دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی
مهندسی کامپیوتر

عنوان:

بخش‌بندی تصاویر توده مغزی با استفاده از شبکه‌های عصبی ژرف

نگارش:

مهسا امانی

استاد راهنما:

دکتر شهره کسائی

بهمن ۱۴۰۱


بسم الله الرحمن الرحیم

سپاس

از استاد بزرگوارم، دکتر شهره کسائی که با کمک‌ها و راهنمایی‌های بی‌دریغشان، بنده را در انجام این پروژه یاری داده‌اند، تشکر و قدردانی می‌کنم. هم‌چنین از آقای فرنوش عارفی که در این مسیر نهایت همکاری را داشتند، صمیمانه سپاس‌گزارم.

چکیده

توده‌های مغزی جان بسیاری از انسان‌ها را تهدید می‌کنند. بخش‌بندی دقیق این توده‌ها و درمان آن‌ها می‌تواند کمک شایانی در نجات جان بسیاری از انسان‌ها داشته باشد، اما انجام این کار توسط تیم درمانی و پزشکان کار دشواری است. روش‌های مبتنی بر یادگیری ژرف می توانند دقت بخش‌بندی را افزایش دهند. شبکه‌های مختلفی همانند شبکه‌های عصبی پیچشی و شبکه‌های مولد تخاصمی برای بخش‌بندی توده‌های مغزی مورد استفاده قرار گرفته‌اند. چنین مدل‌هایی نیازمند منابع سخت‌افزاری و زمان زیاد برای همگرا شدن و یادگیری هستند. امروزه روش‌های یادگیری ماشین بر پایه معماری‌ای به نام «ترنسفرمر» در کاربردهای مختلف، همانند دسته‌بندی و بخش‌بندی، نتایج خیره‌کننده‌ای داشته‌اند. چنین شبکه‌هایی نیازمند دادگان زیاد برای عملکرد بهتر هستند. در این پروژه با استفاده از مجموعه دادگان BraTS شبکه‌ای بر پایه ترنسفرمر برای بخش‌بندی تصاویر سه بعدی توده‌های مغزی آموزش می‌دهیم و عملکرد آن را با امتیاز Dice می‌سنجیم. سپس با استفاده از تنظیم پارامترها و افزایش مجموعه دادگان در جهت بهبود مدل گام برمیداریم و با هر تغییر نتایج را مقایسه و ارزیابی می‌کنیم. در مجموع این تغییرات عملکرد مدل به میزان ۱۴ درصد در طی ۱۰ دور بهبود می‌یابد.

کلیدواژه‌ها: توده مغزی، بخش‌بندی معنایی، شبکه عصبی ژرف، ترنسفرمر، شبکه مولد تخاصمی، مجموعه دادگان BraTS

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱

مقدمه

در این فصل به شرح مختصری از مسئله‌ای که با آن روبرو هستیم، اهمیت پرداختن به این مسئله، ادبیات متداول استفاده شده در بررسی این مسئله و اهداف اصلی این تحقیق می‌پردازیم. در انتها نیز خلاصه‌ای از ساختار کلی این نوشتار و آنچه در فصل‌های آینده مورد بررسی قرار می‌گیرد، آورده شده است.

۱ـ۱ تعریف مسئله

یکی از انواع شایع توده‌ها، توده‌های مغزی هستند که تعداد آن‌ها به بیش از ۱۵۰ نوع می‌رسد. درحالیکه بخش‌بندی[۱] تعدادی از این توده‌ها به آسانی انجام‌پذیر است، برخی دیگر از این توده‌ها به علت پیچیدگی‌هایی که دارند، تجزیه و تحلیل دستی آن‌ها کاری بسیار زمانبر خواهد بود. امروزه با توجه به پیشرفت‌های وسیعی که در زمینه یادگیری ژرف انجام شده است، بینایی ماشین به یکی از فناوری‌های اصلی در زمینه تحلیل تصاویر پزشکی تبدیل شده است. امروزه سازوکار توجه[۲] به نقطه عطفی در عمده روش‌های یادگیری ژرف تبدیل شده است. شبکه‌های متنوعی بر پایه سازوکار توجه برای کاربردهای پردازش تصویر، به طور خاص بخش‌بندی، معرفی و پیاده‌سازی شده‌اند. تا پیش از ظهور سازوکار توجه و با توجه به دشواری‌های بخش‌بندی، عمده این شبکه‌ها نیازمند معماری پیچیده، ورودی‌های دستی، پیش‌پردازش[۳]

[1] Segmentation
[2] Attention Mechanism
[3] Preprocessing

و یا پس‌پردازش[۴] بودند. اما پس از معرفی سازوکار توجه این نیاز به طرز قابل توجهی کاهش پیدا کرد. در این پروژه به بررسی و ارزیابی نتایج بخش‌بندی توده‌های مغزی با استفاده از معماری‌ای بر پایه‌ی سازوکار توجه می‌پردازیم.

۱ـ۲ اهمیت موضوع

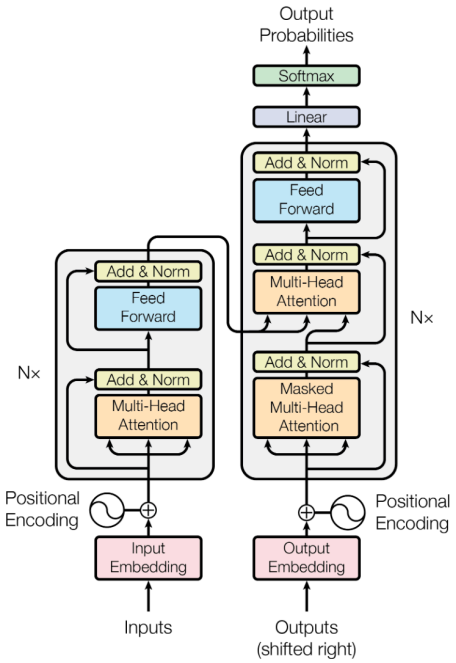
بخش‌بندی معنایی توده‌های مغزی یک امر اساسی در تجزیه و تحلیل تصاویر پزشکی است که می‌تواند به پزشکان در دسته‌بندی بهتر، تخمین نرخ رشد و برنامه‌ریزی درمان کمک شایانی داشته باشد. یکی از مشکلات این توده‌ها این است که می‌توانند در همه جای مغز و در هر شکل و اندازه‌ای رشد کنند و کار بخش‌بندی را از این هم دشوارتر بکنند. برای دست‌یافتن به این مهم و نجات جان بسیاری از انسان‌ها، نیاز داریم تا عمل بخش‌بندی را با نهایت دقت و صحت انجام دهیم. روش‌های یادگیری ژرف که از سازوکار توجه کمک می‌گیرند و بدون نیاز به مدل‌های پیچیده، پیش پردازش و یا پس‌پردازش خاص می‌توانند شیءها را با دقت قابل‌قبولی دسته‌بندی و بخش‌بندی کنند، باعث ایجاد تحول در عرصه پزشکی شده‌اند. با استفاده از این روش‌ها می توان در زمان کمتر، نتایج بهتر یا یکسانی را تولید کرد و به حفظ جان بسیاری از مردم کمک کرد.

۱ـ۳ ادبیات موضوع

ترنسفرمرها[۵] یک ساختار شبکه‌ای مبتنی بر سازوکارهای توجه برای ترجمه‌ی ماشینی[۶] هستند. با توجه به یک عنصر پرسش[۷] (به عنوان مثال، یک کلمه هدف در جمله‌ی خروجی) و مجموعه‌ای از عناصر کلید[۸](به عنوان مثال، کلمات منبع در جمله‌ی ورودی)، واحد چندسر ـ توجه[۹] محتوای کلید را با استفاده از وزن‌های توجه که با جفت پرسش و کلید سازگار است، جمع می‌کند. این مدل از یک رمزگذار[۱۰] و یک رمزگشا[۱۱] تشکیل شده است که هر کدام از این دو می‌توانند به تعداد دلخواه بلوک‌های پردازشی با

[4] Postprocessing
[5] Transformers
[6] Machine Translation
[7] Query
[8] Key
[9] Multi-Head Attention
[10] Encoder
[11] Decoder

ساختار مشخصی داشته باشند. معماری دقیق این شبکه در شکل ۱ ـ ۱ به تصویر کشیده شده است [۱].
ترنسفرمرها در ابتدا برای کارهای مرتبط با پردازش زبان طبیعی معرفی شدند، اما به تازگی راه به پردازش



شکل ۱ ـ ۱: معماری شبکه ترنسفرمر [۱].

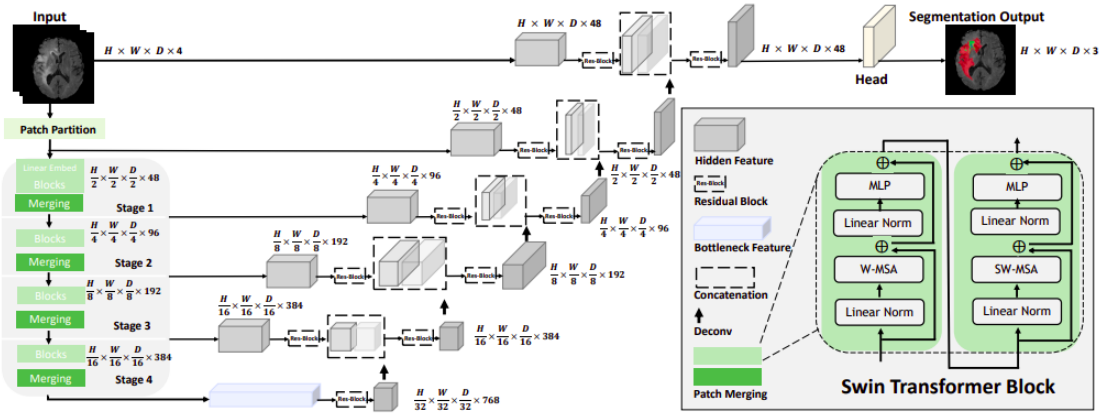
تصویر نیز برده‌اند و در وظایف مختلف همچون دسته‌بندی و بخش‌بندی بسیار موفق هستند. به طور خاص، شبکه ViT از وصله‌های[۱۲] دوبعدی تصویر، به عنوان یک توالی ورودی، مشابه توالی ورودی به یک مدل ترنسفرمر زبان طبیعی استفاده می‌کند. عملکرد این شبکه با شبکه‌های عصبی پیچشی[۱۳] برای دسته‌بندی تصویر در حالی که از قبل روی مجموعه داده‌های تصویر بزرگ آموزش دیده است، قابل مقایسه است [۲]. شبکه SwinUNETR نیز مدل دیگری بر پایه‌ی ترنسفرمر است که در زمینه‌ی دسته‌بندی و بخش‌بندی تصاویر، موفق ظاهر شده است [۳].

مدل SwinUNETR که ساختار رمزگذار ـ رمزگشا دارد و بخش رمزگذار آن بر پایه ترنسفرمر بوده و با استفاده از ۱۲ بلوک Swin ترنسفرمر و ورودی تصاویر مغزی، نقشه‌های ویژگی[۱۴] $\boldsymbol{x} \in \mathbb{R}^{H \times W \times D \times C}$ را استخراج می‌کند. در بخش رمزگشای این شبکه نیز نقشه‌های ویژگی استخراج شده توسط رمزگذار و

[12] Patch
[13] Convolutional Neural Network
[14] Feature Maps

نیز نقشه‌های ویژگی که توسط اتصالات ردشونده[۱۵] از بلوک‌های ترنسفرمر به رمزگشا پل زده شده‌اند، به یک شبکه پیچشی ورودی داده می‌شوند تا خروجی که همان تصویر بخش‌بندی شده توده مغزی است تولید شود.

در بلوک‌های مبتنی بر ترنسفرمر موجود در رمزگذار، از دو بلوک Swin ترنسفرمر پشت سر هم استفاده شده است که خروجی یکی ورودی دیگری است. ترنسفرمر اول به ترتیب از یک عادی‌سازی خطی[۱۶]، چندسر_توجه، عادی‌سازی خطی و Perceptron چندلایه تشکیل شده است. معماری ترنسفرمر دوم در این بلوک نیز تقریبا مشابه با ترنسفرمر اول است، با این تفاوت که بجای چندسر_توجه معمولی، از چندسر_توجه با سازوکار پنجره‌بندی[۱۷] استفاده می‌شود. معماری این مدل در شکل ۱_۲ مشخص شده است.



شکل ۱_۲: معماری شبکه SwinUNETR [۳].

بخش چندسر_توجه با سازوکار پنجره‌بندی استفاده شده در این معماری، از جابجایی یک پنجره سه‌بعدی بصورت چرخه‌ای استفاده می‌کند که این مورد به نوبه خود در عملکرد مدل اثر بسزایی دارد. تصویری از خروجی این مدل در شکل۱_۲ قابل مشاهده است. SwinUNETR یک مدل جذاب برای بخش‌بندی تصاویر سه‌بعدی است که نیاز به اجزای طراحی شده با دست را برطرف می‌کند.

[15] Skip Connections
[16] Linear Normalization
[17] Window Partitioning Multi-head Self-attention

۱ـ۴ اهداف تحقیق

در این پایان‌نامه سعی می‌شود مدلی مبتنی بر سازوکار توجه و بر پایه‌ی معماری ترنسفرمر برای بخش‌بندی توده‌های مغزی آموزش داده شود. سپس تلاش می‌شود نتایج این مدل با تغییراتی در معماری مدل و افزایش مجموعه دادگان بهبود داده شود.

۱ـ۵ ساختار پایان‌نامه

این پایان‌نامه شامل پنج فصل است. فصل دوم دربرگیرنده‌ی مفاهیم اولیه مرتبط با پایان‌نامه به همراه مرور ادبیات موضوع مرتبط و کارهای پیشین در زمینه مدل‌های بخش‌بندی تصاویر توده مغزی است. در فصل سوم به توضیح مدل، دادگان و توابع هزینه پرداخته می‌شود. در فصل چهارم نتایج جدیدی که در این پایان‌نامه به دست آمده‌اند ارائه می‌گردند. در این فصل، با بررسی نتایج کمی و کیفی خروجی‌های مدل، اثر تغییرات اعمال شده با خروجی‌های شبکه اولیه مورد مقایسه قرار می‌گیرند. در فصل پایانی نیز به نتیجه‌گیری و پیشنهادهایی برای کارهای آینده پرداخته می شود.

فصل ۲

ادبیات مربوطه

در این فصل به توضیح برخی مفاهیم اولیه از جمله سازوکار خودـ‌توجه[1]، پیچش تغییرشکل‌پذیر[2] و معیار ارزیابی Dice می‌پردازیم. علاوه بر این موارد بر ابزارها و روش‌های بخش‌بندی این توده‌ها نیز تمرکز خواهیم کرد.

۲ـ۱ سازوکار خودـ‌توجه

فرض می‌کنیم نقشه ویژگی ورودی $\boldsymbol{x} \in \mathbb{R}^{C_{in} \times H \times W}$ با ارتفاع H، عرض W و کانال‌های C_{in} است. خروجی $\boldsymbol{y} \in \mathbb{R}^{C_{out} \times H \times W}$ از یک لایه‌ی خودـ‌توجه با کمک ورودی تصویر شده به شکل زیر با رابطه ۲ـ۱ محاسبه می‌شود [۱]:

$$y_{ij} = \sum_{h=1}^{H} \sum_{w=1}^{W} \text{softmax}\left(q_{ij}^{T} k_{hw}\right) . v_{hw} \qquad (۲ـ۱)$$

که در آن پرسش‌ها $q = W_Q x$، کلیدها $k = W_K x$ و مقدارها $v = W_V x$ همگی تصویرهای محاسبه شده از روی ورودی x هستند. ماتریس‌های W_V, W_Q, W_K قابل یادگیری هستند. دلیل نام‌گذاری این سازوکار به خودـ‌توجه، به دست آمدن بردارهای پرسش، کلید و مقدار از روی یک بردار، همان بردار ورودی، است.

[1] Self-Attention
[2] Deformable Convolution

۲ـ۲ پیچش تغییرشکل‌پذیر

یکی از لایه‌های اصلی در شبکه‌های عصبی پیچشی، لایه پیچشی است. عملکرد این واحد به این صورت است که با در نظر گرفتن یک توری[3] که شامل تعدادی خانه است و هر خانه وزنی دارد، عمل پیچش را با جابجایی‌های متوالی این توری، بین این توری و تصویر ورودی اعمال می‌کند تا ویژگی‌های تصویر استخراج شود. رابطه ۲ـ۲ نحوه اعمال این پیچش را دقیق‌تر بیان می‌کند [۴]:

$$y_{p_0} = \sum_{p_n \in R} w(p_n).x(p_0 + p_n) \qquad (۲ـ۲)$$

که در آن w بیانگر وزن خانه مشخصی از توری، R خانه‌های توری، و x نیز تصویر ورودی است.

این واحد برای استخراج ویژگی از تصویر ورودی، فقط از نقطه تصویر[4]های مجاور هم استفاده می‌کند و ممکن است نتواند ویژگی‌هایی را که از هم فاصله مکانی دارند، تشخیص دهد. برای حل این مشکل، لایه‌های پیچش تغییرشکل‌پذیر معرفی شدند که عمل پیچش بین توری و نقطه‌تصویرهایی از تصویر که ممکن است با فاصله از هم قرار داشته باشند، انجام می‌شود [۴]. رابطه ۲ـ۳ نحوه عملکرد لایه پیچش تغییرشکل‌پذیر را بهتر بیان می‌کند [۴]:

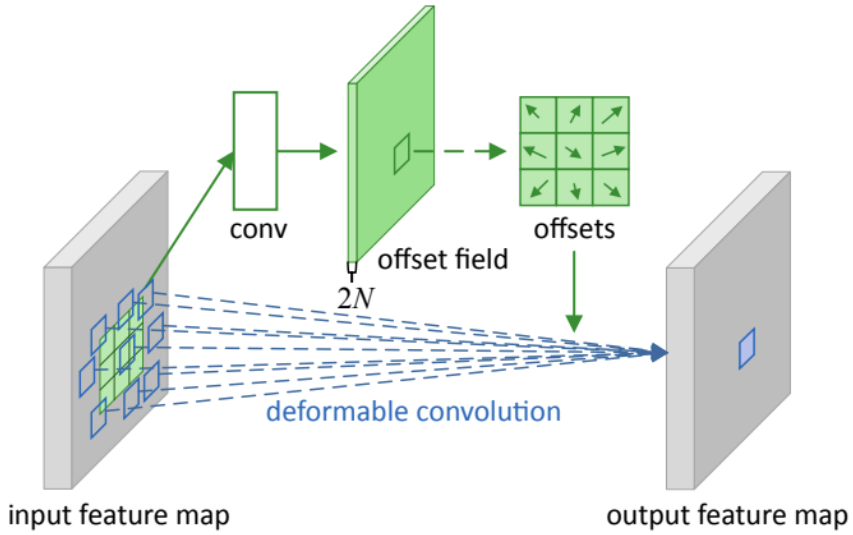
$$y_{p_0} = \sum_{p_n \in R} w(p_n).x(p_0 + p_n + \Delta p_n) \qquad (۲ـ۳)$$

در شکل ۲ـ۱ واحد پیچش تغییرشکل‌پذیر و در شکل ۲ـ۲ تصویری از مقایسه نحوه عملکرد این دو واحد نمایش داده شده است.

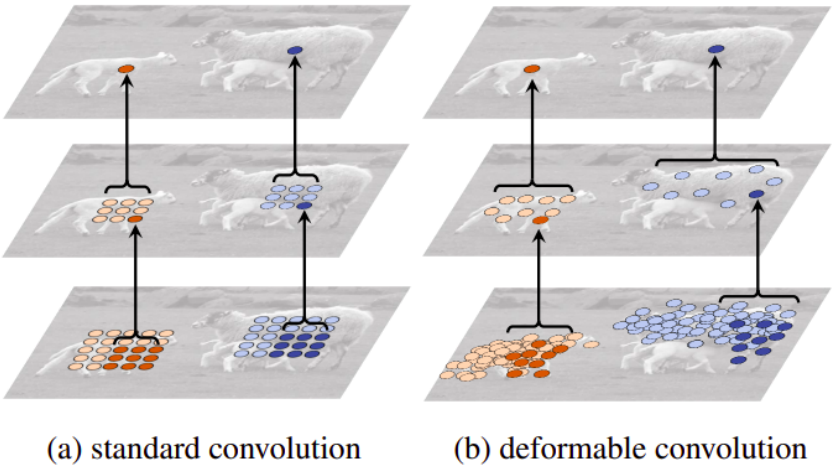
۲ـ۳ شبکه UNet

شبکه عصبی پیچشی UNet، یک ساختار رمزگذار ـ رمزگشا دارد که با اتصالات ردشونده به یکدیگر پل زده‌اند. در این معماری، رمزگزار نقش فشرده‌سازی و رمزگشا نقش گسترده‌سازی را بر عهده دارد و به این صورت عمل می‌کنند که پس از استخراج نقشه ویژگی توسط رمزگذار، این نقشه به رمزگشا داده می‌شود تا بتواند با استفاده از این ویژگی‌های استخراج شده و نیز ویژگی‌هایی که توسط اتصالات ردشونده به آن ورودی داده می‌شوند، این نقشه ویژگی را گسترده و خروجی موردنظر را تولید کند. این شبکه در اصل

[3] Grid
[4] Pixel



شکل ۲ـ۱: شمایی از یک پیچش تغییرشکل‌پذیر با اندازه توری ۳ در ۳ [۴].

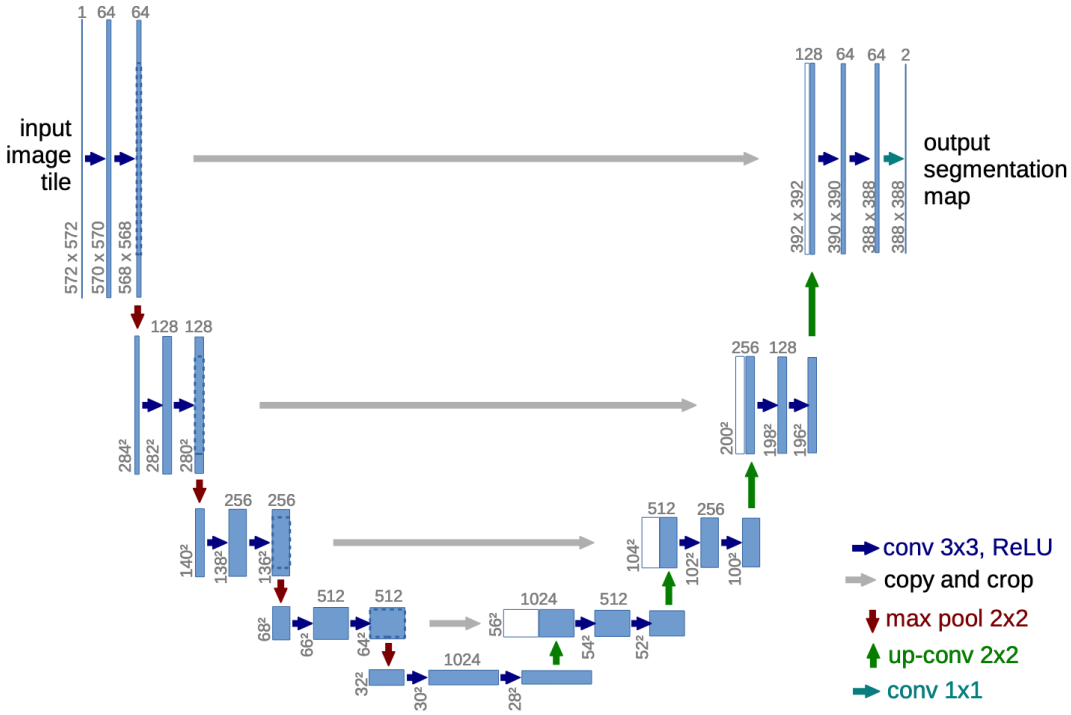


(a) standard convolution (b) deformable convolution

شکل ۲ـ۲: شمایی از مقایسه نحوه عملکرد لایه پیچش عادی در سمت چپ و لایه پیچش تغییرشکل‌پذیر در سمت راست [۴].

برای بخش‌بندی تصاویر بایوپزشکی تعریف شده است و از سالی که معرفی شده است تاکنون دگرگونی‌ها و بهبودهای فراوانی داشته است و در بخش‌بندی تصاویر سایر زمینه‌ها نیز پرکاربرد و موفق بوده است [۵].

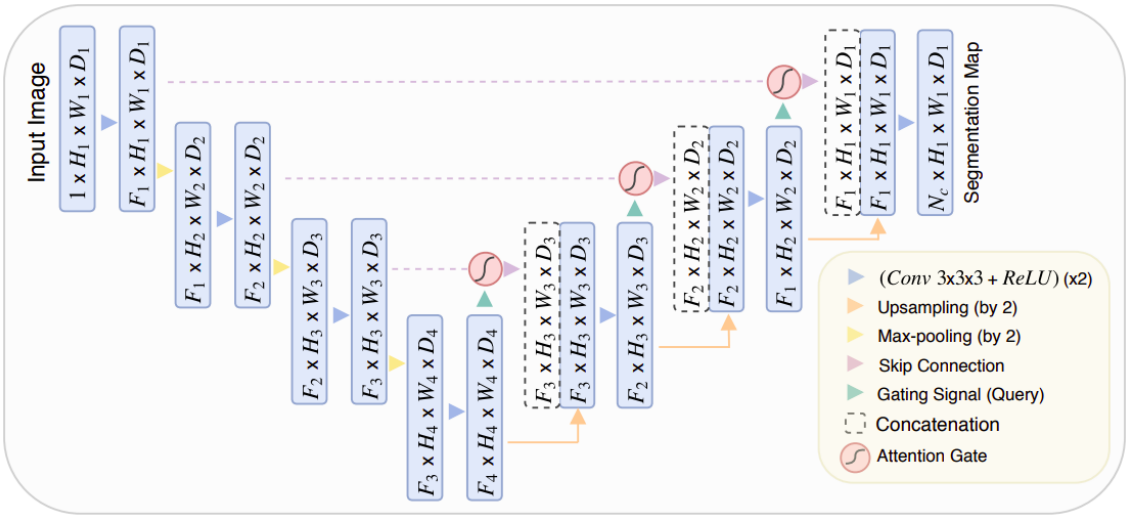
این شبکه تنها از لایه‌های پیچشی تشکیل شده است و به علت عدم استفاده از لایه‌های فشرده، توانایی پردازش تصاویر با ابعاد گوناگون را دارد. تصویری از معماری UNet در شکل ۲ـ۳ نشان داده شده است.



شکل ۲ـ۳: معماری UNet ـ مستطیل‌های آبی نشان‌دهنده‌ی نقشه‌های ویژگی چندکاناله است که تعداد کانال‌های آن در بالایش نوشته شده است. ابعاد تصاویر نیز پایین سمت چپ آنان قابل مشاهده است. عملکرد هر فلش در گوشه پایین در قسمت توضیحات نوشته شده است. مستطیل‌های سفید که در کنار مستطیل‌های آبی دیده می‌شوند حاصل از پل بین رمزگذار و رمزگشا هستند [۵].

۲ـ۴ شبکه AttentionUNet

AttentionUNet همانطور که از نامش پیداست، یک شبکه عصبی پیچشی UNet دارای واحد توجه است. این مدل نیز مانند مدل قبلی ساختار رمزگزارـرمزگشا دارد که توسط اتصالات ردشونده بهم پل زده‌اند. اما برتری این مدل نسبت به مدل UNet این است که در انتهای هر اتصال ردشونده یک دروازه توجه [۵] اضافه شده است تا از نقشه ویژگی بدست آمده از آن اتصال بتواند اطلاعات بیشتر و مفیدتری بدست آورد. چارچوب این مدل در شکل ۲ـ۴ نشان داده شده است [۶].

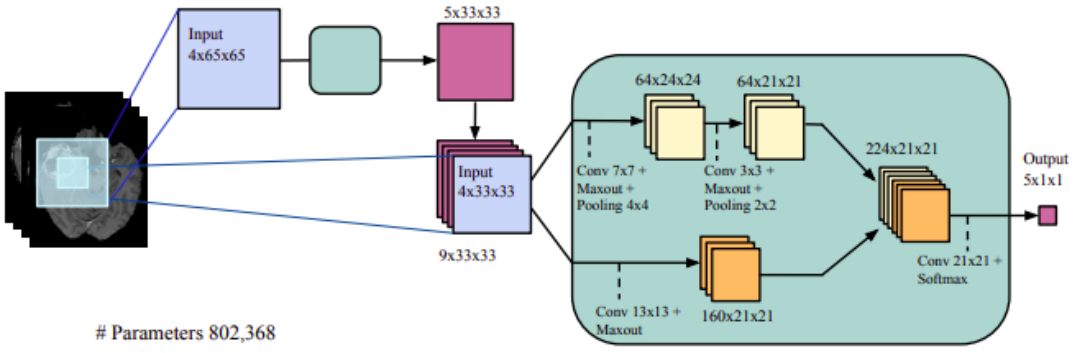


شکل ۲ـ۴: معماری AttentionUNetـ تصویر ورودی در رمزگزار به تدریج و با ضریب ۲ فشرده می‌شود. N_c تعداد دسته‌ها را نشان می‌دهد. دروازه‌های توجه ویژگی‌های ورودی از طریق اتصالات ردشونده را با استفاده از سازوکار توجه استخراج می‌کنند [۶].

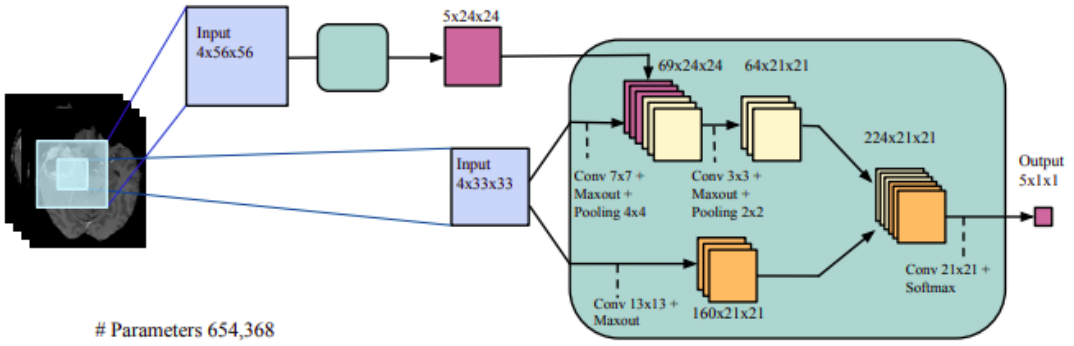
[5] Attention Gate

۲ـ۵ شبکه CascadedCNN

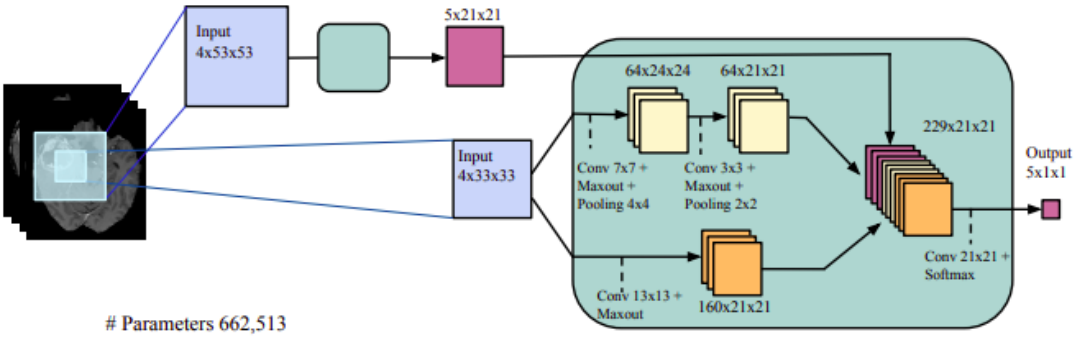
CascadedCNN برخلاف تصور عمده ما از شبکه‌های عصبی پیچشی که بصورت پشته‌ای از چندین لایه پیچشی هستند، معماری متفاوتی دارد. این شبکه با الحاق نقشه ویژگی‌های بدست آمده از لایه‌های مختلف معماری‌ای را می سازد که در آن مسیرهای محاسباتی متعدد، هدف های مختلفی را دنبال می‌کنند. دو بخش اصلی این شبکه عبارتند از: Two-pathway architecture و Cascaded architecture. شبکه Two-pathway architecture از ۲ مسیر تشکیل می‌شود، مسیر اول که مسیر محلی نام دارد برای استخراج ویژگی‌های نواحی نزدیک به نقاط تصویر است. مسیر دوم نیز که مسیر سراسری نام دارد، برای استخراج اطلاعات ناحیه بزرگتری در اطراف نقاط تصویر بکار می‌رود. شبکه Cascaded architecture خروجی‌های یک شبکه عصبی پیچشی را به عنوان ورودی به یک شبکه عصبی پیچشی دیگر می‌دهد و در هر کدام از این شبکه‌های عصبی از Two-pathway architecture استفاده می‌کند. بسته به بخشی که در آن خروجی شبکه اول وارد شبکه دوم می‌شود، ۳ معماری متفاوت پدید می‌آیند که در شکل۲ـ۵ نمایش داده شده‌اند [۷].



(a) Cascaded architecture, using input concatenation (InputCascadeCNN).



(b) Cascaded architecture, using local pathway concatenation (LocalCascadeCNN).

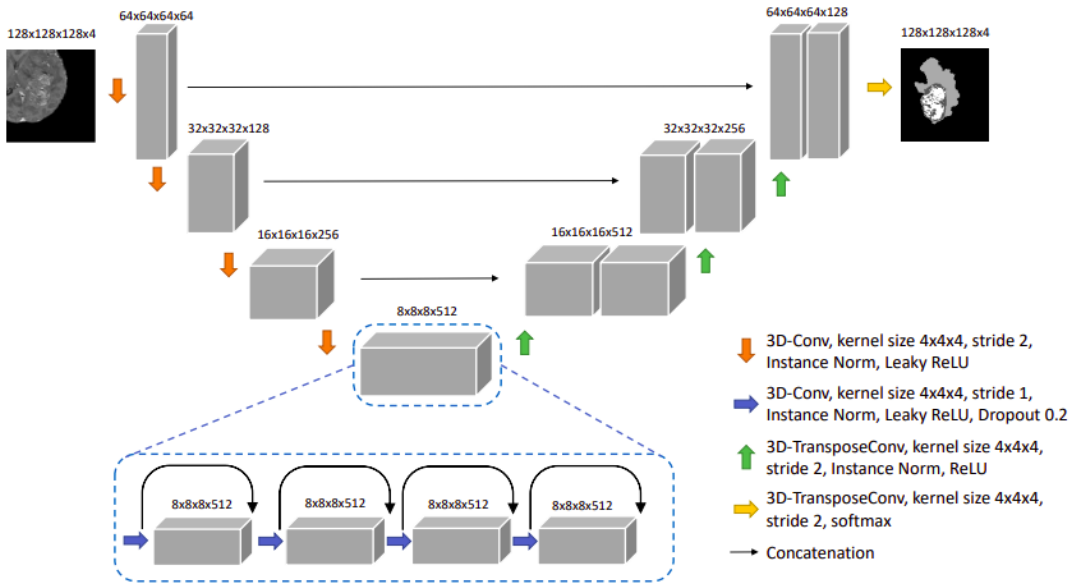


(c) Cascaded architecture, using pre-output concatenation, which is an architecture with properties similar to that of learning using a limited number of mean-field inference iterations in a CRF (MFCascadeCNN).

شکل ۲ـ۵: معماری‌های Cascade ـ در هر کدام از این ۳ معماری، خروجی بخش اول(ناحیه صورتی رنگ)، وارد واحد متفاوتی از بخش دوم می‌شود [۷].

۲ـ۶ شبکه Vox2Vox

امروزه شبکه‌های مولد تخاصمی به علت تواناییشان در تولید تصاویر مجوبیت زیادی در زمینه بینایی ماشین بدست آورده‌اند. شبکه Vox2Vox که از یک تولید کننده[۶] و یک تمییز دهنده[۷] تشکیل شده است، یک شبکه مولد تخاصمی است. بخش تولید کننده، یک ساختار رمزگذار_رمزگشا دارد که با اتصالات ردشونده به یکدیگر پل زده‌اند و بخش تمییز دهنده نیز یک شبکه پیچشی است. این مدل به این صورت عمل می‌کند که تولید کننده تلاش می‌کند تصاویری را تولید کند که تمییز دهنده را در تشخیص واقعی بودن تصویر به اشتباه بیاندازد. هر دو بخش تولید کننده و تمییز دهنده به ترتیب در شکل‌های ۲ـ۶ و ۲ـ۷ نمایش داده شده‌اند [۸].



شکل ۲ـ۶: معماری بخش تولید کننده شبکه Vox2Vox [۸].

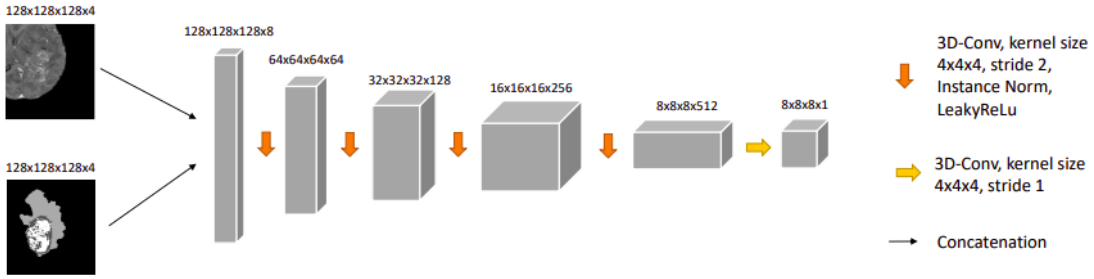
علاوه بر معماری نمایش داده شده در تصویر، تغییراتی نیز بر روی این شبکه اعمال شده است که از جمله آن‌ها افزودن واحد توجه در انتهای اتصالات ردشونده در تولید کننده و استفاده از پیچش تغییر شکل‌پذیر بجای لایه پیچشی عادی است.

[6] Generator
[7] Discriminator

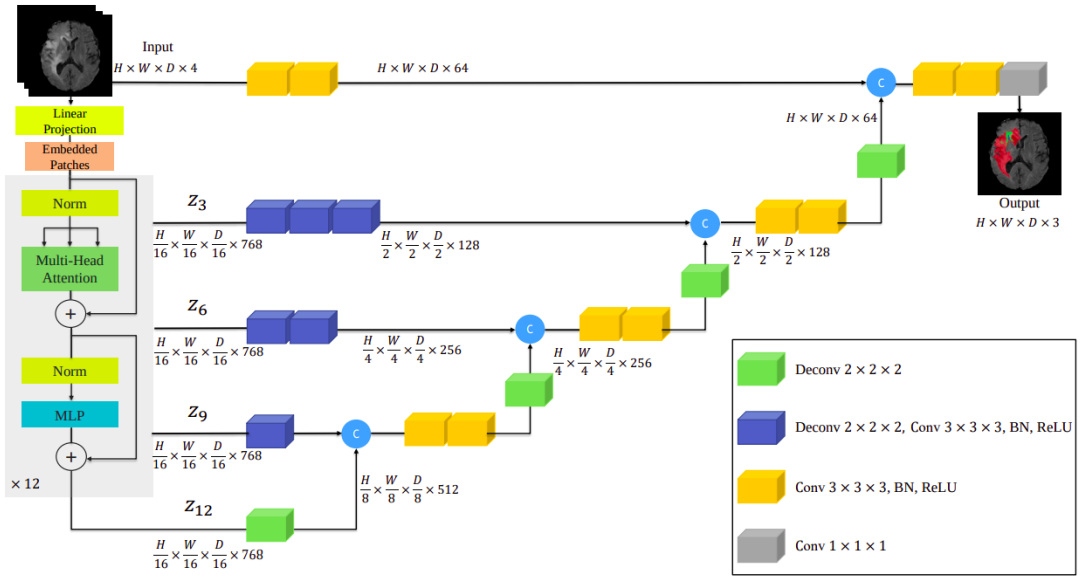
۲ـ۷ شبکه UNETR

شبکه UNETR با معماری نشان داده شده در شکل ۲ـ۸ مدلی بر پایه‌ی ترنسفرمر است، که در اواخر سال ۲۰۲۱ معرفی شده است. از این مدل برای بخش‌بندی تصاویر سه‌بعدی پزشکی استفاده شده است. در این مدل تصویر پس از تعبیه شدن در یک فضای یک‌بعدی و جمع شدن با جاسازی موقعیت[8] به مدل ورودی داده می‌شود. ساختار UNETR یک الگوی انقباضی‌ـ انبساطی متشکل از پشته‌ای از ترنسفرمرها در رمزگذار است. رمزگشای این مدل نیز برای تولید خروجی قطعه بندی شده، از لایه‌های پیچشی و خروجی رمزگذار از طریق اتصالات ردشونده استفاده می‌کند. شمایی از این مدل در شکل ۲ـ۸ نشان داده شده است [۹].

[8] Position Embedding



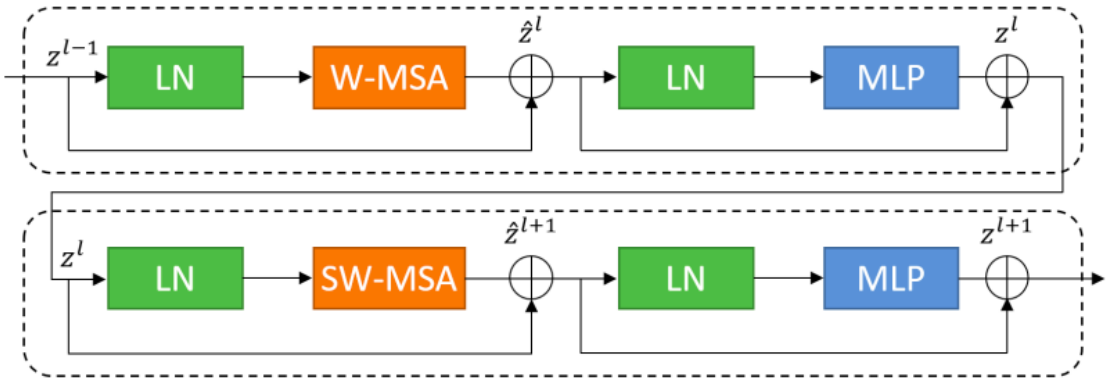
شکل ۲ـ۷: معماری بخش تمییز دهنده شبکه Vox2Vox [۸].



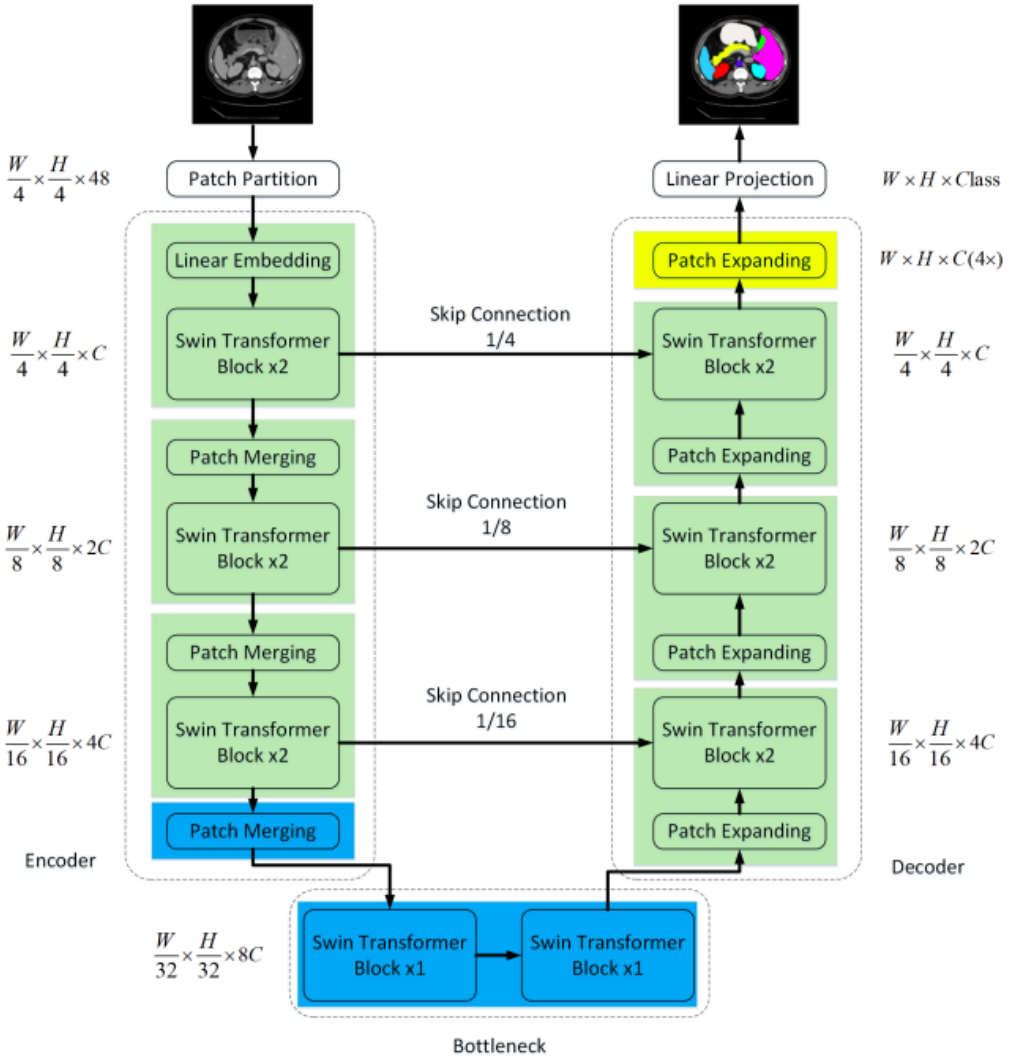
شکل ۲ـ۸: معماری شبکه UNETR ـ تصویر سه‌بعدی ورودی به دنباله‌ای از وصله‌های غیر هم‌پوشان تقسیم می‌شود و با استفاده از یک لایه خطی در یک فضای یک‌بعدی تعبیه می‌شود. این دنباله تعبیه شده، با جاسازی موقعیت جمع بسته شده و به عنوان ورودی به ترنسفرمر داده می‌شود. نمایش‌های رمزگذاری شده لایه‌های مختلف در ترنسفرمر استخراج می‌شوند و از طریق اتصالات ردشونده با رمزگشا ادغام می‌شوند تا بخش‌بندی نهایی را پیش‌بینی کند [۹].

۲ـ۸ شبکه Swin-unet

شبکه‌های پیچشی به علت استفاده از لایه‌های پیچشی و محلی بودن عملیات پیچش، توانایی یادگیری اطلاعات سراسری و دوربرد را ندارند. این در حالی است که یکی از ویژگی‌های اصلی ترنسفرمرها قابلیت آن‌ها در یادگیری اطلاعات سراسری است. شبکه Swin-unet یک ترنسفرمر خالص شبیه UNet است که برای بخش‌بندی تصاویر پزشکی بکار می‌رود. این مدل از رمزگذارـ رمزگشای ترنسفرمری با نوع ترنسفرمر Swin، تشکیل شده است. بین رمزگذار و رمزگشا نیز اتصالات ردشونده برقرار است. تصویر ساختار ترنسفرمر Swin و معماری کلی شبکه در شکل ۲ـ۹ و ۲ـ۱۰ نمایش داده شده است [۱۰].



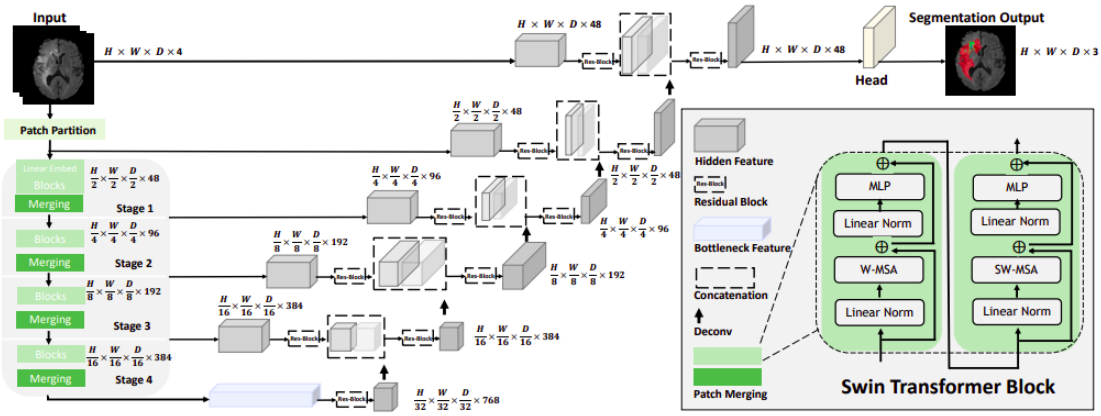
شکل ۲ـ۹: ترنسفرمر Swin [۱۰].



شکل ۲ـ۱۰: معماری Swin-unetـ از اتصالات رمزگذار، گلوگاه، رمزگشا و ردشونده تشکیل شده است. رمزگذار، گلوگاه و رمزگشا همگی از ترنسفرمرهای Swin تشکیل شده‌اند [۱۰].

۲ـ۹ شبکه SwinUNETR

با معرفی بلوک‌های ترنسفرمر Swin در سال ۲۰۲۱ انقلاب عظیمی در دنیای بینایی ماشین رخ داد. توانایی این بلوک‌ها در تشخیص ویژگی‌های سراسری و دوربرد باعث ایجاد شبکه‌هایی با چنین ستون فقراتی[۹] شد. شبکه SwinUNETR برای بخش‌بندی تصاویر سه‌بعدی پزشکی بکار می‌رود. تصویر ورودی پس از تعبیه شدن در یک فضای یک بعدی و جمع شدن با جاسازی موقعیت به عنوان ورودی به شبکه داده می‌شود. معماری شبکه نیز ساختار رمزگذار‌ـرمزگشا دارد و ورودی پس از عبور از لایه‌های ترنسفرمر رمزگذار، در هر لایه توسط اتصالات ردشونده به رمزگشا داده می‌شود و با استفاده از لایه‌های پیچشی،خروجی بخش‌بندی شده تولید می‌شود. تصویر این معماری در شکل ۲ـ۱۱ نشان داده شده است [۳].



شکل ۲ـ۱۱: معماری SwinUNETR ـ وصله‌های غیرهم‌پوشان تصویر سه‌بعدی به عنوان ورودی داده می‌شوند. ویژگی‌های رمزگذاری شده در لایه‌های مختلف ترنسفرمر Swin از طریق اتصالات ردشونده و با وضوح‌های مختلف به رمزگشای متشکل از لایه‌های پیچشی ورودی داده می‌شوند [۳].

[9] BackBone

۲ـ۱۰ جمع‌بندی

به طور کلی، تا به امروز مدل‌های فراوانی برای وظیفه‌ی بخش‌بندی تعریف شده است، که به توضیح و تفسیر ۷ عدد از آنان در بالا پرداخته شد. اکثریت مدل‌های بخش‌بندی که بر پایه شبکه‌های پیچشی هستند، از ساختار رمزگذار و رمزگشا به همراه اتصالات ردشونده تشکیل شده‌اند تا بتوانند نقطه تصویرهای ورودی را به خوبی پردازش کنند و اشیا با ابعاد متفاوت را دسته‌بندی کنند اما مدل‌هایی که بر پایه ترنسفرمر هستند علاوه بر این ساختار، از سازوکار توجه نیز برای یافتن اشیا استفاده می‌کنند. خود بلاک ترنسفرمر نیز می‌تواند بسته به معماری‌اش در روند بخش‌بندی تاثیر داشته باشد.

فصل ۳

روش پیشنهادی

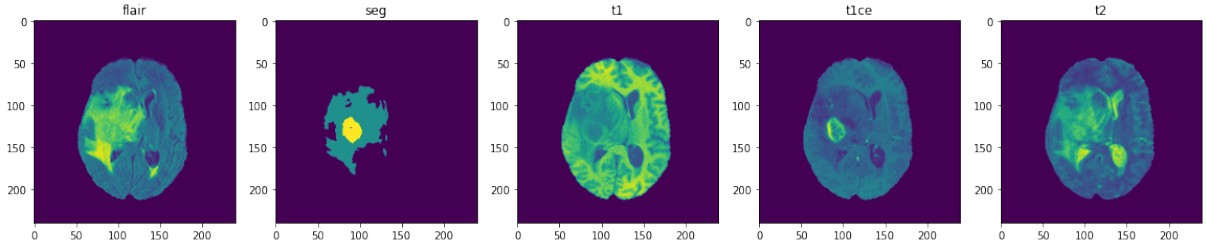
حال که مفاهیم اولیه و ادبیات موضوعی مربوطه شرح داده شدند، به توضیح روش پیشنهادی می‌پردازیم. در این فصل ابتدا دادگان مورد استفاده و پیش‌پردازش‌های انجام شده بر روی آن‌ها توضیح داده می‌شوند. سپس مدل پیشنهادی مبتنی بر ترنسفرمر معرفی می‌شود و در پایان توابع هزینه مختلفی که در روند آموزش این مدل بکار می‌روند توضیح داده خواهند شد.

۳ـ۱ دادگان

دادگان مورد استفاده در آزمایش‌ها، دادگان چالش برخطِ BraTS-2020 بوده است. این چالش هر ساله با هدف یافتن مدل‌های جدید برای بخش‌بندی معنایی[۱] توده‌های مغزی برگزار می‌شود، که اولین مرحله برای ساخت ابزار دسته‌بندی و بخش‌بندی توده‌های مغزی است. دادگان آموزش این مجموعه داده شامل تصاویر سه‌بعدی MRI مربوط به ۳۶۹ بیمار است که به ازای هر بیمار ۴ تصویر که هر کدام با روش‌های مختلفی تصویربرداری شده‌اند وجود دارد. این روش‌های مختلف تصویربرداری عبارتند از: T1[۲]، T1Gd[۳]، T2[۴] و FLAIR[۵]. علاوه بر این تصاویر، تصویر بخش‌بندی شده مغز به ازای هر بیمار نیز وجود دارد [۱۱].

[1] Semantic Segmentation
[2] Native
[3] Post-contrast T1-weighted
[4] T2-weighted
[5] T2 Fluid Attenuated Inversion Recovery

در زمان اجرا هیچ‌گاه از داده‌های آزمون در فرایند یادگیری مدل استفاده نشده است و از آن‌ها فقط برای گزارش نتایج استفاده شده است. از میان این مجموعه دادگان، ٪۷۵ آن‌ها که شامل ۲۷۷ نمونه توده با ابعاد ۱۵۵×۲۴۰×۲۴۰ است برای آموزش و ۹۲ نمونه توده باقی‌مانده برای آزمون استفاده می‌شود. نمونه‌ای از دادگان در شکل ۳ـ۱ نشان داده شده است.



شکل ۳ـ۱: نمونه‌ای از تصاویر دادگان این پروژه [۱۱].

۳ـ۲ پیش‌پردازش تصاویر

ماسک‌های موجود برای دادگان شامل ۴ دسته هستند که یکی از این دسته‌ها نشانگر پس زمینه و بخش‌هایی از مغز که تودهی در آن‌ها وجود ندارد، است. ۳ دسته دیگر نیز هر کدام بیانگر نوع خاصی از توده مغزی هستند که عبارتند از: ED[۶]، NCR/NET[۷] و ET[۸]. به عنوان پیش‌پردازش و با توجه به اینکه بخش‌بندی مهم برای ما تشخیص بخش هسته توده، قسمت‌های در حال رشد توده و کل توده است، برای اینکه بخش‌بندی بر اساس این ۳ بخش قابل اهمیت صورت گیرد، دسته‌های مربوط به هر بخش با یکدیگر تجمیع شدند و بدین ترتیب دسته WT از اجتماع دسته‌های ED، ET و NCR/NET، دسته TC از اجتماع دسته‌های ET و NCR/NET و نهایتا دسته ET نیز که معادل همان دسته ET است، حاصل شدند.

همانطور که بالاتر نیز اشاره شد ابعاد تصاویر در این دادگان ۱۵۵×۲۴۰×۲۴۰ است که با انجام عملیات برش و حذف بخش‌های کناری تصاویر به ابعاد ۶۴×۱۲۸×۱۲۸ تبدیل شدند. برای برخی از تصاویر و ماسک‌های نظیرشان به صورت تصادفی از فیلپ [۹] و سپس در مرحله بعد برای تصاویر

[6] Peritumoural Edema
[7] Necrotic and Non Enhancing Tumour Core
[8] GD Enhancing Tumour
[9] Flip

از عادی‌سازی شدت نور نقاط تصویر [10] استفاده شد. در انتها نیز شدت نور تصاویر به صورت تصادفی مقیاس شدند و انتقال‌هایی[11] بر روی آن‌ها صورت گرفت. تصاویر و ماسک‌های خروجی پیش‌پردازش‌های ذکر شده نیز در داده‌ساختار تنسور قرار گرفتند تا برای مدل قابل استفاده شوند.

۳ـ۳ مدل

مدل استفاده شده در این پروژه، مدل SwinUNETR است. این مدل مبتنی بر شبکه‌ای بر پایه‌ی ترنسفرمر است. معماری این شبکه که در شکل ۳ـ۲ قابل مشاهده است از دو بخش اصلی تشکیل شده است :

۱. رمزگذار: با شروع از عکس ورودی اولیه $\mathcal{X} \in \mathbb{R}^{H\times W\times D\times S}$، وصله‌هایی از آن با وضوح (H', W', D') و ابعاد $H' \times W' \times D' \times S$ به مدل ورودی داده می‌شوند. به منظور جدا کردن وصله‌ها از تصویر اصلی، از یک لایه جداکننده برای ساختن دنباله‌ای از توکن[12] های سه‌بعدی با ابعاد $\lceil \frac{H}{H'} \rceil \times \lceil \frac{W}{W'} \rceil \times \lceil \frac{D}{D'} \rceil$ استفاده می‌کنیم و آن‌ها را به یک فضای تعبیه‌سازی[13] با ابعاد C تصویر می‌کنیم. برای این که بتوان ارتباطات بین این توکن‌ها را بهتر مدل کرد، از سازوکار خودـ توجه بر روی پنجره‌هایی که هم‌پوشانی ندارند، استفاده می‌شود. به منظور تقسیم توکن‌های سه‌بعدی بطور مساوی به بخش‌هایی با ابعاد $\lceil \frac{H'}{M} \rceil \times \lceil \frac{W'}{M} \rceil \times \lceil \frac{D'}{M} \rceil$ در لایه l پنجره‌هایی با ابعاد $M \times M \times M$ بکار می‌روند. در ادامه و در لایه $l+1$ بخش‌های تقسیم‌بندی شده با استفاده از این پنجره‌ها به اندازه $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ شیفت داده می‌شوند. در لایه‌های بعدی پس از l و $l+1$ در رمزگذار، خروجی‌ها بصورتی که در رابطه ۳ـ۱ نوشته شده است، محاسبه می‌شوند [۳].

$$
\begin{aligned}
\hat{z}^{l} &= \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \\
z^{l} &= \text{MLP}(\text{LN}(\hat{z}^{l})) + \hat{z}^{l} \\
\hat{z}^{l+1} &= \text{SW-MSA}(\text{LN}(z^{l})) + z^{l} \\
z^{l+1} &= \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}
\end{aligned}
\qquad (۱ـ۳)
$$

در روابط بالا W-MSA و SW-MSA، به ترتیب بلاک‌های چندـسر خودـ توجه معمولی و با استفاده از پنجره‌های جداکننده؛ $\hat{z}^{l}$ و $\hat{z}^{l+1}$ خروجی‌های W-MSA و SW-MSA؛ LN نشان‌دهنده

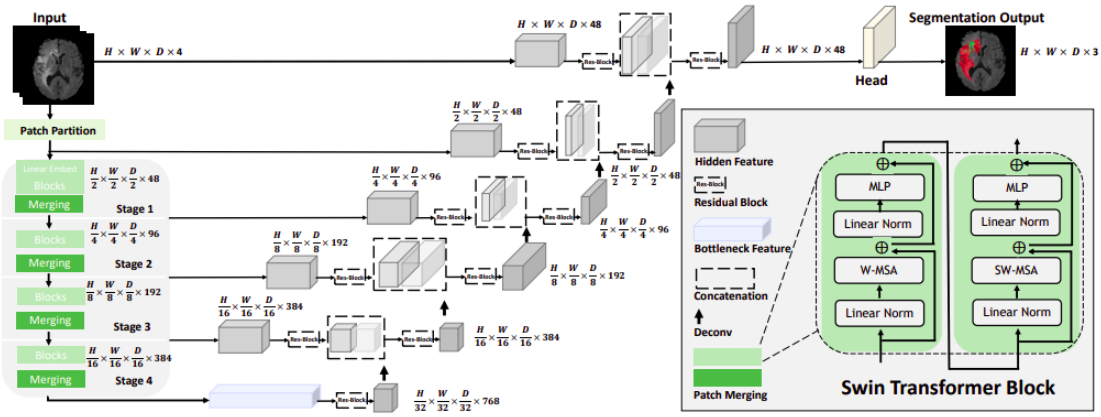
[10] Intensity Normalization
[11] Shift
[12] Token
[13] Embedding Space

لایه عادی‌سازی[۱۴] و MLP نشان دهنده پرسپترون چند لایه[۱۵] هستند.
اندازه وصله‌های استفاده شده در این رمزگذار $2\times2\times2$ و بعد ویژگی $2\times2\times2\times4 = 32$، که چهار نشان‌دهنده چهار کانال[۱۶] حاصل از روش‌های مختلف عکس برداری برای هر بیمار است. اندازه فضای تعبیه‌سازی یعنی C، ۴۸ است. به‌علاوه، این رمزگذار چهار مرحله دارد که هر مرحله دو بلاک ترنسفرمر را شامل می‌شود که باعث می‌شود در کل هشت لایه داشته باشد. در مرحله یک، یک لایه تعبیه‌سازی خطی[۱۷] برای ایجاد توکن‌های سه‌بعدی با ابعاد $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$ استفاده می‌شود. سپس یک لایه ادغام‌کننده وصله‌ها[۱۸] با ضریب دو برای کاهش وضوح ویژگی‌های استخراج شده و پس از آن، گروه‌بندی این وصله‌ها و الحاق کردن آن‌ها بهم در انتهای هر مرحله بکار می‌رود که باعث می‌شود بعد ویژگی‌های تعبیه‌شده به $4C$ برسد. ابعاد این بردار ویژگی در انتهای هر مرحله نصف مرحله قبل و به ترتیب در مراحل دو و سه و چهار این ابعاد $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$، $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$ و $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$ است.

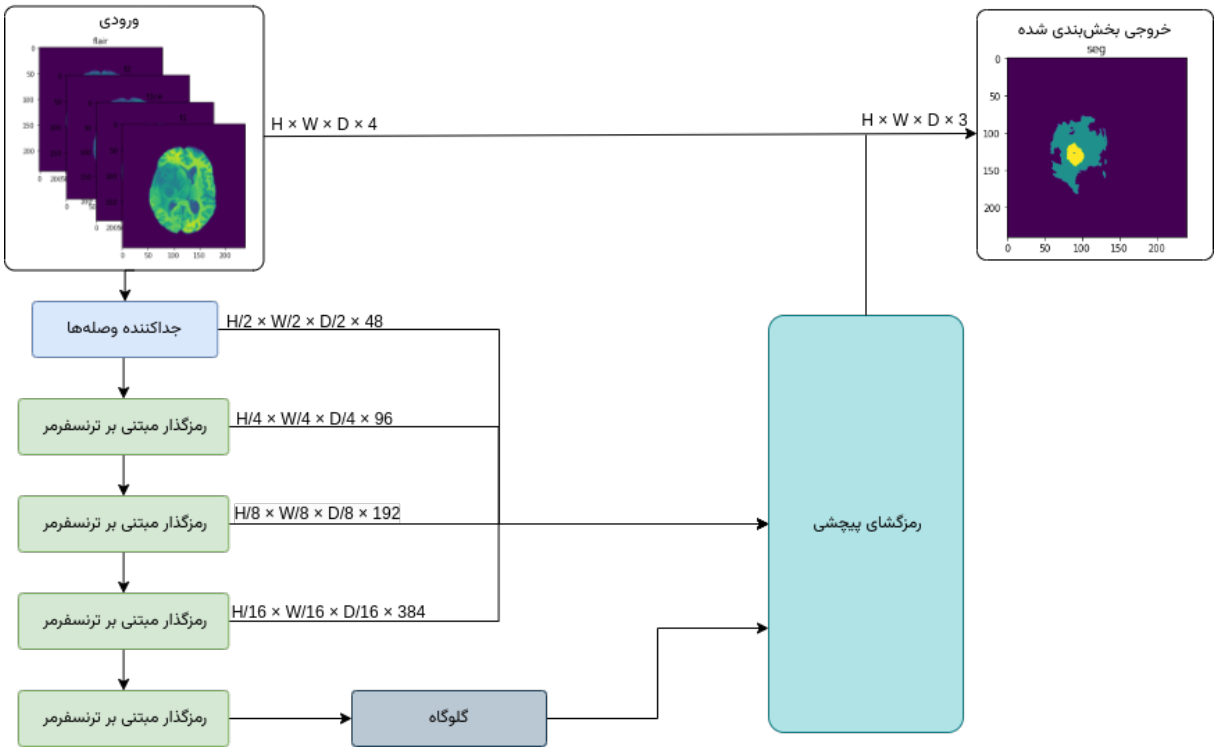
۲. رمزگشا: شبکه Swin UNETR یک طراحی U شکل دارد که در آن نمایش ویژگی های استخراج شده رمزگذار در رمزگشا از طریق اتصالات ردشونده در هر وضوح استفاده می شود. در هر مرحله i $(i \in \{0,1,2,3,4\})$ در رمزگذار و گلوگاه[۱۹] $(i = 5)$، بردار ویژگی استخراج شده به اندازه $\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i}$ تغییر شکل می‌دهد و به یک بلاک باقی‌مانده [۲۰] متشکل از دو لایه پیچشی $3 \times 3 \times 3$ و یک لایه عادی‌ساز ورودی داده می‌شود. پس از آن، وضوح نقشه های ویژگی با استفاده از یک لایه دکانولوشن[۲۱] ۲ برابر افزایش می یابد و خروجی ها با خروجی های مرحله قبل الحاق می شوند. سپس ویژگی های به هم الحاق شده، همانطور که قبلاً توضیح داده شد به یک بلوک باقی‌مانده دیگر وارد می شوند. خروجی‌های بخش‌بندی نهایی با استفاده از یک لایه پیچشی $1 \times 1 \times 1$ و تابع فعال‌ساز سیگموید[۲۲] محاسبه می‌شوند.

شمای کلی این مدل در تصویر ۳ـ۳ قابل مشاهده است.

[14] Layer Normalization
[15] Multi-Layer Perceptron
[16] Channel
[17] Linear Embedding Layer
[18] Patch Merging Layer
[19] Bottleneck
[20] Residual Block
[21] Deconvolution
[22] Sigmoid



شکل ۳ـ۲: معماری SwinUNETR ـ وصله‌های غیرهم‌پوشان تصویر سه‌بعدی به عنوان ورودی داده می‌شوند. ویژگی‌های رمزگذاری شده در لایه‌های مختلف ترنسفرمر Swin از طریق اتصالات ردشونده و با وضوح‌های مختلف به رمزگشای متشکل از لایه‌های پیچشی ورودی داده می‌شود [۳].



شکل ۳ـ۳: شمایی از معماری پیشنهادی.

۳ـ۴ توابع هزینه

برای آموزش شبکه SwinUNETR از توابع هزینه مختلفی برای بخش‌بندی تصاویر استفاده شده است که به اختصار به توضیح برخی از آنان می‌پردازیم.

۳ـ۴ـ۱ تابع هزینه Dice

این تابع هزینه که برای مسائل مربوط به بخش‌بندی بکار می‌رود، بسیار پرکاربرد است. برای محاسبه آن از رابطه ۳ـ۲ استفاده می‌شود.

$$\mathcal{L}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} \qquad (۳ـ۲)$$

در اینجا I نشان‌دهنده تعداد نقاط تصاویر، J نشان‌دهنده تعداد دسته‌ها و $Y_{i,j}$ و $G_{i,j}$ به ترتیب نشان‌دهنده احتمال خروجی مدل و هدف مدل برای دسته j و نقطه i تصویر را نشان می‌دهند.

حالت خاصی از این تابع هزینه نیز وجود دارد که به ترتیب بجای $Y_{i,j}^2$ و $G_{i,j}^2$ موجود در مخرج از $Y_{i,j}$ و $G_{i,j}$ استفاده می‌شود.

۳ـ۴ـ۲ تابع هزینه DiceCE

در این تابع برای محاسبه هزینه از ترکیب خطی تابع هزینه Dice و Cross Entropy استفاده می‌شود که در رابطه ۳ـ۳ نشان داده شده است.

$$\lambda_{\text{Dice}} L_{\text{Dice}} + \lambda_{\text{CE}} L_{\text{CE}} \qquad (۳ـ۳)$$

در رابطه فوق λ نشان دهنده یک ضریب ثابت بین ۰ و ۱ است که برای هر کدام از این توابع هزینه مقدار متفاوتی می‌گیرد. تابع هزینه Cross Entropy نیز از رابطه ۳ـ۴ محاسبه می‌شود.

$$L_{\text{CE}} = -\sum_{j=1}^{J} G_j \times \log Y_j \qquad (۳ـ۴)$$

۳ـ۴ـ۳ تابع هزینه Focal

این تابع اولین بار برای شبکه RetinaNet تعریف شد که برای دسته‌بندی اشیا در تصاویر به کار می‌رفت. هنگامی که دسته‌های مجموعه دادگان دارای عدم تعادل باشند، این تابع می‌تواند کمک‌کننده باشد. از طرفی این تابع هزینه نسبی دسته‌هایی که تقریبا درست طبقه‌بندی شده‌اند را نسبت به حالت عادی کمتر حساب می‌کند. رابطه ۳ـ۵ نحوه محاسبه آن را نشان می‌دهد [۱۲].

$$L_{\text{Focal}} = -\sum_{j=1}^{J} G_j \times (1 - Y_i)^{\gamma} \log Y_j \qquad (۳ـ۵)$$

۳ـ۴ـ۴ تابع هزینه Lovász-Softmax

تابع هزینه Lovász-Softmax گونه‌ای از تابع هزینه softmax است که به طور خاص برای بخش‌بندی تصاویر طراحی شده است و زمانی استفاده می‌شود که مدل در تلاش است تا از داده‌های آموزشی یاد بگیرد و یا عدم تعادل قابل توجهی در دسته‌های مجموعه دادگان وجود دارد. این تابع هزینه در تشخیص اشیا کوچک و کمتر کردن منفی غلط مفید است [۱۳].

۳ـ۵ جمع‌بندی

در این فصل ابتدا توضیحاتی در ارتباط با دادگان استفاده شده داده شد که شامل مواردی مثل نوع تصاویر، تعداد آن‌ها و نحوه تقسیم بندی این دادگان در مراحل آموزش و ارزیابی بود. سپس پیش پردازش‌های انجام شده بر روی آن‌ها که مواردی از قبیل فیلیپ، انتقال و عادی‌سازی شدت نور تصاویر بودند، شرح داده شدند. در آخر نیز پس از توضیح معماری مدل پیشنهادی مبتنی بر ترنسفرمر، توابع هزینه استفاده شده در مرحله آموزش این مدل توضیح داده شدند.

فصل ۴

نتایج تجربی

۴ـ۱ معیار ارزیابی

در این قسمت معیار اصلی ارزیابی بخش‌بندی تصاویر که در گزارش نتایج در این پایان‌نامه استفاده شده است، توضیح داده می‌شود.

تعریف ۴ـ۱ (معیار Dice) *معیار $Dice$ یک روش برای تعیین هم‌پوشانی بین ماسک[1] هدف و ماسک پیش‌بینی شده است. این معیار نسبت بین دو برابر اشتراک تعداد نقاط تصویر ماسک هدف و پیش‌بینی شده و کل نقاط ماسک‌های آن‌ها را اندازه‌گیری می‌کند. اشتراک بین دو ماسک، نقاط تصویری است که در ماسک هدف و ماسک پیش‌بینی وجود دارند. به عبارتی برای محاسبه $Dice$ بین دو ماسک هدف (T) و پیش‌بینی (P) داریم:*

$$Dice(T, P) = \frac{2 \times |T \cap P|}{|T + P|} \qquad (۴ـ۱)$$

که در رابطه‌ی (۴ـ۱) عملگر $||$ نشان‌دهنده‌ی تعداد نقاط تصویر در ناحیه است.

با توجه به تعریف معیار Dice ،برای هر ورودی مشخص، ماسک پیش‌بینی با ماسک هدف مقایسه می‌شود:

[1] Mask

- یک **مثبت صحیح (TP)** زمانی مشاهده می‌شود که جفت ماسک پیش‌بینی و هدف توده مغزی را نشان دهند.
- یک **مثبت غلط (FP)** زمانی مشاهده می‌شود که یک ماسک پیش‌بینی توده مغزی را نشان دهد درحالیکه ماسک هدف توده نباشد.
- یک **منفی غلط (FN)** زمانی مشاهده می‌شود که ماسک پیش‌بینی بخشی را بدون توده دسته‌بندی کند، در حالیکه در ماسک هدف آن بخش توده باشد.

بنابراین با استفاده از موارد ذکر شده معیار Dice، به صورت زیر محاسبه می‌شود:

$$Dice = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \qquad (۴\_۲)$$

در صورتی که یک مدل بتواند ماسک هدف را به طور دقیق پیش‌بینی کند، امتیاز Dice دقیقا برابر با یک می‌شود.

۴_۲ تنظیمات

کلیه آزمایش‌ها با استفاده از کتابخانه‌ی متن‌باز پایتورچ[۲] روی کارگزار Google Colaboratory با حافظه اصلی ۷۷ گیگابایت و با بهره‌گیری از پردازنده گرافیکی Tesla T4 با حافظه ۱۲ گیگابایت انجام شده‌اند. هر یک از مدل‌ها با تنظیمات مختلف طی ۱۰ دور یا ایپاک[۳] آموزش داده شده‌اند که تعداد داده‌های هر دسته، ۱ و بهینه‌ساز استفاده شده در تمامی آزمایش‌ها، AdamW است. عملکرد مدل در هر کدام از این آزمایش‌ها با استفاده از امتیاز Dice بر روی مجموعه دادگان آزمون گزارش می‌شود.

۴_۳ آزمایش‌ها

در این فصل، نتایج ارزیابی مدل SwinUNETR در طی آزمایش‌های مختلف آورده شده است. در اولین دسته از آزمایش‌ها اثر تغییر تعداد کارگرهای موجود در Data Loader را بررسی می‌کنیم. این کارگران

[2] PyTorch
[3] Epoch

موازی سازی عمل بارگزاری دادگان از حافظه را انجام می‌دهند. در جدول ۴ـ۱ با بررسی ۳ مقدار ۱، ۲ و ۸ مشاهده می‌کنیم که مقدار ۲ نتیجه بهتری نسبت به سایر مقادیر دارد.

جدول ۴ـ۱: بررسی نتایج مدل SwinUNETR برای مقادیر مختلف تعداد کارگران Data Loader.

تعداد کارگران	امتیاز دسته WT	امتیاز دسته TC	امتیاز دسته ET	میانگین امتیازات
۱	۰/۴۵	۰/۳۹	۰/۱۴	۰/۳۳
۲	۰/۴۶	۰/۴۱	۰/۱۴	۰/۳۴
۸	۰/۴۱	۰/۳۶	۰/۱۲	۰/۳۰

در جدول ۴ـ۲ به بررسی نتایج مدل SwinUNETR با توابع هزینه مختلف می‌پردازیم. همانطور که از این نتایج برمی‌آید، تابع هزینه DiceSquared نتایج بهتری نسبت به بقیه حالات دارد. به طور کلی در توضیح عملکرد هر کدام از این توابع می‌توان گفت همانطور که در بخش روش پیشنهادی نیز توضیح داده شد، توابع DiceCELoss، DiceLoss و DiceLossSquared به علت تفاوتی اندکی که در پیاده‌سازی‌شان وجود دارد و اینکه بخش اصلی هر ۳، تابع Dice است، عملکرد نسبتا مشابهی نیز دارند. تابع Lovász-Softmax به علت استفاده از شاخص جاکارد[۴] برای معیار ارزیابی در پیاده‌سازی خود، عملکرد مطلوبی در آزمایشاتی که از تابع امتیاز Dice استفاده می‌کنند، ندارد. تابع Focal Loss نیز بخاطر توزیع احتمالاتی خروجی لایه آخر مدل، عملکرد خوبی ندارد و بیشتر ماسک‌ها را جزو پس‌زمینه قطعه بندی می‌کند.

جدول ۴ـ۲: بررسی نتایج مدل SwinUNETR برای توابع هزینه متفاوت.

تابع هزینه	امتیاز دسته WT	امتیاز دسته TC	امتیاز دسته ET	میانگین امتیازات
DiceCE	۰/۴۶	۰/۴۱	۰/۱۴	۰/۳۳۸
Dice	۰/۴۸	۰/۴۱	۰/۱۴	۰/۳۴۳
DiceSquared	۰/۴۹	۰/۴۳	۰/۱۸	۰/۳۶۶
Lovász-Softmax	۰/۱۲	۰/۰۹	۰/۰۲	۰/۰۸
Focal ($\gamma = 0$)	۰/۳۵	۰/۱۶	۰	۰/۱۷
Focal ($\gamma = 1$)	۰/۳۶	۰/۰۲	۰	۰/۱۴
Focal ($\gamma = 2$)	۰	۰	۰	۰

دسته دیگر آزمایشات بر روی آستانه[۵] تبدیل مقادیر احتمالی به مقادیر دودویی (جهت تشخیص نوع

[4] Jaccard's Loss
[5] Threshold

دسته) انجام شده‌است. در این آزمایشات علاوه بر مقدار اولیه آن در آزمایشات فوق که ۰/۶ بود، دو آستانه ۰/۵ و ۰/۷ نیز بررسی شده‌اند که نتایج هر کدام در جدول ۴ـ۳ برای مقایسه مدل قابل مشاهده است. با توجه به توزیع خروجی‌های احتمالی مدل، مقدار ۰/۵ نتیجه بهتری دارد.

جدول ۴ـ۳: بررسی نتایج مدل SwinUNETR برای مقادیر مختلف آستانه.

مقدار آستانه	امتیاز دسته WT	امتیاز دسته TC	امتیاز دسته ET	میانگین امتیازات
۰/۵	۰/۴۹	۰/۴۳	۰/۱۹	۰/۳۷
۰/۶	۰/۴۹	۰/۴۳	۰/۱۸	۰/۳۶۶
۰/۷	۰/۴۸	۰/۴۱	۰/۱۶	۰/۳۵

در دسته بعدی آزمایشات اثر تغییر تعداد دورهایی که بعد از آن مقدار نرخ یادگیری[6] شروع به کاهش می‌کند که به آن Warmup Epochs نیز گفته می‌شود، بررسی می‌شود. در تمامی آزمایشات تاکنون این مقدار مساوی ۱ بود که علاوه بر آن مقدار ۵ را نیز بررسی می‌کنیم که تاثیر مثبت آن در جدول ۴ـ۴ قابل مشاهده است.

جدول ۴ـ۴: بررسی نتایج مدل SwinUNETR برای Warmup Epochs متفاوت.

شماره دور	امتیاز دسته WT	امتیاز دسته TC	امتیاز دسته ET	میانگین امتیازات
۱	۰/۴۹	۰/۴۳	۰/۱۹	۰/۳۷
۵	۰/۵۲	۰/۴۶	۰/۲۱	۰/۴۰

دسته آخر آزمایشات اثر افزایش تعداد دادگان را بر روی عملکرد مدل بررسی می‌کند. در این آزمایش علاوه بر مجموعه دادگان کنونی، از دادگان چالش برخطBraTS2021 نیز استفاده می‌کنیم که شامل دادگان ۳۰۰ بیمار مبتلا به توده مغزی است. با ترکیب این دو مجموعه داده، تعداد دادگان ۶۶۹ مورد افزایش می‌یابد. همانطور که از نتایج جدول ۴ـ۵ نیز بدست می‌آید، افزایش تعداد دادگان بر روی عملکرد مدل اثر مثبتی داشته است.

جدول ۴ـ۵: بررسی نتایج مدل SwinUNETR برای تعداد دادگان متفاوت.

مجموعه دادگان	امتیاز دسته WT	امتیاز دسته TC	امتیاز دسته ET	میانگین امتیازات
BraTS2020	۰/۵۲	۰/۴۶	۰/۲۱	۰/۴۰
BraTS2020+2021	۰/۶۰	۰/۵۲	۰/۲۸	۰/۴۷

به طور کلی در این ۴ دسته آزمایش به ترتیب اثر تعداد کارگران Data Loader، توابع هزینه، مقادیر

[6] Learning Rate

مختلف برای آستانه، مقدار Warmup Epochs و افزایش تعداد دادگان بررسی شدند تا عملکرد مدل بدون ایجاد تغییرات در معماری بهبود داده شود. در این آزمایشات که بهترین خروجی هر دسته در دسته بعدی آزمایشات استفاده می‌شد، به ترتیب به میزان ۱، ۳، ۰/۰۰۴، ۳ و ۷ درصد بهبود حاصل شد.

فصل ۵

جمع‌بندی و راه‌کارهای آتی

۵ـ۱ جمع‌بندی

در طی مراحل این پایان‌نامه، شبکه‌ای بر پایه‌ی معماری ترنسفرمر که امروزه نتایج خیره‌کننده‌ای در زمینه‌های بخش‌بندی و دسته‌بندی اشیا در تصاویر، از زمینه‌های مختلف داشته‌اند، برای بخش‌بندی توده‌های مغزی طراحی و پیاده‌سازی شد. هدف از این بخش‌بندی برداشتن اولین گام برای مقابله با توده‌های مغزی بوده است.

روش‌های مبتنی بر ترنسفرمر به منابع سخت‌افزاری زیاد و همچنین مجموعه دادگان قوی برای یادگیری نیازمندند. به طور خاص مدل SwinUNETR برای آموزش به تعداد هسته‌های GPU زیاد و زمان زیاد برای یادگیری و همگرا شدن نیازمند است، همچنین این مدل حجم دادگانی که نیاز دارد در حدود چند ۱۰ هزارتا می‌باشد. همانطور که در بخش دادگان مطرح شد، حجم کل این مجموعه دادگان به ۴۰۰ تصویر نیز نمی‌رسد و این حجم برای این مدل برای همگرایی کامل کم بوده است. همچنین با توجه به حجیم بودن مدل SwinUNETR نمی‌توان دسته‌هایی که به مدل ورودی می‌دهیم را متناسب با کمبود منابع سخت‌افزاری زیاد گرفت، در نتیجه آموزش این مدل نیز بسیار زمانبر می‌باشد.

با توجه به موارد مطرح شده در بالا و مقادیر گزارش شده در بخش نتایج می‌توان مشکلات مطرح شده را تا حدودی با تنظیم پارامترهای مدل مرتفع ساخت و عملکرد مدل را بهبود بخشید.

۵ـ۲ راه‌کارهای آتی

در آینده برای بهبود شبکه‌های مبتنی بر ترنسفرمر برای بینایی ماشین می‌توان از ایده‌های زیر استفاده کرد:

- از لایه پیچش تغییر شکل پذیر به‌جای به لایه پیچشی عادی می‌توان استفاده کرد. این لایه‌ها به علت حجیم بودن نیاز به منابع سخت‌افزاری زیادی دارند اما در بسیاری از موارد باعث بهبود عملکرد مدل شده‌اند [۴].
- در واحدهای چندسرـ توجه از سازوکار توجهِ محوریِ دردار[1] به جای سازوکار توجه عادی استفاده شود. این سازوکار در مواردی که دادگان کمی در دسترس است، موفق عمل کرده است و در وظایفی مانند بخش‌بندی معنایی تصاویر فراصوت مغزی و همچنین تصاویر میکروسکوپی نتایج قابل توجهی داشته است [۱۴].
- برای حل دشواری‌های موجود در بخش‌بندی این تصاویر می‌توان از روش‌های یادگیری ماشین کوانتومی[2] که باعث انقلاب در دنیای بینایی ماشین شده‌اند، استفاده کرد [۱۵].

[1] Gated Axial-Attention
[2] Quatum Machine Learning (QML)

مراجع

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[3] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022.

[4] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks, 2017.

[5] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[6] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[7] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, Jan. 2017.

[8] M. D. Cirillo, D. Abramian, and A. Eklund. Vox2vox: 3d-gan for brain tumour segmentation, 2020.

[9] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation, 2021.

[10] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021.

[11] Multimodal brain tumor segmentation challenge 2020. https://ipp.cbica.upenn.edu/. Accessed: 2023-01-04.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] M. Berman, A. R. Triki, and M. B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, 2018.

[14] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel. Medical transformer: Gated axial-attention for medical image segmentation, 2021.

[15] S. Pramanik, M. G. Chandra, C. V. Sridhar, A. Kulkarni, P. Sahoo, V. C. D, H. Sharma, A. Paliwal, V. Navelkar, S. Poojary, P. Shah, and M. Nambiar. A quantum-classical hybrid method for image classification and segmentation, 2021.

واژه‌نامه

الف

آستانه ..... Threshold

اتصالات ردشونده ..... Skip Connections

انتقال ..... Shift

ب

بلاک باقی‌مانده ..... Residual Block

بینایی ماشین ..... Computer Vision

پ

پرسپترون چندلایه ..... Multi-Layer Perceptron

پرسش ..... Query

پس‌پردازش ..... PostProcessing

پیچش تغییرشکل‌پذیر . Deformable Convolution

پیش‌پردازش ..... PreProcessing

ت

ترجمه ماشینی ..... Machine Translation

ترنسفرمر ..... Transformer

تمییز دهنده ..... Discriminator

توجهِ محوریِ دردار ..... Gated Axial-Attention

توری ..... Grid

تولید کننده ..... Generator

ج

جاسازی موقعیت ..... Position Embedding

چ

چندسر_توجه ..... MultiHead-Attention

چندسر_توجه با مکانیزم پنجره‌بندی ..... Window Partitioning Multi-head Self-attention

خ

خود_توجه ..... Self-Attention

د

دروازه توجه ..... Attention Gate

س

ستون فقرات ..... BackBone

ش

شبکه عصبی پیچشی ..... Convolutional Neural Network

ع

عادی‌سازی خطی ..... Linear Normalization

عادی‌سازی شدت نور .. Intensity Normalization

عادی‌سازی لایه ..... Layer Normalization

ف

فضای تعبیه‌سازی ..... Embedding Space

ق

بخش‌بندی ..... Segmentation

بخش‌بندی معنایی ..... Semantic Segmentation

ک

کدگذار ..... Encoder

کدگشا ..... Decoder

کلید ..... Key

گ

گلوگاه ..... Bottleneck

ل

لایه ادغام کننده وصله‌ها ... Patch Merging Layer

لایه تعبیه‌سازی خطی .. Linear Embedding Layer

م

ماسک ..... Mask

مکانیزم توجه ..... Attention Mechanism

ن

نرخ یادگیری ..... Learning Rate

نقطه تصویر ..... Pixel

نقشه ویژگی ..... Feature Map

و

وصله ..... Patch

ی

یادگیری ژرف ..... Deep Learning

یادگیری ماشین ..... Machine Learning

یادگیری ماشین کوانتومی ..... Quatum Machine Learning

Abstract

Brain tumors pose a significant threat to the lives of many individuals. Accurate segmentation and treatment of these tumors can greatly improve patient outcomes, however, achieving this can be challenging for medical teams. The use of deep learning-based methods, such as convolutional neural networks and generative adversarial networks, have shown promise in increasing segmentation accuracy for brain tumors. These methods, however, require extensive computational resources and a large amount of training data. Recently, architectures such as the "Transformer" have demonstrated exceptional performance in various applications, including image classification and segmentation. In this project, we will employ a transformer-based network to segment 3D images of brain tumors using the BraTS dataset and evaluate its performance using the Dice score. By adjusting parameters and increasing the dataset, we will take steps to improve the model and evaluate the results through comparison. In total, these changes improve the model's dice score by 14% over 10 epochs.

Keywords: Brain Tumor, Semantic Segmentation, Deep Neural Network, Transformer, Generative Adversarial Network, BraTS Dataset

Sharif University of Technology

Department of Computer Engineering

B.Sc. Thesis

Brain Tumor Segmentation Using Deep Neural Networks

By:

Mahsa Amani

Supervisor:

Dr. Shohreh Kasaei

January 2023