal{O} \longrightarrow [-1, 1], \ \ \forall O \in \mathcal{O} \ S_O(O, O) = 1$$

  Такой функцией может выступать например косинусное сходство в некотором латентном пространстве.

- Для оценки образов введем функцию *оценки* или *совместимости* его элементов:

$$\mathcal{S} : \ 2^{\mathcal{X}} \longrightarrow [0, 1]$$

  причем выполнено следующее:

$$\forall O \in \mathcal{O} : \ \mathcal{S}(O) > 0$$

$$\forall O' \in 2^{\mathcal{X}} \setminus \mathcal{O} : \mathcal{S}(O') = 0$$

  Совместимостью или *оценкой совместимости* образа O будем называть результат применения функции совместимости к этому образу $\mathcal{S}(O)$

Постановка задачи

Оценка образа

Задача оценки образа – это классическая задача регрессии направленная на получения аппроксимации функции оценки $\mathcal{S}$:

- **Дано:**
  $\{O_1 \dots O_n\} \subset \mathcal{O}$
  $\{\mathcal{S}(O_1) \dots \mathcal{S}(O_n)\}$
- **Требуется:**
  Найти наилучшую в некотором смысле аппроксимацию функции $\mathcal{S}$ функциями заданного класса, т.е. решить задачу оптимизации:

$$\hat{\mathcal{S}} = \underset{S \in \mathscr{S}}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\mathcal{S}(O_i), S(O_i)) \right]$$

  где $\mathcal{L}(\cdot, \cdot)$ некоторая метрика, например евклидова, а $\mathscr{S}$ – рассматриваемое множество функций, например нейросети заданной архитектуры.

Постановка задачи

Описание образа

Задача описания образа есть задача построения наилучшего текстового описания для данного образа по изображениям его элементов. Полагая функцию оценки известной получаем:

- **Дано:**
  $\{X_i\} = O_n,\ |O_n| = n$ — образ, где $X_i = (I_i, \emptyset),\ I_i \in \mathcal{I},\ i = \overline{1, n}$ — его элементы с пустым текстовым описанием.
  $O_n(T) = \{X_i(T) = (I_i, T)\}_{i=1}^{n}$ для некоторого общего описания T
- **Требуется:**
  Найти оценку $\hat{T}$ общего для всех элементов описания T, максимизирующую значение функции оценки $\mathcal{S}(O_n)$, т.е.:

$$\hat{T} = \operatorname*{argmax}_{T} \mathcal{S}(O_n(T))$$

В данном случае задача генеративная — лучшее описание находится, из решения задачи максимизации функции оценки, а не просто выбирается из заранее заданного конечного множества.

Постановка задачи

Дополнение (восстановление) образа

Задача дополнения образа – есть задача выбора в некотором смысле наилучшего набора элементов из $\mathcal{X}$ для дополнения данного образа:

- **Дано:**
  $O_n \in \mathcal{O},\ |O| = n$
  $k \in \mathbb{N},\ k > n$ — количество недостающих элементов
  $J \subset \mathbb{N},\ |J| = m \geqslant k$ — индексы категорий недостающих элементов
  $\{\hat{T}_i\}_{i=1}^{k}$ — текстовые представления недостающих элементов, возможно пустые. В случае если предлагается только текстовое описание всего образа T, рассматривается $\forall i \in \overline{1,k}:\ \ T_i = T$.
- **Требуется:**
  Найти наилучшее в смысле максимизации функции оценки дополнение образа O_n до $O_k \in \mathcal{O}, |O_k| = k$ элементами из категорий $\{C_j\}_{j \in J}$, т.е. решить следующую задачу:

Постановка задачи

Дополнение (восстановление) образа

$$\{\hat{X}_i\}_{i=1}^k = \underset{\{X_i\}_{i=1}^k \in \bigcup\limits_{j \in J} \mathcal{X}_{C_j}}{\operatorname{argmax}} \Big[\alpha \cdot \mathcal{S}\left(O_n \cup \{X_i\}_{i=1}^k\right) +$$

$$+(1-\alpha) \cdot \sum_{i=1}^k S_X((I_i, T_i), (I_i, \hat{T}_i)) \cdot \mathbb{I}\{\hat{T}_i \neq \emptyset\}\Big]$$

здесь $X_i = (I_i, T_i)$, $\alpha \in [0, 1]$. Второе слагаемое отвечает за соответствие предсказанного элемента предъявленному текстовому представлению и равно нулю, если представление пусто. Задача, в отличие от предыдущей, дискриминативная и может быть решена точно полным перебором всех объектов из $\mathcal{X}$.

Постановка задачи

Задача генерации состоит в выборе образа произвольного размера, наиболее подходящего к предоставленному текстовому описанию. В терминах введенных выше получаем:

- **Дано:**
  T — текстовое описание образа.
- **Требуется:**
  Найти наилучший в смысле максимизации функции оценки образ $O \in \mathcal{O}$ элементы которого наилучшим образом соответствуют предложенному описанию T, т.е.:

$$\hat{O} = \underset{k, O=\{X_i\}_{i=1}^{k}, O \in \mathcal{O}}{\operatorname{argmax}} \left[\alpha \cdot \mathcal{S}(O) + (1-\alpha) \cdot \sum_{i=1}^{k} S_X((I_i, T_i), (I_i, T)) \cdot \mathbb{I}\{\hat{T} \neq \emptyset\} \right]$$

здесь $X_i = (I_i, T_i)$, $\alpha \in [0, 1]$. Стоит заметить, что если зафиксировать $k = 1$, получаем обычную задачу поиска наиболее подходящего под описание элемента в коллекции. Учитывая то, что множество образов определено множеством элементов и определением образа, задача генерации образа также является дискриминативной и состоит в переборе всех возможных образов.

Теоретическая часть

- Элементы равнозначны и задачи симметричны к перестановке $\Longrightarrow$ разумно рассматривать операции эквивариантные относительно группы перестановок.
- Опустим вопрос выбора способа получения латентных представлений элементов образа и будем использовать полученные с помощью предобученной модели
- В качестве функции оценки возьмем модель из статьи[1].
- Рассмотрим задачу дополнения образа. В случае пустого описания недостающих элементов без заранее определенных категорий, с учетом формальной постановки получаем задачу:
  - **Дано:**
    $O_n \in \mathcal{O},\ |O| = n$ — исходный образ
    $k \in \mathbb{N},\ k > n$ — количество недостающих элементов
  - **Задача:**

$$\{\hat{X}_i\}_{i=1}^k = \underset{\{X_i\}_{i=1}^k \subset \mathcal{X}}{\operatorname{argmax}}\ \mathcal{S}\left(O_n \cup \{X_i\}_{i=1}^k\right)$$

[1] https://doi.org/10.48550/arXiv.2204.04812

Дополнение образа

Дискретный подход

- Один из подходов к задаче – приближенный перебор.
- В таком случае, можно рассмотреть полный граф на вершинах $\mathcal{X} \setminus O_n \cup \{X_{init}\}$, где X_{init} – дополнительная начальная вершина. Тогда задача эквивалентна максимизации *оценки* пути $X_{init}, X_1, X_2 \dots X_k$ в таком графе. Где под *оценкой* пути понимается оценка образа $O_n \cup \{X_1 \dots X_k\}$
- Жадные бейзлайны:
  - $X_1 = \operatorname*{argmax}_{X \in \mathcal{X}} \mathcal{S}(O_n \cup X), \ \dots, X_k = \operatorname*{argmax}_{X \in \mathcal{X} \setminus \bigcup_{i=1}^{k-1} X_i} \mathcal{S}(O_n \cup X)$
  - $X_1 = \operatorname*{argmax}_{X \in \mathcal{X}} \mathcal{S}(O_n \cup X), \ \dots, X_k = \operatorname*{argmax}_{X \in \mathcal{X} \setminus \bigcup_{i=1}^{k-1} X_i} \mathcal{S}(O_n \cup X_1 \dots X_{k-1} \cup X)$
- Предложение: использовать алгоритм beam-search, активно применяемый в языковых моделях. Beam-search в граничных случаях вырождается либо в полный перебор, либо во второй из двух жадных алгоритмов выше.

Дополнение образа

Непрерывный подход

- Задача дополнения — задача дискретной оптимизации. Однако:
  - Фукнция оценки по постановке непрерывна по всем элементам и дифференцируема почти всюду (задана нейросетью)
  - Есть доступ не только к значению функции оценки, но и градиенту по любому параметру в любой точке
- Идея: заменим дискретную задачу непрерывной следующего вида:

$$\{\hat{X}_i\}_{i=1}^k = \underset{\{X_i\}_{i=1}^k \subset \mathbb{R}^d}{\operatorname{argmax}} \mathcal{S}\left(O_n \cup \{X_i\}_{i=1}^k\right)$$

- Полученная задача разрешима за разумное время какой либо модификацией стохастического градиентного спуска
- При некоторых ограничениях на функции активации функция оценки к тому же липшицева
- Таким образом, в случае довольно богатой доступной коллекции $\mathcal{X}$ стоит ожидать, что выбор $X_i = \underset{X \in \mathcal{X}}{\operatorname{argmin}}\, \rho(X, \hat{X}_i),\ i = \overline{1, k}$, где ρ – некоторая мера близости между X и $\hat{X}_i$, позволит получить хорошее приближенное решение исходной задачи

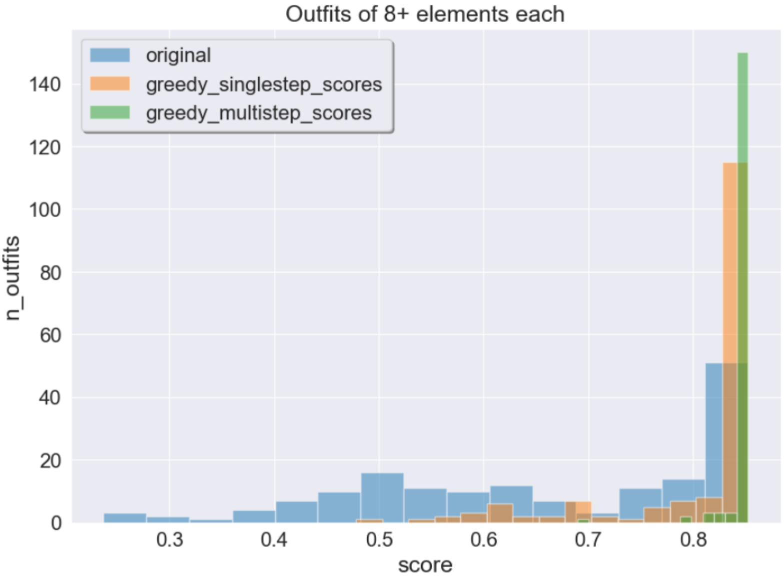
Базовый эксперимент

Жадные алгоритмы

- Датасет Polyvore, 17000 образов из 65000 объектов
- Рассмотрим небольшое подмножество
- Считаем распределение оценок исходных образов
- Из каждого образа случайным образом удаляем 2 элемента
- Рассматриваем дополнение получившихся образов 2-мя элементами жадным перебором

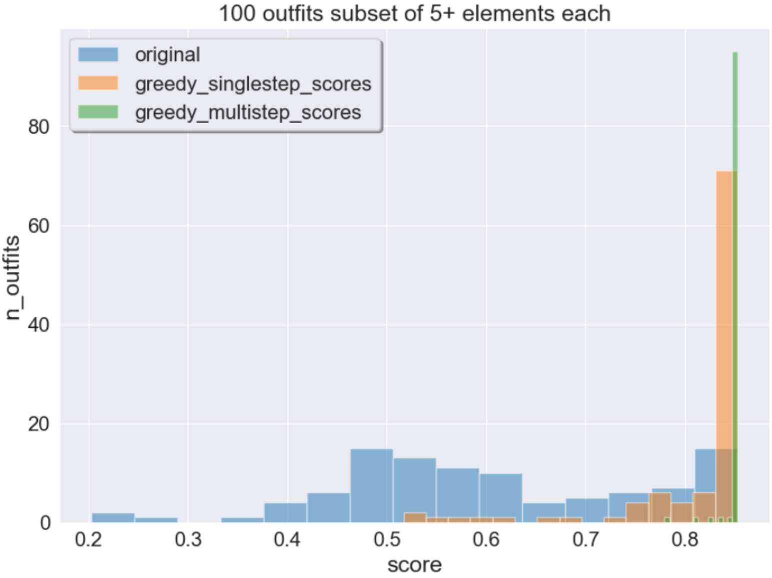
Базовый эксперимент

Жадные алгоритмы



Базовый эксперимент

Жадные алгоритмы



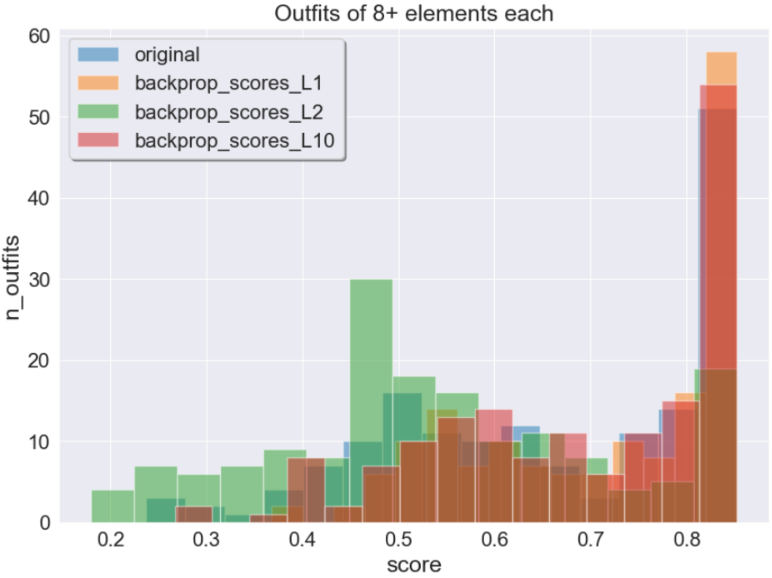
Базовый эксперимент

Непрерывная аппроксимация

- Рассматриваем те же самые образы, удаляем из каждого 2 элемента и рассматриваем дополнение получившихся образов
- Замораживаем веса модели оценки
- Добавляем 2 обучаемых эмбединга для недостающих элементов
- С помощью оптимизатора Adam получаем эмбединги, максимизирующие оценку
- Выбираем из всей коллекции элементы, ближайшие к полученным по некоторой метрике, в данном случае взяты L_1, L_2, L_{10}
- Вычисляем оценку получившегося образа

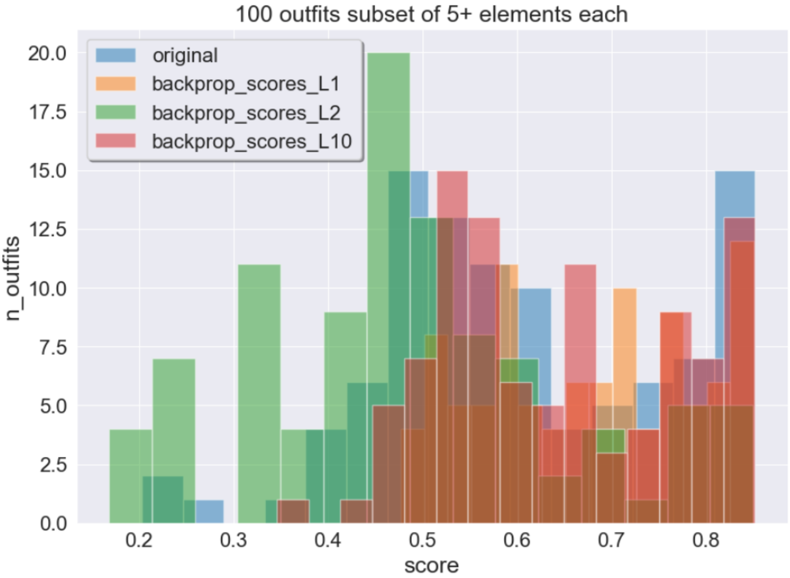
Базовый эксперимент

Непрерывная аппроксимация



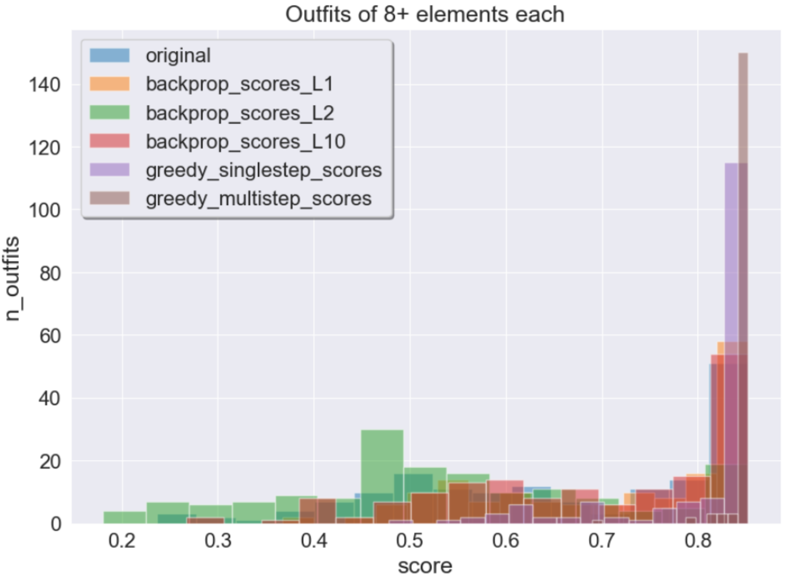
Базовый эксперимент

Непрерывная аппроксимация



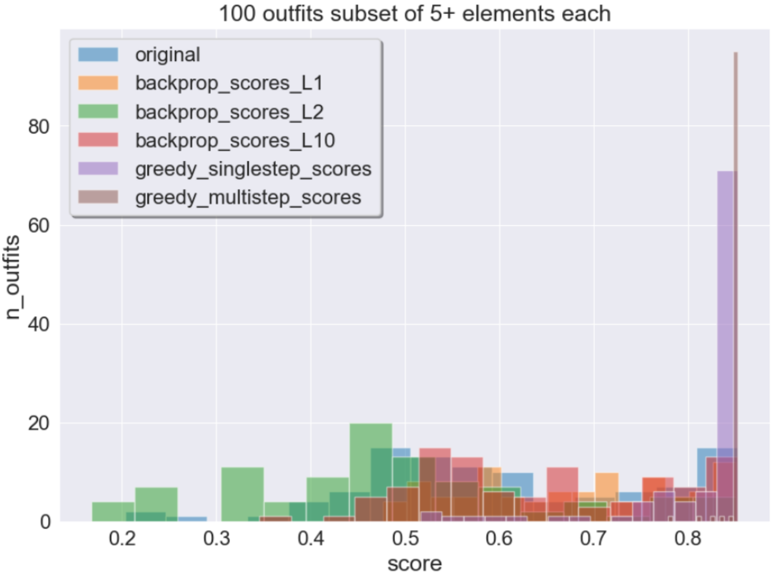
Базовый эксперимент

Сравнение



Базовый эксперимент

Сравнение



Базовый эксперимент

Выводы

- Жадные алгоритмы показывают хороший результат, но вычисления крайне не эффективны и занимают слишком много времени
- Метод непрерывного восстановления векторных представлений недостающих элементов серьезно уступает жадным
- Структура пространства представлений элементов слишком сложна, чтобы простые метрики близости позволяли выбрать лучший элемент коллекции
- Предлагается рассмотреть возможности агрегации представлений всех элементов перед выбором ближайшего для учета структуры пространства и взаимодействия элементов между собой.
- Агрегация может быть обучаемой. С учетом симметрии задачи, предлагается рассмотреть графовую нейронную сеть
- Исходя из постановки задачи, необходимо рассмотреть способы поощрения инвариантности к порядку выбора элементов
- Для обучения в дальнейшем можно применять элементы выбранные жадным образом, поскольку более точное решение задачи вряд ли достижимо за разумное время.