Е. В. Верховский, Г. А. Зорин, Т. Ю. Чернышева
Тюменский государственный университет, г. Тюмень
УДК 004.853

ПРИМЕР РЕАЛИЗАЦИИ МЕТОДА ПОИСКА ИЗОБРАЖЕНИЙ ПО ТЕКСТУ С ПОМОЩЬЮ НЕЙРОННОЙ СЕТИ

Аннотация. В данной работе описывается реализация модели поиска изображений по тексту на естественном языке с помощью метода "Dual Encoder" на русском языке. Приведено исследование качества получившейся модели.

Ключевые слова: поиск изображений, нейронные сети, компьютерное зрение, трансформер.

Введение. Поиск изображения — это классическая проблема области компьютерного зрения. Он осуществляется тремя подходами к поиску: контекстный, контентный и контекстно-контентный [1].

Контекстный поиск заключается в анализе текстовой информации, сопровождающей изображение, например, теги и текст к изображению. К этому подходу применяются следующие методы: boolean retrieval (булевой поиск), vector space retrieval (векторный поиск), probabilistic retrieval (вероятностный поиск) [1] и NLP (нейронная обработка естественного языка).

Контентный метод представляет из себя поиск по содержанию изображения. Его можно разделить на следующие методы: вариационные ряды и поиск по векторам (локальные, особых точек, хеш-функции и нейросетевые).

Вариационные ряды представляют из себя описания изображения по фрагментам используя гистограммы и ориентированные градиенты. Разделяются на алгоритмы описания прямоугольных (R-HOG) и круглых (C-HOG) фрагментов изображений [2]. Поиск по векторам можно разделить по способу получения вектора. Вектор особых точек — это вектор, выделенный из особых точек, полученных, например, гауссовым масштабированием. Самый известный метод формирования вектора признаков на основе «ключевых» точек является SIF (Scale-Invariant Feature Transform) [3]. Вектор на основе хеш-функции — это вектор, полученный с помощью алгоритма параметризации изображения, основанного на преобразовании изображений. Нейросетевой вектор — это метод получения вектора с помощью машинного обучения, используя сверточные сети или vision transformer [4].

Контекстно-контентный — это объединение векторов предыдущих подходов или же получение векторов с помощью Dual Encoder.

Dual Encoder (двойной кодировщик) также известный как "two-tower" — это особый тип поиска на основе внедрения, где одна башня глубокой нейронной сети производит внедрение запроса, а вторая вычисляет кандидата на встраивание. Вычисление скалярного произведения между двумя векторами внедрения определяет, насколько похож кандидат к запросу [5]. Обучается проекционная голова видео- и текстового кодера, чтобы вектора были схожи. Благодаря этому можно осуществлять поиск на естественном языке.

Проблема исследования. Проблема — существующие методы поиска изображений по тексту ориентированы на английский язык, что усложняет взаимодействие с сервисами, использующими эти методы, для пользователей, не знающих английский язык на должном уровне. Задача заключается в исследовании метода поиска изображений по тексту на естественном языке с помощью метода "Dual Encoder" и реализации его для поиска на русском языке.

Материалы и методы. Для реализации поиска изображений по тексту с помощью двойного кодировщика выбран датасет MS COCO, который содержит 5 аннотаций к каждому изображению. Данный набор представляет собой zip файл с 77 тысячами изображений и Json-файл, содержащий имя изображения и аннотации. Решено перевести его на русский с помощью переводчика DeepL API и создать из этого базу данных для обучения модели.

Для исследования использована модель нейронной сети с двойным кодировщиком для поиска изображений с использованием естественного языка представленная TensorFlow, которая вдохновлена CLIP. Идея модели состоит в том, чтобы обучить видео-кодер и текстовый кодер совместно для проецирования векторов изображений и их подписей в одно и то же пространство так, чтобы вектора подписей были расположены рядом с векторами изображений, которые они описывают. Для основы используются готовые видео-кодер и текстовый кодер, а обучается проекционная голова, размещая векторы в одно пространство и размерность (рис. 1).

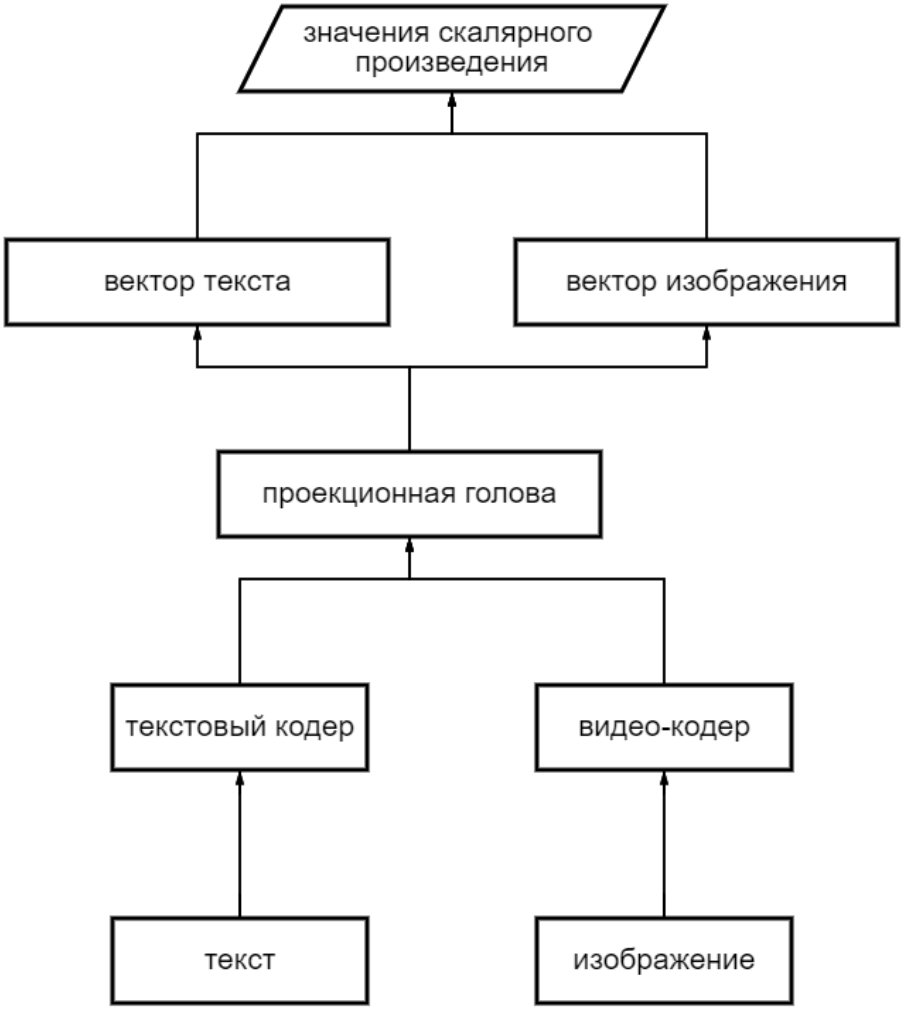

Рис. 1. Асимметричный "Dual Encoder" с устройством встраивания замороженных кодировщиков [4]

Основу для видео-кодера составляет Xception от Keras Applications, для текстового кодера — BERT от google. Объединяем их в проекционную голову для преобразования изображения и текста в одно и то же пространство вложения с одинаковой размерностью.

Вычисление потерь будет осуществляться следующим образом. Для этого авторами вычисляется сходство попарного скалярного произведения между каждой аннотацией и предсказанным изображением как прогноз. Целевое сходство между аннотацией и предсказанным изображением вычисляется как среднее значение сходства скалярного произведения между аннотацией и предсказанной аннотацией и сходства скалярного произведения между изображением и предсказанным изображением. Затем используется кроссэнтропия для вычисления потерь между целями и прогнозами.

При обучении замораживаются базовые кодеры для текста и изображений, обучаем только проекционную голову. Модель обучалась на 60 000 пар изображений — текст и 5 эпохах.

Результат. Чтобы оценить модель с двойным кодировщиком, используем подписи в качестве запросов. Прогноз засчитывается как верный, если для данной подписи соответствующее изображение извлекается в пределах k лучших совпадений. Для тестирования взято k = 10. Результат составил 7.61% и 3.36% для тренировочных и вариационных выборках соответственно.

Чтобы уточнить результат, проведены эксперименты с аннотациями, которых нет в датасете. Для этого созданы вектора изображений, используя получившуюся модель, из них создаем набор данных. Используя текстовый энкодер, получаем вектора запросов, которые будут написаны. Для каждого запроса будем получать 9 соответствующих изображений. Данное количество удобно для визуализации и достаточно информативно.

Самые интересные результаты получены из запросов: «дикие животные стоят в поле», «грибы растут в лесу», «корова на льду», «корова стоит на берегу океана», «птица сидит у воды». В ходе эксперимента оказалось, что модель в основном обращает внимание на отдельные слова по типу: лес, поле, человек, животное, корова и т. д. Например, запрос «птица сидит у воды» выдал правильно 9 вариантов. Запрос «корова стоит на берегу океана» выдавал все фотографии с океаном, и только на двух были крупные животные, а именно лошади. В датасете нет фотографии коровы, стоящей на берегу. На запрос «корова на льду» выдавались изображения с коровами и лошадьми, либо лыжниками. Это свидетельствует о сильном влиянии слова «корова» на поиск, а также связь слова «лед» с «лыжником» и «снегом». В датасете нет фотографии коровы на льду, и вообще нет слова «лед». А в последнем результате на запрос «грибы растут в лесу» все изображения содержат только лес. В датасете нет фотографии грибов, растущих в лесу, и слово «грибы» фигурирует как продукт среди других продуктов (мясо, овощи, рыба).

Заключение. Реализованная модель для поиска изображений в тестировании качества поиска в пределах значения k показала низкий результат, который составил 7.61% и 3.36% для тренировочных и вариационных наборов данных соответственно. Но в ответ на запросы из модели находятся приблизительные изображения. В основном поиск получился по словам, которые имели больший вес: корова, лес, лед, океан и т. д. Поэтому можно сказать, что модель может проводить приблизительный поиск изображений в массиве данных.

Для улучшения качества модели можно увеличить количество эпох обучения, объем датасета изображений и аннотаций, участвующих в обучении, но это все приведет к значительному увеличению затрачиваемых вычислительных ресурсов на обучение. Из других способов улучшения качества можно привести замену базовых видео- и текстовых кодеров, а также улучшить качество самих аннотаций.

СПИСОК ЛИТЕРАТУРЫ

1. Папулин С.Ю. Поиск изображений с использованием семантических признаков: специальность 05.13.01 «Системный анализ, управление и обработка информации (по отраслям)»: дис. … канд. техн. наук / С.Ю. Папулин. — 2015. — 214 с. — Текст: непосредственный.
2. Barskaya G.B. Creation of a painting dataset for use in artificial intelligence tasks / G.B. Barskaya, T.Y.Chernysheva, I.A. Krupkin, A.A. Lesiv.- Direct text // Third International Conference on Optics, Computer Applications, and Materials Science (CMSD-III 2023). — Washington, 2024. — P. 1306502.
3. Прозоров Д.Е. Применение легковесной сиамской нейросети для формирования вектора признаков в системе васкулярной аутентификации / Д.Е. Прозоров, А.В. Земцов. — Текст: непосредственный // Компьютерная оптика. — 2023. — Т. 47, № 3. — С. 433-441.
4. Lowe D.G. Distinctive Image Features from Scale-Invariant Keypoints / D.G. Lowe // International Journal of Computer Vision. — 2004. — Vol. 60, No. 2. — P. 91-110.
5. Dong Z. Exploring Dual Encoder Architectures for Question Answering / Z. Dong, J. Ni, D. Bikel, E. Alfonseca, Y. Wang, C. Qu, I. Zitouni // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. — 2022. — P. 9414-9419.