

ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ДЛЯ ФОРМИРОВАНИЯ ОБУЧАЮЩЕГО НАБОРА ПРИ СЕГМЕНТАЦИИ ЯДЕР КЛЕТОК РАКОВЫХ ТКАНЕЙ С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ ГЛУБОКОГО ОБУЧЕНИЯ

Сылунь Сьюй, В.В. Скакун

Белорусский государственный университет, Минск, Беларусь;

E-mail: xusilun@hotmail.com

Сегментация и подсчет ядер клеток на гистологических и иммуногистохимических изображениях срезов раковых тканей позволяет сделать оценку степени развития заболевания. Основным инструментом для выполнения сегментации изображений являются нейронные сети глубокого обучения. В данной работе произведено исследование применения метода главных компонент для расширения обучающего набора при сегментации ядер клеток на флуоресцентных изображениях раковых тканей молочной железы. Наилучшие результаты получены при добавлении первой главной компоненты с последующей стандартизацией данных.

Ключевые слова: глубокое обучение; изображения раковых клеток; предварительная обработка; метод главных компонент; U-Net.

ВВЕДЕНИЕ

В настоящее время цифровая обработка изображений стала важной частью медицинской диагностики и имеет широкий спектр применения. Анализ гистологических и иммуногистохимических изображений составляет основу диагностики многих видов рака. Важным шагом в процессе анализа таких изображений является сегментация ядер раковых клеток, целью которой является отнесение каждого пикселя к классу пикселей, образующему или область ядра, или фона. Искусственные нейронные сети, особенно сверточные нейронные сети (CNN), сегодня являются основным инструментом решения данной задачи.

Основной проблемой применения нейронных сетей к решению задачи сегментации является необходимость формирования сравнительно большого набора изображений, размеченных экспертами. Поскольку мы имеем ограниченный набор размеченных изображений, всего 9 изображений размером 2048×2048 пикселей, критическим является проведение предварительной подготовки данных. Нами уже было проведено исследование различных методов предварительной подготовки данных для обучения нейронной сети [1]. Наилучшим явился метод, заключающийся в выделении областей размером 256×256 пикселей с перекрытием наполовину как по вертикали, так и по горизонтали, с последующим выделением из каждой полученной области по 8 фрагментов размером 128×128 пикселей со случайным смещением. Поскольку в обучающем наборе вследствие

перекрытия имеется значительное число одинаковых фрагментов, для устранения эффекта переобучения необходимо применять аугментацию изображений.

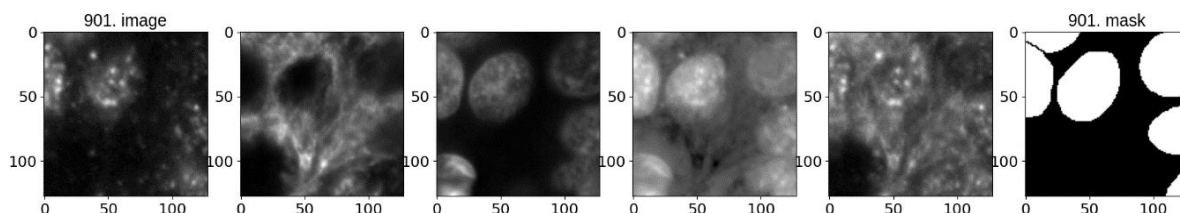
Исследования показали, что двухкратное увеличение полученной обучающей выборки путем повторной подачи изображений на вход нейронной сети (с выполнением аугментации) позволяет улучшить качество сегментации (с средним на 1-2% по метрике IoU) и не переобучить модель. Следовательно, подготовленный набор данных еще не является оптимальным. Одним из способов увеличения объема полезной информации является применение различного рода преобразований исходных данных. Если исходными данными являются изображения, то такими преобразованиями являются аффинные преобразования, изменение яркости, контрастности и т.д., входящие в набор методов аугментации изображений. К таким преобразованиям также можно отнести методы снижения размерности, в частности метод главных компонент (МГК). Применение МГК позволило значительно улучшить качество сегментации ядер раковых клеток на флуоресцентных изображениях с использованием классических методов сегментации изображений [2]. Так как мы работаем с трехканальными флуоресцентными изображениями раковых клеток, где вследствие перекрытия спектров флуорофоров имеется корреляция данных в соседних каналах, то добавление еще одного канала, являющегося первой главной компонентой, построенной на основе трех исходных каналов, может предоставить дополнительную информацию для модели и улучшить качество сегментации. Соответственно, целью работы является исследование применения МГК для расширения обучающего набора при сегментации ядер клеток раковых тканей с помощью нейронных сетей глубокого обучения.

РЕЗУЛЬТАТЫ

Исходными данными являются 9 флуоресцентных изображений срезов ткани опухоли молочной железы в формате RGB, полученные при 10-кратном увеличении с использованием эпифлуоресцентного инвертированного микроскопа Nikon TE200, оснащенного ПЗС-камерой Photometrics серии 300. Изображения 3-х канальные, 2048x2048 пикселей с разрешением 0,2 мкм/пиксель [2]. Синий краситель 4,6-диамидино-2-фенилиндол дигидрохлорид (DAPI) использовался для маркировки всех ядер. Красный цианиновый краситель Cy5 использовался для маркировки белка рецептора эстрогена, содержащегося в основном в ядрах раковых клеток. Белок цитокератин, содержащийся в цитоплазме раковых клеток, был помечен зеленым цианиновым красителем Cy3. Таким образом, маркерами раковых клеток являются два красителя, Cy5 и Cy3. Размеченные

экспертами изображения, на которых выделены только маски ядер клеток (ground truth, GT), представляют собой бинарные изображения, в которых пикселям ядер клеток присвоено значение 1.

В процессе предварительной подготовки данных, заключающейся в выделении областей размером 256×256 пикселей с перекрытием наполовину как по вертикали, так и по горизонтали с последующим выделением из каждой полученной области по 8 фрагментов размером 128×128 пикселей со случайным смещением, получено 12064 фрагмента размером 128×128 . Для каждого фрагмента независимо был применен МГК. Признаками в преобразовании являются каналы изображения и результатом применения МГК являются изображения той же размерности, что и исходное. Перед применением МГК была произведена стандартизация данных (приведение к нулевому среднему и единичной дисперсии). Так как мы имеем всего 3 признака и дисперсия, приходящаяся на главные компоненты, быстро падает, информативными являются только первые две компоненты. На рисунке показан результат применения МГК для фрагмента изображения раковых клеток размером 128×128 пикселей.



Фрагмент иммуногистологического флуоресцентного изображения раковых клеток размером 128×128 пикселей. Слева направо: красный канал, зеленый канал, синий канал, первая и вторая главные компоненты, размеченные маски ядер.

Таким образом, исходными данными для обучения являются уже 4 каналные изображения, где 4-ый канал есть 1 главная компонента, вычисленная на основе каналов исходных изображений. Были рассмотрены 4 варианта: 1) исходные данные (R,G,B); 2) исходные данные и первая главная компонента (R,G,B,PC1(RGB)); 3) исходные данные плюс первая главная компонента, вычисленная на основе нелинейных комбинаций компонент (G,B,PC1(G,B,G/B,B/G,1/G,1/B)), показавшая наилучшие результаты в работе [2]; 4) первые две главные компоненты (PC1(RGB), PC2(RGB)). Во всех вариантах проводилась стандартизация исходных данных (приведение к нулевому среднему и единичной дисперсии). Поскольку дисперсия первой главной компоненты более чем в 2 раза превышала дисперсию исходных стандартизированных данных, то дополнительно были исследованы варианты выравнивания дисперсий по всем каналам, исключая первую главную компоненту (обозначим их как 2* и 3*).

Для исследования применения МГК для расширения обучающего набора данных использовалась модель U-Net. Архитектура сети и ее параметры подробно описаны в работе [1]. После каждого слоя свертки в кодере и декодере дополнительно был добавлен слой Batch normalization. Все тесты проводились на компьютере, оснащённом графической картой RTX-3060 с объемом оперативной памяти 12 ГБ. Программная реализация выполнена на языке Python с использованием 2 версии фреймворка TensorFlow. Для всех рассматриваемых случаев размер выборки BatchSize был установлен в 32, количество итераций – 30. Для оценки качества сегментации выполнялся расчет двух метрик для тестовых данных: бинарное пересечение над объединением на уровне 0,5 (BIOU) и функция бинарной энтропии loss. Результаты исследования на основе 9 независимых экспериментов представлены в таблице. Наилучшие результаты были получены для вариантов 2 и 3. Наихудшим явился вариант с использованием только преобразованных данных. Добавление главной компоненты позволило значительно улучшить стабильность процесса обучения.

Средние значение и стандартные отклонения метрик BIOU и loss

Вариант	1	2	2*	3	3*	4
BIOU	0,876±0,003	0,878±0,001	0,877±0,003	0,878±0,001	0,877±0,004	0,845±0,005
Loss	0,153±0,007	0,151±0,003	0,154±0,005	0,150±0,001	0,153±0,008	0,197±0,008

ВЫВОД

Проведено исследование применения МГК для расширения обучающего набора при сегментации ядер раковых клеток с помощью нейронных сетей глубокого обучения. Наилучший результат был получен для обоих вариантов с добавлением первой главной компоненты и стандартизацией исходных данных (результаты сравнимы), что подтверждает идею о целесообразности применения МГК для расширения обучающей выборки.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Xu, S., Skakun, V. Comparison of Deep Learning Preprocessing Algorithms of Nuclei Segmentation on Fluorescence Immunohistology Images of Cancer Cells // PRIP 2021. Communications in Computer and Information Science, 2022, vol 1562. Springer, Cham. P. 166–177.
2. Lisitsa Y.V., Yatskou M.M., Apanasovich V.V., Apanasovich T.V. Algorithm for automatic segmentation of nuclear boundaries in cancer cells in three-channel luminescent images. // Journal of Applied Spectroscopy. 2015. 82(4) C. 598-607.