СРАВНИТЕЛЬНЫЙ АНАЛИЗ АРХИТЕКТУР RNN ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ ПРИ ОБРАБОТКЕ ВИДЕО

Курочка К. С., Башаримов Ю. С.
Кафедра информационные технологии факультета автоматизированных и информационных систем
Гомельского государственного технического университета имени П.О. Сухого
Гомель, Республика Беларусь
E-mail: kurochka@gstu.by, basharymauyury@gmail.com

Выбор архитектуры нейронной сети для решения задачи классификации и распознавания образов является весьма важным этапом, влияющим как на скорость обучения сети, так и на точность и уровень ошибок полученной модели. В данной работе проводится сравнительных анализ различных архитектур реккурентных нейронных сетей на примере решения задачи классификации объектов в видеопотоке.

Введение

Классификация видео является сложной задачей, требующей обработки большого количества данных и выявления скрытых закономерностей. В последние годы нейросети стали неотъемлемой частью инструментария для решения задач классификации видео. Архитектуры нейросетей, такие как Convolutional Neural Networks (CNN) и Recurrent Neural Networks (RNN) проявили себя как мощные инструменты для обработки видеоданных и выявления их характеристик [1, 2].

I. Обзор основных архитектур

В последние годы было разработано множество архитектур нейросетей, специально предназначенных для классификации видео [1]. CNN – это популярная архитектура нейросети для классификации видео. Она извлекает пространственные признаки из кадров видео и моделируют их зависимости. Для классификации видео с помощью CNN сначала нужно подготовить данные для обучения и тестирования, а затем выбрать или создать подходящую архитектуру CNN для видео. После этого нужно обучить модель на тренировочных данных, минимизируя функцию потерь, и протестировать модель на тестовых данных, вычисляя метрики оценки. В результате модель сможет предсказывать метки классов действий для тестовых фрагментов видео.

LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit) – это два типа (RNN) (см. рис. 1), которые хорошо подходят для моделирования последовательных данных, таких как текст или речь. Они способны запоминать долгосрочные зависимости между элементами последовательности и избегать проблемы затухания или взрыва градиента. Эти архитектуры также могут быть адаптированы для классификации картинок и видео.

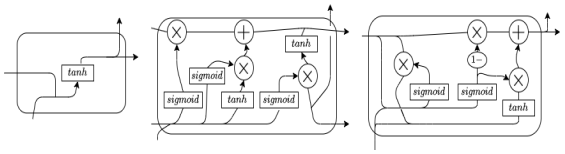

Рис. 1 – Слева на право RNN LSTM GRU

Для обработки видеоданных можно использовать LSTM или GRU по-разному [3]: либо к выходам других нейросетей, извлекающих пространственные признаки, либо к пикселям или патчам напрямую. Это позволяет учитывать как пространственную, так и временную информацию. Но эти подходы имеют недостатки: вычислительную сложность, большой объем данных и сложность интерпретации [4]. Поэтому нужно выбирать архитектуру нейросети в зависимости от задачи.

Для улучшения производительности и точности классификации видео, часто применяются комбинированные архитектуры, которые объединяют свойства CNN и RNN. Например, архитектура Convolutional Recurrent Neural Network (CRNN) [5] (см. рис. 2) комбинирует свёрточные слои CNN для извлечения пространственных признаков с рекуррентными слоями RNN для моделирования временных зависимостей в видео.

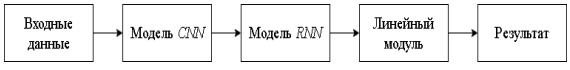

Рис. 2 – Общая архитектура модели CNN RNN

Такой подход заключается в одновременном выполнении сегментации и классификации. Это называется моделью CNN RNN, специально разработанной для задач прогнозирования последовательности с пространственными входными данными, такими как изображения или видео. Эта архитектура включает в себя использование слоев CNN для извлечения признаков на входных данных в сочетании с RNN для выполнения предсказания последовательности векторов признаков.

Свёрточный автоэнкодер сжимает и восстанавливает изображения или видео. Его можно

использовать вместо CNN в CRNN, подавая на вход скрытое представление изображения или видео. Это дает более абстрактные и семантические признаки, но требует больше усилий для обучения и стабилизации. Модель CNN RNN сочетает компьютерное зрение и обработку естественного языка и решает сложные задачи, такие как преобразование видео.

3D-архитектуры нейросетей представляют собой расширение CNN или RNN для работы с трехмерными пространственными и временными данными видео [6]. Они позволяют учитывать изменения во времени и пространстве одновременно, что может быть полезным для классификации видео (см. рис. 3).

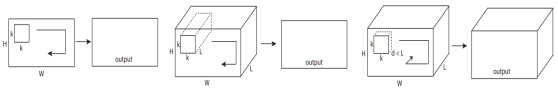

Рис. 3 – Слева на право 2D CNN, 2D CNN с 3D данными и 3D CNN

Архитектура представляет видео как объемные данные и расширяет CNN до третьего, временного измерения заменяя 2D-фильтры 3D-эквивалентами. В отличие от 2D CNN, 3D CNN принимают последовательность видеокадров в качестве входных данных. 3D-архитектуры нейросетей имеют преимущества перед CNN в классификации видео. Они учитывают пространственно-временные зависимости, изучают временные паттерны, улучшают точность классификации и могут быть компактными. Однако требуют больше ресурсов и данных, и их настройка сложнее.

II. Сравнительный анализ

Оценим производительность архитектур нейронных сетей на тестовом наборе данных Kinetics (табл. 1) [7, 8].

Таблица 1 – Сравнение разных архитектур нейросетей

Архитектура	Набор-данных	Точ-ность (%)	F1-мера (%)	Ско-рость (кадро-в/сек)
RNN	Kinetics	51.01	50.9	15.1
2D CNN	Kinetics	48.91	48.8	30.1
2D CNN с 3D данными	Kinetics	46.11	46.0	25.1
3D CNN (C3D)	Kinetics	51.62	51.5	10.2
3D CNN (I3D)	Kinetics	71.13	71.0	53
CRNN	Kinetics	51.01	50.9	47.1

III. Заключение

В завершение, выбор между 3D и 2D нейросетями зависит от задачи, данных и ресурсов. Для распознавания действий в видео лучше 3D CNN или CRNN, которые учитывают пространственные и временные зависимости. Для маленьких данных лучше классические архитектуры. 3D-архитектуры требуют больше ресурсов, чем CRNN. Для классификации видеопотока предлагается CRNN с автоэнкодером вместо CNN. Это ускоряет обучение и повышает качество. Параметры модели: размер изображения – 64 x 64, размер видео - 16 кадров, размер скрытого слоя – 128, размер mini-batch – 32, скорость обучения – 0.001, количество эпох – 50, функция потерь – перекрестная энтропия, оптимизатор – Adam.

Список литературы

1. Z. Zou, Z. Shi, Y. Guo, and J. Ye Object Detection in 20 Years: A Survey // March 2023 Proceedings of the IEEE. — IEEE, 2023. -– С. 257–276.
2. Kurachka K. S., Tsalka I. M. Vertebrae detection in X-ray images based on deep convolutional neural networks //2017 IEEE 14th International Scientific Conference on Informatics. — IEEE, 2017. — С. 194–196.
3. Олисеенко, В. Д., Абрамов, М. В., Тулупьев , А. Л. Нейронные сети lstm и gru в приложении к задаче многоклассовой классификации текстовых постов пользователей социальных сетей // Вестник ВГУ. Серия: Системный анализ и информационные технологии. — СПБ, 2021. -– С. 130–141.
4. Kurochka K., Panarin K. Algorithm for real-time binary classification of adenomas and norms images obtained by confocal microscopy //15th International Conference Mechatronic Systems and Materials, MSM 2020. -– 2020. -– С. 9202107–9202107.
5. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description // IEEE Trans Pattern Anal Mach Intell. – IEEE, 2017. -– С. 130–141.
6. Tran Du, Bourdev L. Learning Spatiotemporal Features with 3D Convolutional Networks // 2015 IEEE International Conference on Computer Vision (ICCV). – IEEE, 2015. -– С. 4489-4497.
7. Yue-Hei Ng J., Hausknecht M Beyond Short Snippets: Deep Networks for Video Classification // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – IEEE, 2015. -– С. 4694–4702.
8. Simonyan K, Zisserman A Very Deep Convolutional Networks for Large-Scale Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – IEEE, 2014. -– С. 3578-3584.