САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Кафедра технологии программирования

Пугин Кирилл Витальевич

Отчет по научно-исследовательской работе

Методы сегментации медицинских МРТ снимков с помощью алгоритмов deep learning

Направление 01.03.02
«Прикладная математика и информатика»

Научный руководитель,
Кандидат технических наук,
доцент
Блеканов И.С.

Санкт-Петербург
2021

Содержание

Введение 3
1. Сегментация и ее применение 3
2. Актуальность задачи 4
3. Цели исследования 5
Глава 1. Постановка задачи 6
Глава 2. Обзор данных 7
Глава 3. Обзор методов 9
3.1. Модель U-Net 9
3.2. Модель TransUNet 10
3.3. Модель Swin-Unet 12
Глава 4. Результаты эксперимента 17
Заключение 18
Список литературы 19

Введение

1. Сегментация и ее применение

Сегментация изображения – это выделение на изображении классов, схожих по определенному критерию. Цель сегментации состоит в упрощении или изменении представления изображения, чтобы его было легче анализировать в дальнейшем. Результатом сегментации является множество сегментов, которые покрывают все изображение. Иначе говоря, каждый пиксель отмечен некоторой меткой некоторого класса.

Существует много областей пр именения сегментации: медицинские изображения, распознавание лиц, выделение объектов на спутниковых снимках, камерах и так далее. Методы сегментации изображений сильно разнятся, выделим несколько из них.

- Метод порогового значения (Thresholding method) – фокусируется на поиске пороговых значений на основе гистограммы изображения для поиска похожих пикселей;
- Метод на основе регионов (Region-Based method) – основан на разбиении изображения на однородные области;
- Метод водораздела (Watershed method) – основан на топологической интерпретации границ изображения;
- Нейронные сети (Neural networks) – основан на алгоритмах глубокого обучения (deep learning), в основном с использованием сверточных нейронных сетей (convolutional neural networks).

В рамках данной работы будет рассмотрена автоматическая сегментация **медицинских снимков магнитно-резонансной томографии (МРТ)** с помощью **нейронных сетей**. Стоит отметить, что подобный подход можно применить также и к **снимкам компьютерной томографии (КТ)**.

Эта задача по праву является одной из самых сложных и востребованных в анализе медицинских изображений из-за своей специфики и отличия от других областей применения методов сегментации. Снимки могут быть нечеткими, размытыми. Также на медицинских изображениях часто нужно сегментировать маленькие фигуры со сложной геометрией, вдобавок к этому цена ошибочной работы алгоритма довольно высока.

Если говорить о сложностях со стороны машинного обучения, то в данной сфере при разработке алгоритмов искусственного интеллекта всегда имеется острый недостаток референсных, размеченных данных. Это обусловлено тем фактом, что для разметки важных областей на снимках, например опухолей, необходимо привлечение медицинских специалистов, а это отнимает много времени и иногда может привести к ошибкам из-за человеческого фактора. Еще одна серьезная проблема в анализе медицинских изображений – несбалансированность классов. Очевидно, что МРТ-снимков здоровых людей куда больше, чем снимков людей с, например, диссекцией аорты.

2. Актуальность задачи

Несмотря на все трудности, описанные выше, сегментация часто играет ключевую роль в компьютерной диагностике различных заболеваний благодаря повышению точности диагностики и использование автоматизированных подходов полностью оправдано ввиду трудоемкости ручной обработки. Популярные задачи из этой области включают в себя сегментацию печени, опухолей печени [1], головного мозга, опухолей головного мозга [2], легких [3], сердца [4] и другие.

3. Цели исследования

В данной работе были поставлены следующие задачи:

1. Сделать обзор на тип и формат данных, использующихся в медицине, а затем и в машинном обучении;
2. Исследовать архитектуры *U-Net* [5], *TransUNet* [6], *Swin-Unet* [7] сверточных нейронных сетей для сегментации медицинских изображений;
3. Провести сравнение результатов для указанных выше моделей нейронных сетей на МРТ-снимках.

Глава 1. Постановка задачи

Постановка задачи сегментации медицинских изображений не отличается от постановки задачи сегментации любых других изображений.

Пусть дано множество изображений $\{I_i\}_{i=1}^{n} \in \mathbb{R}^{w \times h \times c}$, где w – ширина изображения, где h – высота изображения c – количество каналов изображения, и множество классов $C = \{1, \dots, k\}$ (например, целые органы, опухоли органов). Пусть каждому изображению I_i сопоставлена разметка (маска) $M_i \in \mathbb{R}^{w \times h}$ так, что каждому пикселю $w_i(x, y) \in I_i$ сопоставлена метка класса $M_i(x, y) \in C$.

Общая задача сегментации – для входного изображения I_i требуется восстановить разметку M_i, т.е. сегментировать на нем все классы, так, что полученный алгоритмом результат будет слабо отличаться от ручной разметки, в идеале получить более точную разметку. В терминах машинного обучения необходимо достигнуть наивысшей метрики качества (о метрике качества см. главу 4).

Глава 2. Обзор данных

Данные – это основа для построения методов машинного обучения и именно от их качества и количества зависит результат полученного алгоритма.

В данной работе в качестве опорного датасета был использован датасет из соревнования по машинному обучению ACDC (Automated Cardiac Diagnosis Challenge) [8]. Этот набор изображений включает в себя МРТ снимки 150 пациентов, которые были получены за 6 лет. 150 пациентов были разделены на две выборки: одна включает в себя 100 человек и отнесена к тренировочной, другая 50 и отнесена к тестовой. В тренировочной выборке для каждого пациента имеются снимки сердца, а также разметка. Разметка (маска) приведена экспертами-кардиологами для конечной систолической фазы и конечной диастолической фазы.

Цель соревнования состояла в двух задачах:

- Сравнить эффективность автоматических методов сегментации эндокарда и эпикарда левого желудочка сердца и эпикарда правого желудочка как для систолы, так и для диастолы
- Сравнить эффективность автоматических методов классификации пациентов по пяти классам;

Нас, конечно же, интересует первая из нах. Исходный формат файлов – *.nifti* (Neuroimaging Informatics Technology Initiative) и файлы хрянятся в 3D и 4D массивах, но с помощью библиотеки языка python *nibabel* мы можем преобразовать их в numpy array, чтобы работать с 2D изображениями. Ниже на рис. 1 представлен пример 2D среза вместе с его маской.

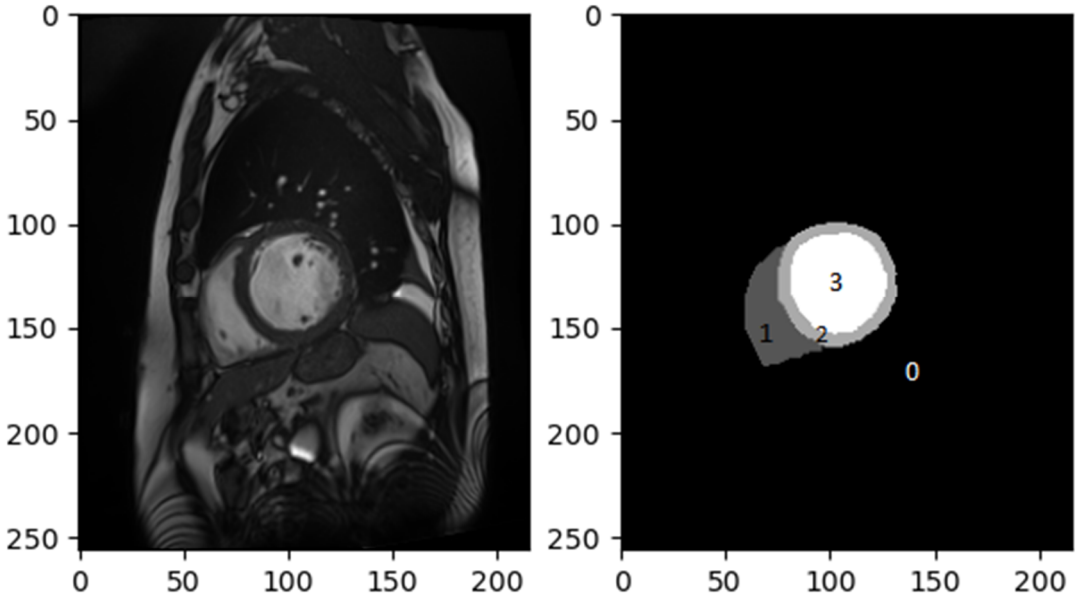

Рис. 1: пример МРТ снимка сердца из датасета ACDC. Слева чернобелое изображение в конечной диастолической фазе, справа четырехклассовая маска: 0 – фон (background), 1 – правый желудочек (RV), 2 – миокард (MYO), 3 – левый желудочек (LV)

Глава 3. Обзор методов

Исходя из изложенного выше, наша задача – обучить алгоритм автоматически сегментировать снимки на 4 класса, можно назвать этот этап предобработкой для дальнейшего обследования и диагностирования пациента (например, по сегментированным участкам желудочков сердца можно найти их объем). Для решения данной проблемы в настоящей работе будут исследованы 3 модели.

3.1. Модель U-Net

Стоит начать с обзора «базовой» модели глубокого обучения U-Net [5]. Изначально данная архитектура создавалась для сегментации изображений в биомедицинских приложениях, однако она нашла применение и в лругих областях. Архитектура U-Net включает в себя блок энкодера и блок декодера которые связаны между собой. Каждый слой энкодера понижает размер изображения в 2 раза, а количество признаков (features) повышает в 2 раза. Напротив же, каждый слой декодера повышает размер изображения в 2 раза, а количество признаков сокращается вдвое. При каждой операции повышения размера изображения (upsampling) каждый результат на выходе объединяется с соответствующим выходом слоя энкодера, чтобы дополнить промежуточную потерянную информацию о границах нужных классов на изображении. Общая архитектура U-Net представлена на рис. 2.

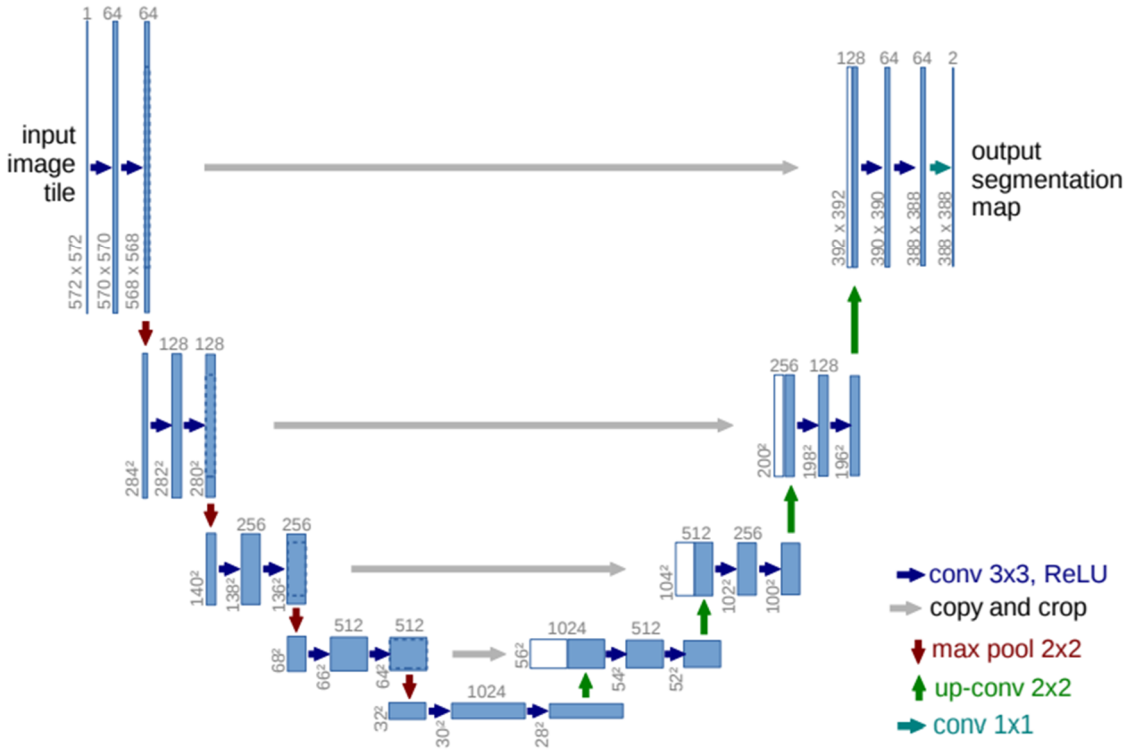

Рис. 2: архитектура U-Net для входного изображения с расширением 572x572

Несмотря на то, что архитектура относительно стара (статья вышла в 2015 году), U-Net по-прежнему является отличным стандартом, на результаты которого стоит опираться при построении более сложных моделей, что, собственно, и будет сделано в данной работе.

3.2. Модель TransUNet

В области обработки естественного языка (NLP) с 2017 года [9]. такая архитектура как Transformers де-факто является стандартом для многих задач: суммаризация текста, перевод, анализ сентимента и т.д. Однако же в задачах компьютерого зрения до недавних пор применение трансформеров оставалось ограниченым в виду специфичности задач. Но, в 2020 году, вдохновленные успехами Трасформеров в NLP авторы статьи *AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE* [10] попытались применить тот же

функционал непосредственно к изображениям. Как известно, Трансформеры в NLP работают с последовательностями токенов, исходя из этой идеи, авторы, разбивая картинки на маленькие части и подавая на вход Трансформерам их эмбеддинги абсолютно так же, как и в задачах NLP, смогли добиться state-of-the-art результатов в классификации изображений. Эту архитектуру, основанную на трансформерах, назвали *Vision Transformers (ViT).*

Модель *TransUNet* [6] появилась в результате совмещения Transformers и U-Net и уже успела показать высокие результаты в медицинских приложения, включая сегментацию нескольких органов и сегментацию сердца. Рассмотрим ее архитектуру подробнее.

Сначала исходное изображение подается на вход сверточной нейронной сети (CNN), которая извлекает признаки (feature extraction) и выдает карту признаков с меньшим разрешением по сравнению с исходным, но с большим количестов каналов. Пока что все так же, как и в части энкодера U-Net. Затем получившуюся карту признаков разбивают на патчи, как описано в [10], и каждый патч изображения с помощью линейной проекции переводится в эмбеддинг. Также вычисляется позиционный эмбеддинг (positional embedding) для сохранения пространственной информации на изображении. Результирующая последовательность векторов-эмбеддингов, которая подается на вход Трансформера, задается формулой (1):

$$\boldsymbol{z}_0 = \left[\boldsymbol{x}_p^1\mathbf{E}; \boldsymbol{x}_p^2\mathbf{E}; \ldots; \boldsymbol{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos}, \quad (1)$$

где $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot D) \times D}$ – patch embedding projection, $\mathbf{E}_{pos} \in \mathbb{R}^{N \times D}$ – positional embedding, $P \times P$ – размер патча, N – количество патчей, D – размерность скрытого пространства, в которое проектируются эмбеддинги.

Полученная на выходе из Трансформера hidden feature последовательность $\boldsymbol{z}_L \in \mathbb{R}^{\frac{HW}{P^2} \times D}$, где L – количество слоев Трансформера, преобразуется к размеру $\frac{H}{P} \times \frac{W}{P} \times D$ и подается на вход Cascaded Upsampler (CUP), который с помощью сверточных и upsampling блоков вкупе с skip-connections из CNN-энкодера преобразует

изображение к исходному разрешению $H \times W$, выдавая на выходе маску изображения. Полная архитектура TransUNet представлена на рис. 3.

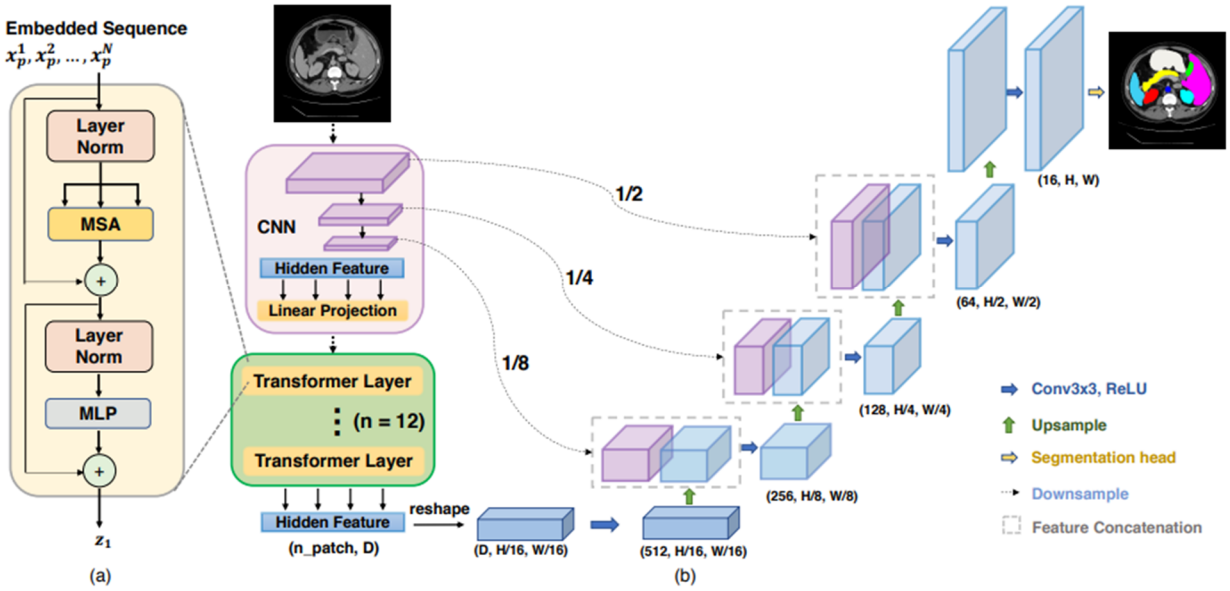

Рис. 3: структура TransUNet.

3.3. Модель Swin-Unet

Еще раз отметим, что архитектура *TransUNet* представляет собой гибрид Transformers и U-Net. То есть Трансформер используется только для извлечения закономерностей/признаков изображения после трех блоков сверточной нейронной сети, а все остальное остается таким же, как и в привычной архитектуре U-Net.

Архитектура *Swin-Unet* [7] – это U-Net подобный «чистый» Transfomer. «Чистый» в том смысле, что архитектура Swin-Unet полностью состоит из блоков Трансформера, в отличии от TransUNet.

Основой новейшей архитектуры для сегментации (12 мая 2021 г.) стал *Swin-Transformer* [11], который показал SOTA результаты в множестве задач компьютерного зрения: классификации, детекции, сегментировании, а к тому же Swin-Transformer имеет линейную вычислительную сложность в зависимости от размеров входного

изображения (к примеру, TransUNet и ViT имеют квадратичную вычислительную сложность).

Рассмотрим структуру Swin-Unet.

Encoder. Вначале, как и в случае с TransUNet, изображение с разрешением $W \times H$ разбивается на патчи размером 4×4. При таком подходе к разбиению размерность признаков (то есть размерность каналов изображения) каждого патча, в случае RGB изображения, становится равной $4 \times 4 \times 3 = 48$. Далее с помощью линейной проекции карты признаков проецируются в пространство произвольной размерности (пусть, C). Измененные патч-токены проходят через несколько последовательных Swin Transformer блоков и слоев слияния патчей (patch merging layers). Здесь можно увидеть сходство со структурой U-Net: patch merging layers используются для понижения размеров изображения (max pooling в U-Net), а Swin Transformer блоки выступают в качестве извлекателя признаков (сверточные слои в U-Net).

Decoder. Декодер состоит из Swin Transformer блоков и слоев расширения патчей (patch expanding layers). Также здесь присутствуют skip-connections между соответствующими Swin Transformer блоками из энкодера и декодера. В отличие от patch merging layers, patch expanding layers необходимы для «апсемплинга»: каждый такой слой увеличивает размер изображения в 2 раза, а последний слой расшерения патчей, вместе с линейной проекцией, увеличивает размер изображения в 4 раза, выдавая на выходе сети сегментированное изображение с исходными размерами $W \times H$.

Swin Transformer block. В отличие от традиционного multi-head self attention (MSA) модуля, который применяется в Трансформерах, Swin Transformer блок основан на сдвигающихся окнах (shifted windows). На рис. 4 представлены 2 последовательных Swin Transformer блока.

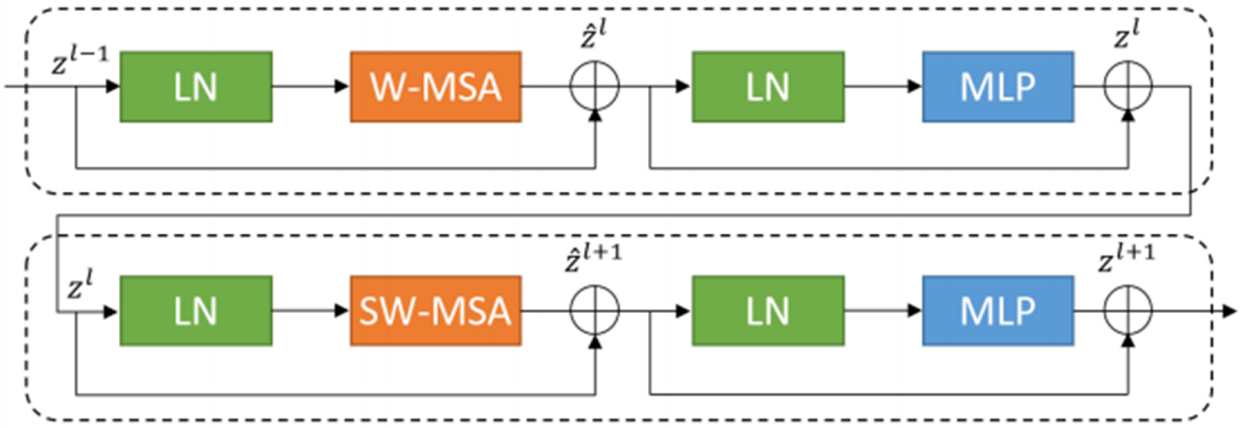

Рис. 4: Swin Transformer блоки *l* и *l+1*

Каждый блок состоит из слоя нормализации (LN), multi-head self attention модуля (W-MSA/SW-MSA), residual connection (+) и двухслойного перцептрона (MLP) с функцией активации GELU. Таким образом, два непрерывных Swin Transformer блока могут быть сформулированы следующим образом:

$$\widehat{z^l} = W\text{-}MSA\left(LN(z^{l-1})\right) + z^{l-1}, \tag{2}$$

$$z^l = MLP\left(LN(\hat{z}^l)\right) + \hat{z}^l, \tag{3}$$

$$\hat{z}^{l+1} = SW\text{-}MSA\left(LN(z^l)\right) + z^l, \tag{4}$$

$$z^{l+1} = MLP\left(LN(\hat{z}^{l+1})\right) + \hat{z}^{l+1}. \tag{5}$$

Shifted Window Self-Attention. Рассмотрим иерархию изображения в рамках этой модели. Изображение, состоящее из пикселей, разбивается на патчи размером $\mathrm{P} \times \mathrm{P}$, затем мы рассматриваем патчи как базовую единицу изображения (будто пиксели в исходном изображении) и разбиваем эти патчи окнами, имеющими размер $\mathrm{M} \times \mathrm{M}$. На рис. 5 приведен пример разбиения изображения на окна размером 4×4.

Рис. 5: Деление исходного изображения на непересекающиеся части – окна.

Применяя «обычный» MSA нам бы нужно было вычислять attention для каждого патча с каждым, что приводит к квадратичной вычислительной сложности. При Window-MSA подходе attention вычисляется только для патчей внутри одного окна, что уменьшает временные затраты. К сожалению, такой метод имеет недостаток: у нас нет связи между непересекающимися окнами, и модель не может найти закономерностей. Для решения данной проблемы предлагается использовать сдвиг окон (shifted window partioning).

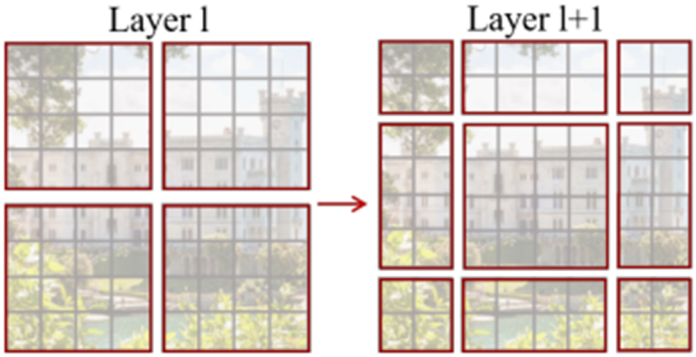

Рис. 6: Иллюстрация подхода со смещенным окном.

Как показано на рис. 6, l −й модуль использует обычную стратегию разбиения изображения на окна, которая начинается с верхнего левого пикселя, и $8 \times 8 = 64$ патчей равномерно разбиваются на $2 \times 2 = 4$ окна размером 4×4 (M $= 4$). После этого, $l + 1$-й модуль принимает конфигурацию окон, которая перемещает их на $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ пикселей относительно окон с обычным разделением (рис. 6 справа).

Окончательная Swin-Unet архитектура изображена на рис. 7.

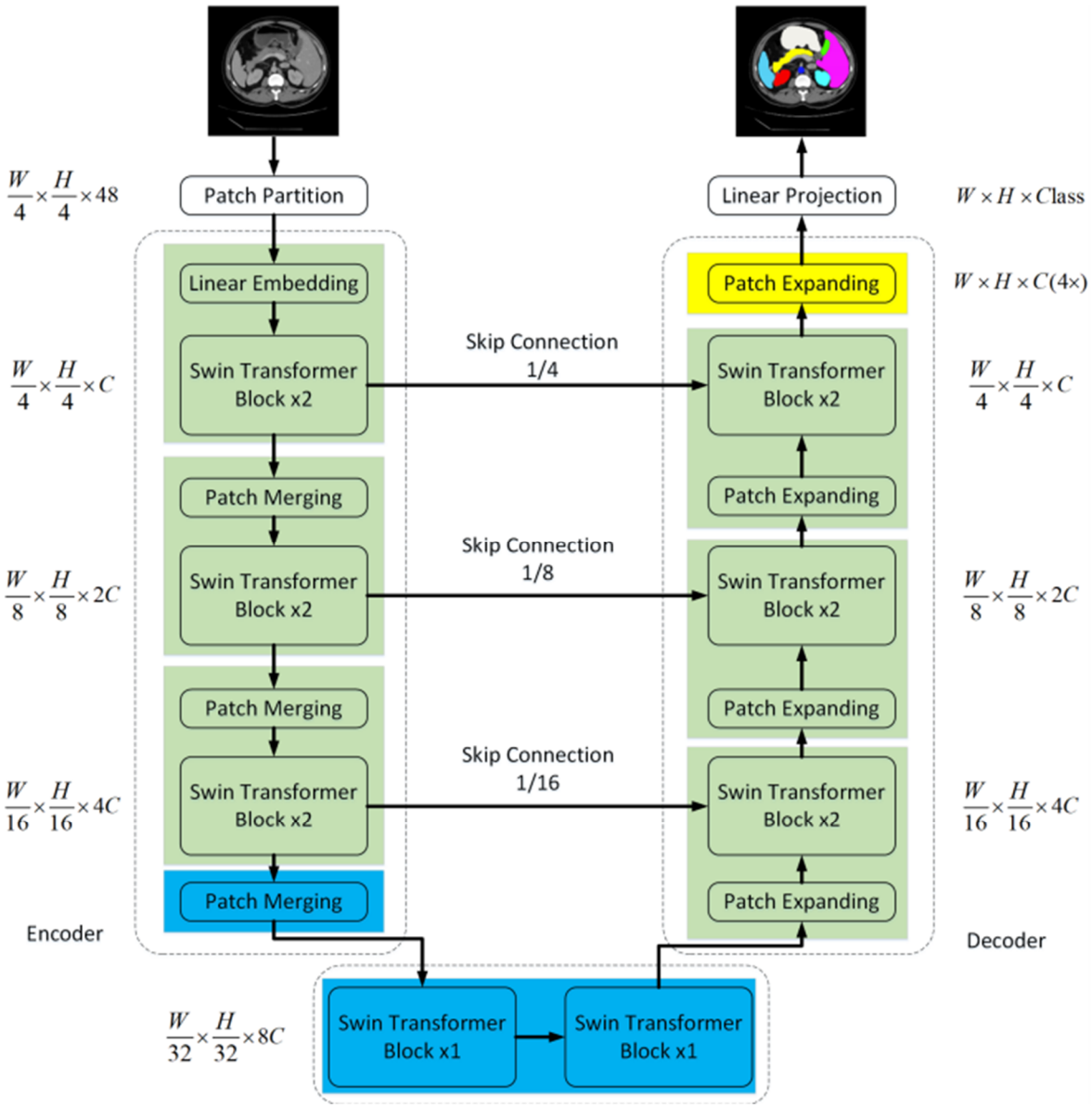

Рис. 7: архитектура Swin-Unet. Слева изображена структура энкодера, справа структура декодера. Обе логические части сети соединены между собой посредством skip connection.

Глава 4. Результаты эксперимента

Для тестирования моделей был использован ACDC датасет (см. главу 2). Исходная выборка из 100 пациентов, содержащая разметку для каждого среза, была разбита в отношении 70:10:20 на тренировочную, валидационную и тестовую соответственно. В качестве метрики качества использован коэффициент Дайса (Dice score), который задается следующим образом:

Пусть у нас есть размеченная маска $\boldsymbol{A}$ бинарного изображения (ground truth), где значение 1 означает, что в соответствующем пикселе находится нужный нам объект сегментирования, а 0 в другом случае. Пусть маска $\boldsymbol{B}$ бинарного изображения это выход нашего алгоритма, тогда Dice score (DSC) задается формулой (6):

$$\mathrm{DSC} = \frac{2 \cdot |A \cap B|}{2 \cdot |A \cap B| + |B \backslash A| + |A \backslash B|}, \tag{6}$$

Значения метрики, полученные на тестовой выборке, приведены в табл. 1

Метод	DSC	RV	LV	MYO
R50 + U-Net	87.55	87.10	80.63	94.92
TransUNet	89.71	**88.86**	84.53	95.73
Swin-Unet	**90.00**	88.55	**85.62**	**95.83**

Табл. 1: Точность сегментации различных методов на датасете ACDC. В столбцах RV, LV, MYO указан Dice score для сегментации правого желудочка, левого желудочка, миокарда сердца соответственно. В столбце DSC указан усредненный Dice score.

R50 + U-Net – архитектура, совмещающая ResNet50 [12] в качестве энкодера (feature extractor) и U-Net описанного в главе 3.

Заключение

Эксперимент, проведенный с использованием МРТ снимков датасета ACDC, показал, что новейшая архитектура Swin-Unet на текущий момент является state-of-the-art подходом в области сегментирования медицинских изображений.

Стоит также отметить, что, как и все архитектуры, использующие Трансформеры, Swin-Unet имеет потенциал на увеличение точности предсказания и обобщающей способности при увеличении количества данных для обучения.

Список литературы

1. Li W., Jia F., Hu Q. Automatic Segmentation of Liver Tumor in CT Images with Deep Convolutional Neural Networks // J. Comput. Commun. 2015. Vol. 03, № 11.
2. Cherukuri V. et al. Learning based segmentation of CT brain images: Application to postoperative hydrocephalic scans // IEEE Trans. Biomed. Eng. 2018. Vol. 65, № 8.
3. Wang S. et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation // Med. Image Anal. 2017. Vol. 40.
4. Chen C. et al. Deep Learning for Cardiac Image Segmentation: A Review // Front. Cardiovasc. Med. 2020. Vol. 7.
5. Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. Vol. 9351.
6. Chen J. et al. Transunet: Transformers make strong encoders for medical image segmentation // arXiv Prepr. arXiv2102.04306. 2021.
7. Cao H. et al. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation // arXiv Prepr. arXiv2105.05537. 2021.
8. Bernard O. et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? // IEEE Trans. Med. Imaging. 2018. Vol. 37, № 11.
9. Vaswani A. et al. Attention is all you need // Advances in Neural Information Processing Systems. 2017. Vol. 2017-December.
10. Dosovitskiy A. et al. An image is worth 16x16 words: Transformers for image recognition at scale // arXiv Prepr. arXiv2010.11929. 2020.
11. Liu Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows // arXiv Prepr. arXiv2103.14030. 2021.
12. He K. et al. Deep residual learning for image recognition // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016. Vol. 2016-Decem.