УДК 004.891

ПОЛУЧЕНИЕ ЭМБЕДДИНГОВ ИЗОБРАЖЕНИЙ В ПРОСТРАНСТВЕ ЭМБЕДДИНГОВ СЛОВ

А. А. УСАТОВ[1], канд. техн. наук А. М. БЕЛОЦЕРКОВСКИЙ[2], д-р техн. наук, проф. А. М. НЕДЗЬВЕДЬ[1]
[1](Белорусский государственный университет, г. Минск)
[2](Объединённый институт проблем информатики Национальной академии наук Беларуси, г. Минск)

Аннотация. *В работе предлагается использовать сочетание подходов свёрточной нейронной сети и векторизации слов для получения эмбеддингов изображений. Приведён пример архитектуры предлагаемого подхода. В результате описывается модель, строящая эмбеддинги изображений в пространстве эмбеддингов слов. Это позволяет реализовать текстовый поиск по изображениям и не требует разметки изображений всеми возможными для использования в поиске словами.*

Ключевые слова: *эмбеддинг, эмбеддинг изображения, эмбеддинг слова, семантический вектор, свёрточная нейронная сеть, Word2Vec, GloVe, FastText, embedding, semantic vector.*

Введение. При работе с изображениями часто возникает необходимость в получении вектора, описывающего изображение. Это может быть семантический вектор или эмбеддинг. Для полученных векторов можно, к примеру, ввести меру близости для фильтрации почти одинаковых фото. Также возникает ситуация, когда картинке нужно сопоставить не только класс изображения на ней, но и некоторое слово. Такое может возникать, например, при текстовом поиске по изображениям. Если используется модель классификации, то поиск может производиться только по заранее заданным словам. Если в выборке был размечен класс «машина», то найти «автомобиль» не получится без дополнительных преобразований. В работе описывается подход, позволяющий получать эмбеддинги изображений, имеющие синтаксический смысл в том же пространстве, что и эмбеддинги слов.

Основные подходы к обработке изображений и слов. Для работы с изображениями чаще всего применяются свёрточные нейронные сети, а для векторизации слов можно воспользоваться такими подходами, как Word2Vec, GloVe или FastText. Преимущество перечисленных подходов для векторизации в том, что вектора описывают «смысл» слова, а не просто его кодируют, это продемонстрировано на рисунке 1.

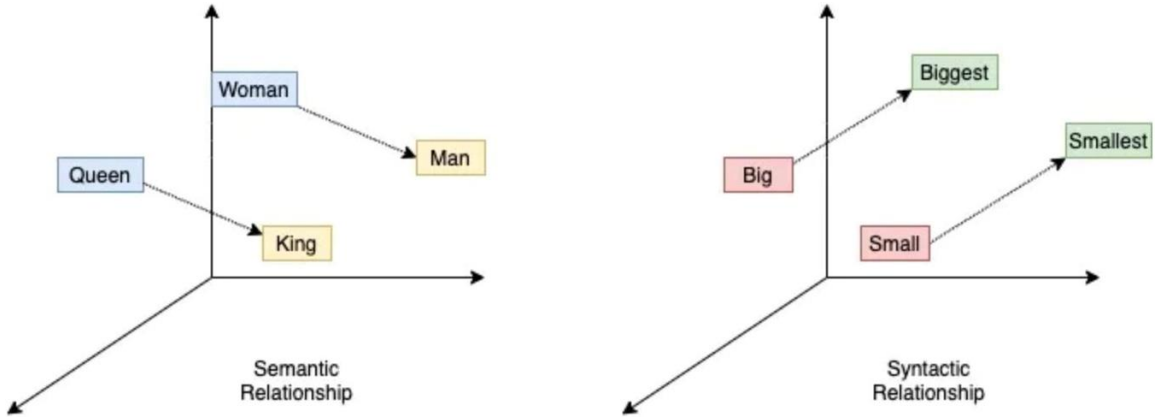

Рисунок 1. – «Смысл» в векторах, полученных Word2Vec

В работе в качестве модели обработки изображений подразумевается классическая свёрточная нейронная сеть (пример архитектуры приведён на рисунке 2) для изображений и Word2Vec для слов, но обе эти модели можно заменить на другие, решающие аналогичную задачу. В частности, можно использовать ResNet для изображений и FastText для слов, однако выбор оптимальных архитектур требует дополнительных экспериментов. Стоит отметить, что для реализации поиска по изображениям, вероятно, не стоит использовать слишком «тяжёлые» модели, поскольку объём обрабатываемой информации крайне велик.

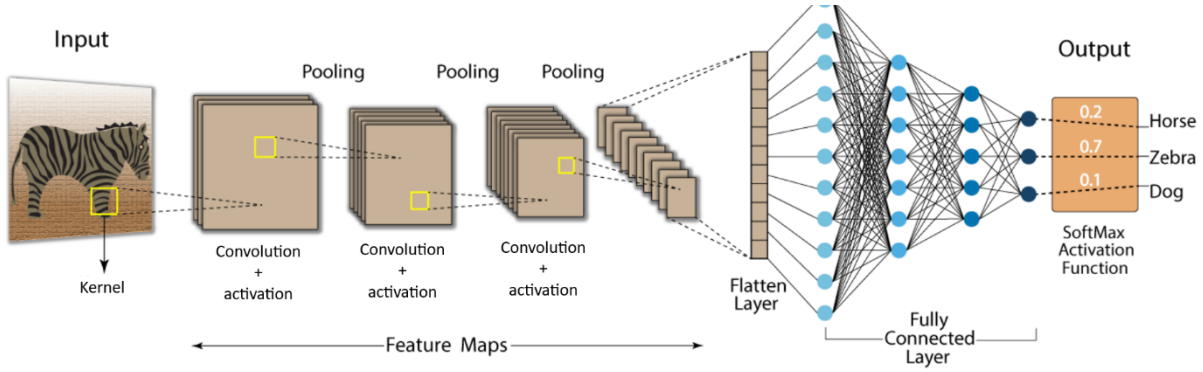

Рисунок 2. – Классическая архитектура свёрточной нейронной сети

Получение эмбеддингов изображений в пространстве эмбеддингов слов. Чтобы решить описанную в предыдущем разделе проблему, в качестве выхода модели можно использовать не вектор, кодирующий класс, а семантический вектор слова, описывающего класс. Получить этот семантический вектор можно моделью векторизации слов, такой как Word2Vec или FastText. Благодаря этому эмбеддинги изображений и эмбеддинги слов будут находиться в одном семантическом пространстве. Модели векторизации слов обучаются без учителя, что позволяет использовать обучающие выборки очень большого размера без затрат на разметку. Кроме того, существуют подходы векторизации слов, которые используют

н-граммы, что позволяет им работать даже со словами, которых не было в обучающей выборке. Примером такого подхода является FastText. Схема предлагаемого подхода изображена на рисунке 3.

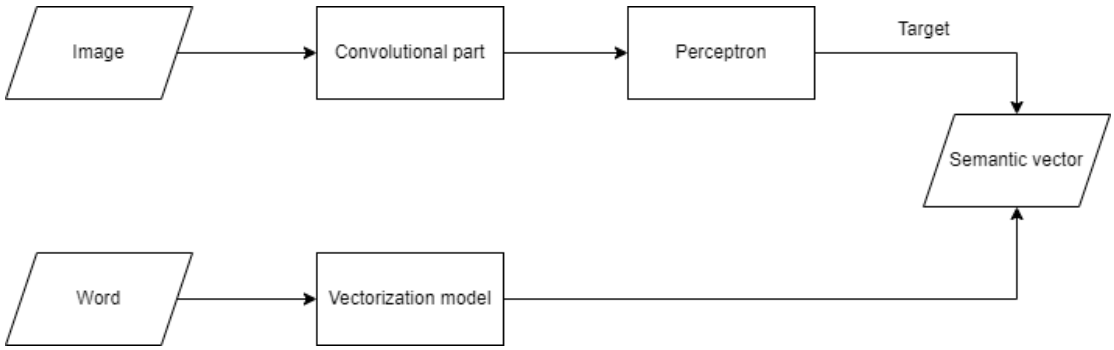

Рисунок 3. – Использование семантических векторов слов в качестве целевой переменной

Возможные улучшения подхода:

1. Поверх эмбеддингов можно добавить классификатор, чтобы не терять возможности сети классифицировать изображения.

2. Можно использовать подход нескольких голов в сети. Тогда сеть будет одновременно обучаться и классифицировать изображения, и аппроксимировать эмбеддингами семантические вектора.

Заключение. Как правило, получаемые из изображений эмбеддинги трудноинтерпретируемы. В работе описывается подход, позволяющий получать эмбеддинги изображений, имеющие синтаксический смысл в том же пространстве, что и эмбддинги слов, что может быть полезно в широком круге задач. В частности, это может применяться при текстовом поиске по изображениям.

ЛИТЕРАТУРА

1. Yann Lecun, Leon Bottou, Y. Bengio, Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. – 1998.
2. Недзьведь, А. М. Анализ изображений для решениязадач медицинской диагностики / А. М. Недзьведь, С. В. Абламейко – ОИПИ НАН Беларуси, 2012. – 240 с.
3. Y. Cai et al., "YOLOv4-5D: An Effective and Efficient Object Detector for Autonomous Driving," in IEEE Transactions on Instrumentation and Measurement, vol. 70, 2021, pp. 1-13.
4. Vapnik V. Principles of Risk Minimization for Learning Theory // Advances in neural information processing systems. – 1992.