Разработка свёрточной нейронной сети для решения задачи семантической сегментации

Р. Р. Отырба, email: rostislav.otyrba.97@gmail.com
А. А. Сирота, email: sir@cs.vsu.ru

Воронежский государственный университет

Ключевые слова: *Данное исследование посвящено разработке конкурентоспособной, мощной и в то же время достаточно легковесной модели семантической сегментации, основанной на свёрточных нейронных сетях. В результате тщательного исследования ряда новейших технологий глубокого обучения и архитектур свёрточных нейронных сетей была спроектирована и построена U-Net-подобная архитектура вида «кодер-декодер». Была построена кодирующая сеть, имеющая пирамидальную структуру, эффективно извлекающая признаки с разными масштабами. В конце кодирующей сети, был построен модуль ASPP, эффективно извлекающий признаки с ещё большим масштабом. После модуля ASPP была построена легковесная декодирующая сеть с пространственным модулем внимания, которая идентична структуре сети в классической архитектуре U-Net. Построенная модель была обучена на Pascal VOC 2012 и тестировалась на валидационных данных.*

Ключевые слова: *Семантическая сегментация, глубокое обучение, свёрточные нейронные сети, обработка изображений*

Введение

Семантическая сегментация – одна из самых фундаментальных и популярных тем исследований в области компьютерного зрения, обладающая широким спектром применений от распознавания медицинских патологий до автономного вождения транспорта. Задача семантической сегментации тесно связана с задачей классификации изображений, поскольку она представляет собой процесс разбиения изображения на сегменты определённых категорий с одновременной их пиксельной классификацией. Впервые это систематически было открыто в основополагающей работе [1], где авторы впервые продемонстрировали полностью свёрточную сеть (FCN), реализующую архитектуру «кодер-декодер» для задачи семантической сегментации. Кодирующая сеть фиксирует более высокую семантическую информацию, постепенной уменьшая карты признаков, в то время как декодирующая сеть постепенно восстанавливает сигнал до желаемого

результата, чтобы получить пиксельную классификацию. С тех пор FCN вдохновила многих исследователей последующих работ, а свёрточные нейронные сети и вид архитектуры «кодер-декодер» стал преобладающим в решениях задач данной области. В последнее время методы на основе трансформеров (SETR, SegFormer, MaskFormer, HRFormer, Segmenter и др.) продемонстрировали большой потенциал и превосходят методы на основе CNN и всё это благодаря мощным и тяжёлым кодирующим сетям с трансформерами.

Успешные современные сегментационные модели в целом обладают следующим характеристиками.

- Эффективная кодирующая сеть, извлекающая признаки.
- Извлечение признаков в разных масштабах. В отличие от задачи классификации изображений, которая в основном работает с одним объектом, семантическая сегментация требует обработки объектов разных размеров на одном изображении, поскольку представляет собой задачу пиксельной классификации.
- Модуль внимания (пространственный и канальный). Модуль позволяет уменьшить вычислительные затраты при обучении, которые тратятся на незначимые активации и обеспечивает лучшую обобщаемую способность сети. Достигается это путём взвешивания различных областей (или каналов) изображения, где наиболее важные области (или каналы) имеют больший вес. Самые современные методы, в том числе основанные на трансформерах используют пространственные механизмы самовнимания и не используют канальное внимание.
- Низкая вычислительная сложность. Это особенно важно при работе с изображениями высокого разрешения, например, с кадрами дистанционного зондирования или городских сцен.

Основываясь на данных идеях и характеристиках, а также вдохновившись идеями и технологиями моделей U-Net с Attention Gate, DeepLabV3+, SegNeXt, Inception ResNet-v2, ResNeXt и MobileNetV2 в данном исследовании преследуется попытка построить конкурентоспособную, мощную и в то же время достаточно легковесную модель семантической сегментации, основанную на свёрточных нейронные сетях.

1. Разработанная архитектура модели семантической сегментации

Предлагаемая архитектура свёрточной нейронной сети имеет U-Net-подобную структуру «кодер-декодер» с пропускными соединениями (рис. 1). Архитектура состоит из следующих основных частей.

- Кодирующая сеть с пирамидальной структурой, извлекающая признаки с разными масштабами, состоящая из 4 блоков.

– Модуль ASPP, который помогает учитывать разные масштабы карт признаков глубоких слоёв и в целом значительно улучшает качество сегментации.
– Декодирующая сеть, представляющая собой модификацию декодера архитектуры U-Net.

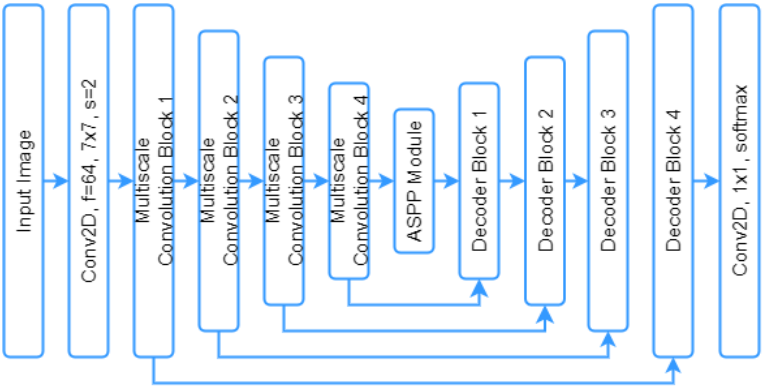

Рис. 1. Разработанная архитектура полностью свёрточной нейронной сети

Кодирующая сеть или кодер – это сеть, принимающая на вход изображение и извлекающая значимые признаки, необходимые для решения желаемой задачи. Многие исследователи в качестве кодирующих сетей используют предварительно обученные классификационные нейронные сети, однако они предназначены для задачи классификации изображений, но не семантической сегментации, которая в свою очередь требует обработки объектов разных размеров на одном изображении. В данном исследовании представляется кодер, каждый блок которого имеет пирамидальную структуру, фиксирующий признаки различного масштаба. Такой блок основан на идеях свёрточной нейронной сети Inception ResNet-v2 [2] и сети ResNeXt [3]. Кроме этого в сети используется функция активации Swish, которая по эффективности соответствует и порой превосходит ReLU, а также помогает предотвратить проблему исчезающего градиента.

Структура блока кодера представляет собой один большой Bottleneck, в котором сперва выполняется компрессия признаков, затем ряд параллельных вычислений как в семействах Inception и обратная декомпрессия (рис. 2). Такая структура позволяет снизить объём вычислений для достаточно тяжёлых операций, не теряя при этом обобщённости. При этом в качестве свёрточных слоёв используются

слои, которые делят каналы ровно на 32 группы, осуществляя затем параллельно для каждой группы свёрточные операции (авторы ResNeXt называют данный параметр мощностью и его использование дало значительный прирост к точности сетей исследователей). Таким образом, последовательность операций представляется следующим образом.

1. К входящим картам признаков применяется лёгкая точечная свёртка, создающая их компрессию, уменьшая количество каналов в определённое количество раза. Стоит отметить, что для каждого блока определяется свой коэффициент компрессии.
2. Полученный результат поступает в свёртку 5×5 для учёта локальной информации, после чего, результат отправляется в параллельные ветви из свёрточных слоёв с размерами фильтров 3×3, 7×7, 11×11 и 15×15 для учёта признаков в разных масштабах. Использование таких размеров основано на исследовании [4], где авторами утверждается, что использование больших фильтров является более эффективной стратегией, вместо использования набора маленьких распространённых фильтров 3×3.
3. Полученные результаты конкатенируются и проходят через точечную свёртку для их агрегации и декомпрессии в необходимое для данного блока количество каналов.
4. Результат суммируется с исходными картами признаков, образуя в данном блоке, таким образом, эффективное остаточное соединение (Residual Connection), как это реализовано в сетях семейства ResNet, чтобы избежать проблему исчезающего градиента и увеличить глубину сетей.
5. К полученным результатам применяется функция активации Swish.

Множество таких блоков образуют последовательность (глубину) и формируют Multiscale Convolution Block (MCB). После данного блока применяется блок понижающей дискретизации, представляющий собой параллельное осуществление операций свёрточным слоем с разделением по глубине (Depthwise Separable Convolution или DC2D) с шагом 2 и размером ядра 3×3, а также операции максимального объединения (MaxPooling2D). После этого, итоговая конкатенация проходит на вход следующего блока (MCB). Более подробная информация о характеристиках разработанной кодирующей сети представлены в таблице 1.

Таблица 1

Характеристики разработанной кодирующей сети

Блоки	**Выходная размерность**	**Кол-во каналов**	**Глубина блока**	**Компрессия каналов в Bottleneck**
Conv (7×7)	128 × 128	64	-	-
MCBlock-1	64 × 64	64	1	2
MCBlock-2	32 × 32	128	3	4
MCBlock-3	16 × 16	256	5	4
MCBlock-4	16 × 16	512	10	8

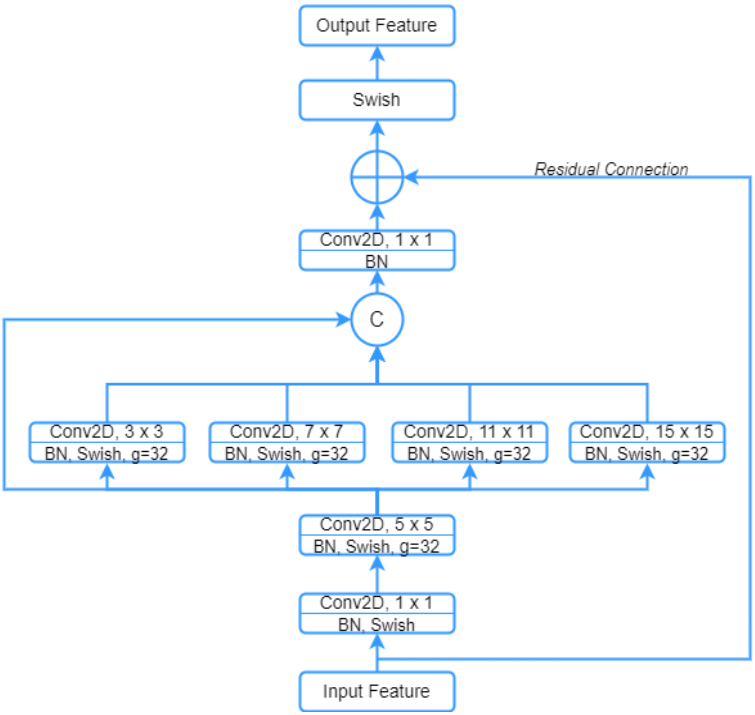

Рис. 2. Блок кодирующей сети из MCB

Разреженный пространственно-пирамидальный пулинг (ASPP Module) [5] изображённый на рисунке 3, изначально предложенный в моделях семейства DeepLab, на практике показал впечатляющие результаты. ASPP используется в конце кодирующей сети для карт

признаков низкого уровня и помогает учитывать большие масштабы объектов и в целом значительно улучшает качество сегментации. Достигается это путём использования и объединения свёрток с разделением по глубине DC2D с разными коэффициентами разреженности (6, 12, 18), а также точечной свёртки и глобального усредняющего пулинга (GlobalAvgPooling2D) для учёта глобального контекста. В данной работе был разработан практически идентичный модуль с небольшими изменениями.

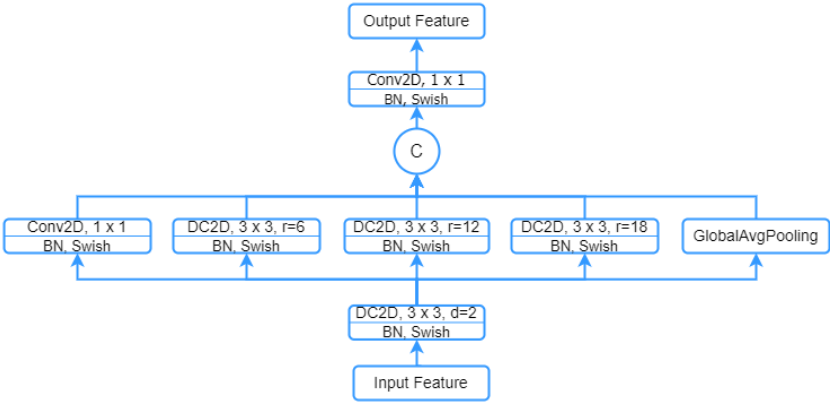

Рис. 3. Разреженный пространственно-пирамидальный пулинг (Module ASPP).

Декодирующая сеть или декодер – это сеть, принимающая извлечённые признаки кодирующей сетью, которые она использует в процессе восстановления пространственной информации, чтобы получить итоговую пиксельную классификацию. Декодирующая сеть, реализованная в данной работе по структуре идентична сети в классической архитектуре U-Net, однако содержит ряд значительных модификаций. Блок такой декодирующей сети представлен на рисунке 4.

Первой значительной модификацией является использование модуля внимания (Attention Block), впервые продемонстрированный в работе [6], но с незначительными модификациями. Как известно, в классическом U-Net пропускные соединения (skip-connections) комбинируют пространственную информацию соответствующих размеров из кодирующей и декодирующей сети, чтобы получить улучшенное представление признаков. Автор утверждает, что признаки, исходящие из кодирующих сетей довольно грубые, поэтому предлагается ввести модуль внимания на этих пропускных соединениях, для подавления активаций на нежелательных областях изображения.

Автор продемонстрировал высокую эффективность данного модуля, поэтому было решено его добавить.

Второй модификацией блока декодирующей сети является использование вместо тяжёлой стандартной пары свёрточных слоёв сети U-Net, легковесных свёрточных операций, используя Inverted Residual Block [7] с DC2D свёртками. Входная точечная свёртка увеличивает количество каналов до соответствующего количества для данного блока, умноженное на коэффициент 3. Затем осуществляется свёртка с разделением по глубине DC2D размером 3×3. Выходная точечная свёртка понижает количество каналов до соответствующего количества для данного блока. Такое резкое повышение каналов объясняется тем, что DC2D очень сильно снижают количество обучаемых параметров, что может привести к плохой обучаемости, поэтому данный блок вместо компрессии каналов, осуществляет их декомпрессию, что показало высокую эффективность на практике.

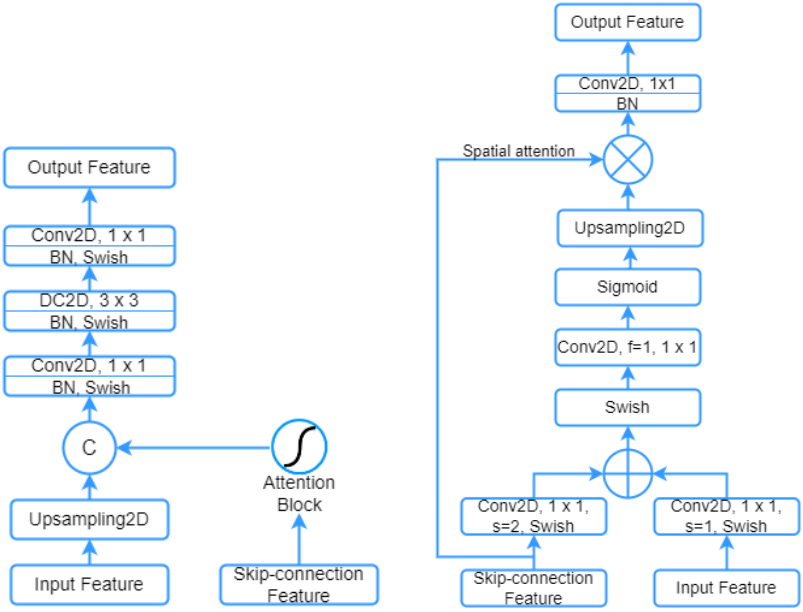

Рис. 4. Блок декодирующей сети (Decoder Block) и Блок внимания (Attention Block)

2. Проведённые эксперименты и результаты

Разработанная модель обучалась на известном эталонном наборе данных Pascal VOC 2012 [8] расширенной версии, имеющий 10582, 1449, 1456 изображений для обучения, валидации и тестирования соответственно. В данном исследовании модель тестировалась на валидационных данных, поскольку для тестовых данных ground truth отсутствует в открытом доступе. Набор данных состоит из 22 классов (фон, самолёт, велосипед, птица, лодка, бутылка, автобус, машина, кот, стул, корова, обеденный стол, собака, лошадь, мотоцикл, человек, горшечное растение, овца, диван, поезд, монитор, контур).

В таблице 1 представлены результаты обучения на проверочных данных в сравнении с другими популярными моделями по метрике средняя взвешенная степень пересечения изображений.

Таблица 2

Сравнение результатов тестирования с другими моделями

Архитектура	**Кодирующая сеть**	**mIoU val set**
FCN-8s	VGG16	62.2%
U-Net	Classic	72.7%
PSPNet	PSPNet	80.9%
DeepLabV3+	Xception-JFT	82.7%
Предлагаемая	Multiscale Convolution Encoder	80.5%

В процессе обучения была применена простейшая аугментация (случайная вырезка, отражение по горизонтали, случайное увеличение масштаба изображения на 0.25). Размеры входных изображений были преобразованы к 256×256, при необходимости для достижения необходимой размерности использовался lancoz5. Размер мини-пакета был установлен на 16. В качестве оптимизатора был выбран SGD Momentum. Скорость обучения определялась по методике описанная в работе [9]. Исследовав рекомендации автора, с помощью Learning Rate Finder осуществлялась одна эпоха со скоростью обучения от 1e-10 до 1, которая возрастает экспоненциально с каждой итерацией обучения. После этого, строился итоговый график зависимости ошибки от скорости обучения. Скорость обучения выбиралась в том промежутке, где она начинает стремительно уменьшаться (в данном случае 1e-7 до 1e-4). После чего с данными диапазонами скорости обучения применялся алгоритм циклической скорости обучения (Cyclical Learning Rate algorithm) вида triangular, который создаёт регуляризационный эффект и помогает ускорить процесс обучения. В качестве функции ошибки была использована комбинация взвешенной функции Дайса и Фокальной ошибки, где функция Дайса – степень соответствия между

предсказанием и полученным результатом, а фокальная ошибка — это модификация кросс-энтропии, занижающий вклад преобладающего класса и фокусируется на более сложных примерах. Модель обучалась в течении 66 тысяч итераций на платформе Kaggle. Количество параметров всей модели составляет 17.4 миллионов.

Заключение

Данное исследование посвящено разработке конкурентоспособной, мощной и в то же время достаточно легковесной модели семантической сегментации, основанной на свёрточных нейронных сетях. В результате тщательного исследования ряда технологий глубокого обучения и новейших архитектур была построена U-Net-подобная архитектура вида «кодер-декодер». Была построена кодирующая сеть, имеющая пирамидальную структуру на основе сети Inception ResNet-v2 и ResNeXt, извлекающая признаки с разными масштабами. В конце кодирующей сети, был построен модуль ASPP, эффективно извлекающий признаки с ещё большим масштабом. После модуля ASPP, была построена легковесная декодирующая сеть с пространственным модулем внимания, которая идентична структуре сети в классической архитектуре U-Net. В последующих исследованиях предстоит провести ряд эмпирических экспериментов для поиска наилучших характеристик архитектуры. Также для большей эффективности необходимо будет предварительно обучить кодирующую сеть на наборе данных ImageNet.

Список литературы

1. Long, J. Fully Convolutional Networks for Semantic Segmentation / J. Long, E. Shelhamer, T. Darrell // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2015. – P. 3431–3440.
2. Szegedy, K. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning / K. Szegedy [et al] // ArXiv – 2016.
3. Xie, S. Aggregated Residual Transformations for Deep Neural Networks / S. Xie [et al] // The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2016. – P. 5987–5995.
4. Ding, X. Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs / X. Ding [et al] // ArXiv – 2022.
5. Chen, L. Rethinking Atrous Convolution for Semantic Image Segmentation / L. Chen, G. Papandreou, F. Schroff, H. Adam // ArXiv – 2017.
6. Oktay, O. Attention U-Net: Learning Where to Look for the Pancreas / O. Oktay [et al] // ArXiv – 2018.

7. Sandler, M. MobileNetV2: Inverted Residuals and Linear Bottlenecks / M. Sandler [et al] // IEEE/CVF Conference on Computer Vision and Pattern Recognition – 2018. – P. 4510–4520.

8. Everingham, M. The Pascal Visual Object Classes (VOC) Challenge // International Journal of Computer Vision – 2010. – V. 88. – №2. – P. 303–338.

9. Smith, L. Cyclical Learning Rates for Training Neural Networks / L. Smith // IEEE Winter Conference on Applications of Computer Vision (WACV) – 2015. – P. 465–472.