دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

گزارش درس شبکه‌های عصبی

شبکه یو-نت و نسخه‌های بهبود یافته آن

نگارش

فردین آیار

استاد درس

دکتر رضا صفابخش

تابستان ۱۴۰۰

چکیده

مسئله قطعه‌بندی معنایی، علارغم سابقه طولانی، به علت کاربردهای فراوان آن همچنان یکی از مهمترین مسائل مطرح در علم بینایی ماشین است. تحت تاثیر پیشرفت‌های اخیر در یادگیری عمیق و شبکه‌های کانولوشنی، اکثر محققین برای حل این مسئله به سمت استفاده از یادگیری عمیق و شبکه‌ها کانولوشنی متمایل شدند. شبکه‌های تمام کانولوشنی، اصلی‌ترین جریان موجود برای پرداختن به مسئله قطعه‌بندی معنایی است. اگرچه این شبکه‌ها عملکرد فوق‌العاده‌ای از خود نشان دادند، اما دو ایراد اساسی در نسخه‌های اولیه آن‌ها وجود داشت. اول، کاهش پیاپی رزولوشن کانال‌های ویژگی به منظور استخراج اطلاعات معنایی، موجب از دست رفتن اطلاعات فضایی می‌شد و دوم، برای آموزش کامل این شبکه‌ها، مانند سایر شبکه‌های عصبی، نیاز به داده‌های آموزشی فراوان بود؛ درحالی که جمع‌آوری داده‌های آموزشی بویژه برای مسئله قطعه‌بندی معنایی دشوار است. یکی از اولین شبکه‌هایی که توانست هر دو ایراد فوق را تا حد زیادی برطرف کند شبکه یو–نت بود. در این گزارش خواهیم دید که شبکه یو–نت و نسخه‌های بهبود یافته آن، چگونه با ساختار خاص خود به این دو ایراد اساسی رسیدگی کردند؛ چیزی که باعث شد سال‌ها پس از ارائه نسخه اولیه شبکه یو–نت، همچنان مورد توجه باشد و محققین مختلف، همچنان در پی ارائه نسخه‌های پیشرفته‌تر از آن باشند.

کلمات کلیدی: قطعه‌بندی معنایی، قطعه‌بندی تصاویر، شبکه‌های تمام کانولوشنی، شبکه یو–نت، ساختار کدگشا–کدگذار

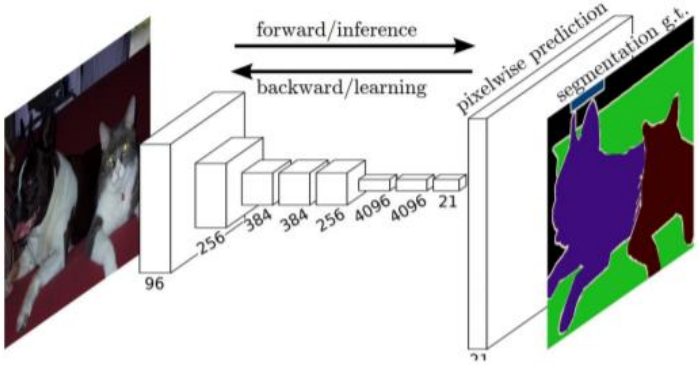
فهرست مطالب

فهرست اشکال

۱- مقدمه

اگرچه قطعه‌بندی تصاویر[۱]، به معنی تقسیم تصویر به قطعات معنادار، یک موضوع با سابقه در علم بینایی ماشین می‌باشد.اما کاربردهای آن در زمینه‌های مختلف از جمله پزشکی، اتوموبیل‌های خودران، واقعیت افزوده و ...، باعث شده که همچنان یک موضوع پویا و بروز باشد. قطعه‌بندی تصاویر می‌تواند به صورت تخصیص یک برچسب معنایی به هر پیکسل(قطعه‌بندی معنایی[۲])، قطعه‌بندی اشیا موجود در تصویر(قطعه‌بندی اشیا[۳]) و یا ترکیبی از هردو(قطعه‌بندی همه‌نما[۴]) تعریف کرد.

کارهای اولیه در این زمینه، از روش‌های مختلفی برای انجام قطعه‌بندی تصاویر استفاده می‌کردند، اما پس از گسترش استفاده از شبکه‌های عصبی و با توجه به عملکرد فوق‌العاده آن‌ها، اکثر محققین به استفاده از آن‌ها در مسئله قطعه‌بندی روی آوردند. به طور خاص برای مسئله قطعه‌بندی معنایی ،که بیشتر مورد نظر این گزارش است، شبکه تمام کانولوشنی[۱] جز اولین روش‌هایی بود که از یادگیری عمیق برای حل مسئله قطعه‌بندی معنایی استفاده کرد(شکل ۱). در این شبکه لایه‌های تماماً متصل[۵] از معماری‌های کانولوشنی رایج، با لایه‌های کانولوشنی جایگزین شده‌اند. استفاده از اتصالات پرشی[۶] به این شبکه اجازه می‌داد که اطلاعات معنایی لایه آخر را با اطلاعات فضایی لایه‌های ابتدایی ترکیب کند و از این طریق قطعه‌بندی نسبتاً دقیق و باجزئیاتی ارائه دهد. همچنین این معماری موجب شد که شبکه ارائه شده، محدودیتی برای سایز تصویر ورودی نداشته باشد. این شبکه نقطه عطفی در مدل‌های قطعه‌بندی بود که کارهای پس از خود را به سمت یادگیری عمیق جهت داد[۲].



شکل ۱-ساختار شبکه تمام کانولوشنی[۱]

علارغم موفقیت شبکه‌های تمام کانولوشنی[۱]، دو ایراد اساسی در آن‌ها وجود داشت. ایراد اول این بود که در شبکه‌های کانولوشنی، برای استخراج اطلاعات معنایی، رزولوشن کانال‌های ویژگی به طور مداوم کاهش می‌یابد؛ در نتیجه بسیاری از جزئیات تصویر از بین می‌رود؛ اگرچه اتصالات پرشی موجود در شبکه‌های تمام کانولوشنی[۱] تا حدی این ایراد را برطرف می‌کند. دومین ایراد این بود که این شبکه‌ها، مانند اکثر شبکه‌های عصبی، برای آموزش به مقدار زیادی داده‌های آموزشی نیاز داشتند که بویژه جمع‌آوری آن‌ها برای مسئله قطعه‌بندی معنایی بسیار مشکل‌است.

---

[۱] Image segmentation
[۲] Semantic segmentation
[۳] Instance segmentation
[۴] Panoptic Segmentation
[۵] Fully-connected
[۶] Skip connections

شبکه یو-نت[۳] به عنوان یکی از مطرح‌ترین شبکه‌های دارای معماری کدگذار-کدگشا[۷]، در ابتدا برای کاربردهای مرتبط با پزشکی معرفی شد؛ اما عملکرد مناسب آن باعث شد به سرعت در زمینه‌های دیگر نیز گسترش یابد و طی سالها، مدل‌های بسیاری بر اساس آن ارائه شود. شبکه یو-نت، با معماری خاص خود، به نحوی که خواهیم دید، هردو ایراد فوق را تا حد زیادی برطرف می‌کند.

در این گزارش شبکه یو-نت[۳] و مدل‌های بهبود یافته آن شامل یو-نت سه‌بعدی[۴]، یو-نت باقی‌مانده‌ای[۵]، یو-نت++[۶]، اتنشن یو-نت[۷] و مولتی‌رز یو-نت[۸] بررسی خواهد شد. در ادامه این گزارش برای هر شبکه یک فصل در نظر گرفته شده که شامل دو بخش برای توضیح ساختار و بررسی نتایج تجربی آن شبکه می‌باشد. از آن‌جا که هدف از این گزارش بررسی ساختار شبکه‌ها، مستقل از کاربرد آن‌هاست، در بخش مربوط به نتایج تجربی از ذکر بسیاری از جزئیات صرف‌نظر می‌شود. خوانندگان علاقه‌مند برای جزئیات بیشتر به می‌توانند به مراجع مربوطه مراجعه کنند.

۲- شبکه یو-نت

در معماری‌های رایج شبکه‌های کانولوشنی، کانال‌های ویژگی به منظور استخراج اطلاعات معنایی به طور متناوب از لایه‌های ادغام[۸] عبور می‌کنند. لایه‌های ادغام با کاهش اطلاعات فضایی، میدان دید لایه‌های بعدی را افزایش می‌دهند و از این طریق شبکه می‌تواند اطلاعات معنایی را از تصاویر استخراج کند. هرچه از ورودی یک شبکه کانولوشنی به سمت خروجی آن حرکت کنیم، اطلاعات فضایی کاسته شده و بر اطلاعات معنایی افزوده می‌شود؛ بنابراین به نظر می‌رسد بین اطلاعات فضایی و معنایی در این شبکه‌ها مصالحه‌ای وجود دارد که افزایش هریک، موجب کاهش دیگری می‌شود. شبکه یو-نت به این چالش بزرگ از طریق معماری خاص خود و اتصالات پرشی رسیدگی می‌کند. این شبکه دارای یک ساختار متقارن کدگذار-کدگشا می‌باشد(شکل ۲) که در آن کانال‌های ویژگی در سطوح مختلفی از بخش کدگذار، از طریق اتصالات پرشی، به کانال‌های ویژگی کدگشا الحاق[۹] می‌شود. این ساختار برخلاف ساختار شبکه تمام کانولوشنی[۱] که خروجی نهایی را در یک مرحله نمونه افزایی[۱۰] ایجاد می‌کند، باعث می‌شود بازیابی اطلاعات فضایی بهتر صورت بگیرد. تفاوت دیگر این دو شبکه این است که اتصالات پرشی در شبکه یو-نت به جای استفاده از عمل جمع، از عمل الحاق استفاده می‌کند. این ساختار متقارن به همراه تکنیک‌های افزونگی داده، باعث می‌شود شبکه یو-نت بتواند علاوه بر کسب دقت بالا، از تعداد بسیار کمی داده‌های آموزشی یاد بگیرد.

۲ -۱ - ساختار شبکه

شکل ۲ ساختار شبکه یو-نت را نشان می‌دهد. این شبکه از یک بخش کدگذار[۱۱](سمت چپ) و یک بخش کدگشا[۱۲](سمت راست) تشکیل شده‌است. هر مرحله از بخش کدگذار، شامل دولایه کانولوشنی 3x3 با یک تابع فعالسازی[۱۳] ReLU در هرکدام و یک لایه ادغام حداکثر[۱۴] با سایز 2x2 و گام ۲ می‌باشد. در هر مرحله از بخش کدگشا، ابتدا یک لایه نمونه‌افزایی به همراه یک لایه کانولوشنی 2x2(کانولوشن افزاینده[۱۵]) تعداد کانال‌های ویژگی را نصف و ابعاد آن‌ها را دوبرابر می‌کند. سپس کانال‌های ویژگی با کانال‌های برش داده شده متناظر از بخش کدگذار الحاق می‌شوند. در ادامه دولایه کانولوشنی، مشابه آنچه در بخش کدگذار وجود دارد، روی کانال‌های ویژگی اعمال می‌شود. در لایه‌های کانولوشنی این شبکه عمل گسترش مرز با صفر[۱۶] انجام نمی‌شود و بنابراین برای الحاق کانال‌های ویژگی، همانطور که اشاره شد، کانال‌های ویژگی بخش کدگذار بریده می‌شوند. در انتهای بخش کدگشا یک لایه کانولوشنی 1x1 تعداد کانال‌های ویژگی(۶۴) را به تعداد کلاس‌های مسئله تبدیل می‌کند.

---

۷ Encoder-Decoder

۸ Pooling

۹ concatenate

۱۰ Up-sampling

۱۱ Encoder – رانبرگر و همکاران اولین بار از اصلاح Contractive path برای این بخش استفاده کردند. اما در این گزارش، از اصطلاح مرسوم Encoder استفاده می‌شود.

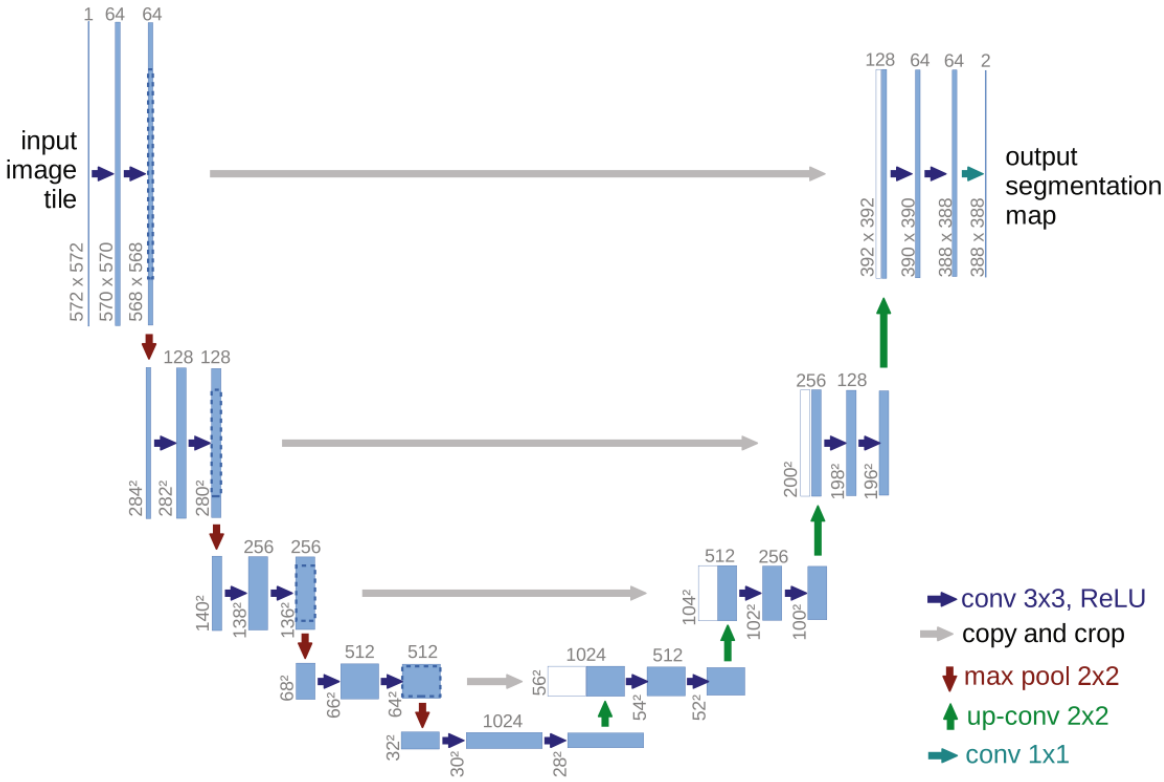
۱۲ Decoder- رانبرگر و همکاران اولین بار از اصلاح Expansive path برای این بخش استفاده کردند.

۱۳ Activation function

۱۴ Max pooling

۱۵ Up-convolution

۱۶ Zero-padding



شکل ۲- ساختار شبکه یو-نت [۳]

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

ب)

Rank	Group name	Warping Error	Rand Error	Pixel Error
	** human values **	0.000005	0.0021	0.0010
1.	u-net	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA [2]	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
⋮				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

الف)

شکل ۳ – نتایج تجربی شبکه یو-نت[۳] - الف) رتبه‌بندی شبکه‌های مختلف در چالش قطعه‌بندی ساختارهای عصبی ب) نتایج شبکه‌های مختلف برحسب معیار IOU روی دو مجموعه دادگان از چالش قطعه‌بندی سلول

۲ -۲ – آموزش شبکه و نتایج تجربی

آموزش شبکه به وسیله تعدادی تصویر و قطعه‌بندی متناظر با آن‌ها و روش گرادیان نزولی تصادفی[۱۷] انجام می‌شود. رانبرگر و همکاران [۳] به منظور محدود نکردن شبکه به تصویر با اندازه ثابت، تصاویر ورودی را به بخش‌های همپوشان تقسیم می‌کنند و خروجی نهایی را از ترکیب آن‌ها بدست می‌آورند. برای افزایش هرچه بیشتر سایز بخش‌های هم‌پوشان و محدود نگه‌داشتن بار محاسباتی، سایز هر دسته[۱۸] به یک کاهش داده شده‌است و بر همین اساس، مقدار ۰.۹۹ برای ضریب مومنتوم انتخاب شده‌است تا تاثیر نمونه‌های قبلی در در بروزرسانی وزن بیشتر باشد. همچنین عمل افزونگی داده‌ها[۱۹]، با ایجاد تعدادی تصویر آموزشی

[۱۷] Stochastic gradient decent

[۱۸] Batch

[۱۹] Data augmentation

جدید با تغییرشکل نمونه‌های فعلی، به منظور مقاوم کردن مدل به تغییرات ظاهری، انجام شده‌است. پس از اعمال تابع سافت–ماکس[۲۰] بر روی کانال‌های ویژگی خروجی، از تابع آنتروپی متقابل وزن‌دار[۲۱] به عنوان تابع انرژی شبکه استفاده شده‌است. از ذکر سایر جزئیات به دلیل وابسته بودن آن‌ها به کاربرد مدنظر نویسندگان، صرفنظر می‌شود.

همانطور که اشاره شد نسخه اولیه شبکه یو–نت برای کاربردهای قطعه‌بندی در زمینه پزشکی توسعه یافت و از این رو، آزمایشات انجام شده توسط رانبرگر و همکاران[۳] همگی مربوط به این کاربردها هستند. به طور دقیق‌تر شبکه یو–نت در مسئله قطعه‌بندی ساختارهای عصبی در تصاویر الکترومیکروسکپی(چالش قطعه‌بندی EM) و مسئله قطعه‌بندی سلول‌ها در تصاویر میکروسکپ نوری(چالش ردیابی سلول ISBI) مورد ارزیابی قرار گرفت. با توجه به شکل ۳ که نتایج این چالش‌ها را نشان می‌دهد، شبکه یو–نت در هر دو چالش موفق به کسب بهترین عملکرد شد.

۳– شبکه یو–نت سه‌بعدی

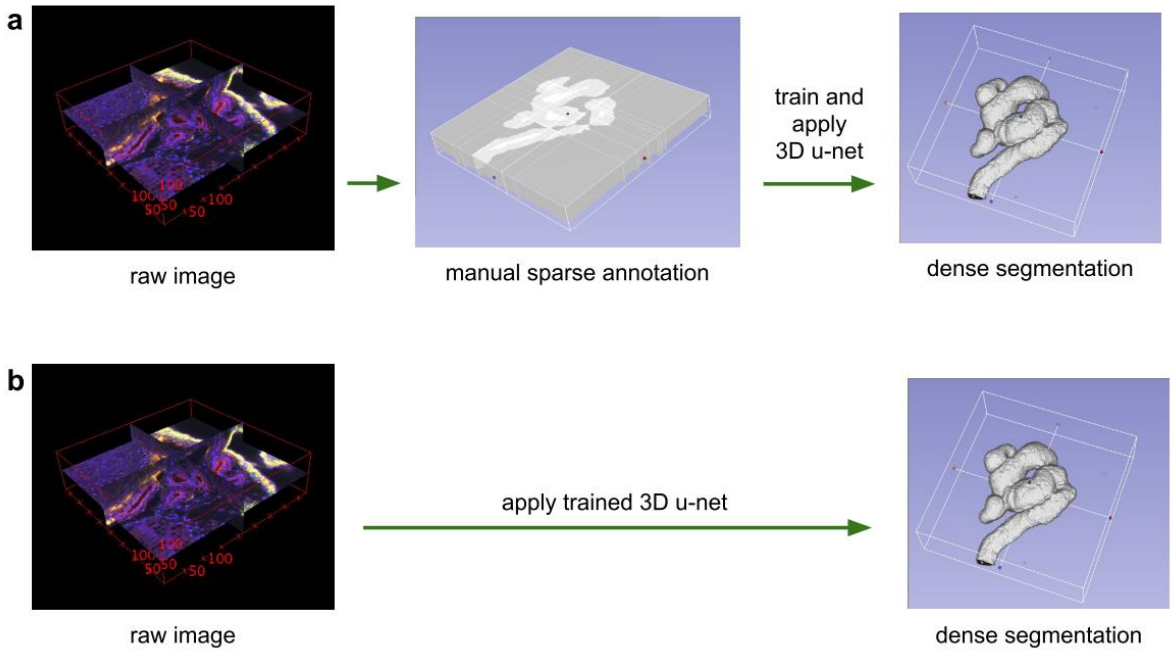
یکی از مهم‌تری چالش‌های بینایی ماشین در علم پزشکی، قطعه‌بندی داده‌های حجمی می‌باشد. برچسب گذاری چنین داده‌هایی باید از طریق برچسب گذاری مجزای برش‌های دوبعدی آن صورت بگیرد. این کار علاوه بر زمان‌بر بودن، ناکارآمد نیز هست؛ زیرا معمولا برش‌های همسایه در یک داده حجمی تقریبا اطلاعات یکسانی دارند. بنابراین مشکل اساسی در این مسائل کمبود داده‌های آموزشی و سخت بودن تهیه آن‌هاست.

همانطور که پیش از این اشاره شد، یکی از آن مهم‌ترین مزیت‌های شبکه یو–نت، توانایی بالای آن در یادگیری از داده‌های آموزشی با تعداد کم می‌باشد. این ویژگی و سایر ویژگی‌های مثبت شبکه یو–نت موجب شد چیچک و همکاران [۴]، برای اولین بار شبکه یو–نت سه بعدی را براساس شبکه یو–نت ارائه کنند. این شبکه می‌تواند به دو روش برای قطعه‌بندی داده‌های حجمی مورد استفاده قرار بگیرد(شکل ۴). در روش اول با استفاده از تعدادی از برش‌های دوبعدی از یک حجم، که به صورت دستی قطعه‌بندی شده‌اند، کل حجم قطعه‌بندی می‌شود(روش نیمه اتوماتیک) و در روش دوم، یک داده حجمی کاملا جدید با استفاده از شبکه آموزش یافته روی نمونه‌های قبلی، قطعه‌بندی می‌شود(روش تمام اتوماتیک). در ادامه این گزارش به بررسی ساختار شبکه یو–نت سه بعدی و برخی از نتایج آن می‌پردازیم.

---

[۲۰] Soft-max

[۲۱] Weighted cross-entropy



شکل ۴–دو روش عملکرد شبکه یو–نت سه‌بعدی برای قطعه‌بندی داده‌های سه‌بعدی a) قطعه‌بندی حجمی با استفاده از چند برش دوبعدی از همان حجم که به صورت دستی قطعه‌بندی شده‌اند. b) قطعه‌بندی سه‌بعدی کاملا خودکار[۴]

۳ –۱ – ساختار شبکه

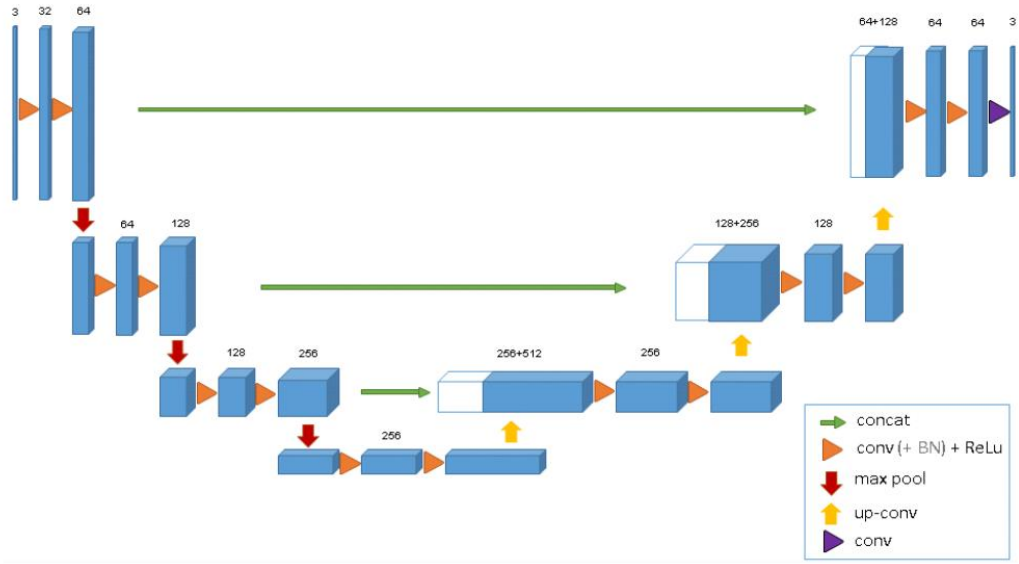
مشابه ساختار یو–نت، شبکه یو–نت سه‌بعدی نیز از دوبخش کدگذار و کدگشا با ساختاری مشابه تشکیل شده‌است(شکل ۵). همه لایه‌های کانولوشنی، ادغام و کانولوشنی افزاینده در شبکه یو–نت با نوع سه بعدی آن(با عمق مساوی سایر ابعاد) جایگزین شده‌اند. تعداد گام‌های بخش‌های کدگذار و کدگشا، به منظور محدود نگه داشتن تعداد پارامترها، کاهش یافته‌است. همچنین برای جلوگیری از گلوگاه‌ها[۲۲]، تعداد فیلترها قبل از هر لایه ادغام دوبرابر شده‌است[۹]. تفاوت مهم دیگر شبکه یو–نت سه‌بعدی انجام عمل نرمال‌سازی دسته‌ای[۲۳] قبل از هر واحد ReLU می‌باشد(در شبکه یو–نت استاندارد، نرمال سازی دسته‌ای وجود ندارد). ورودی شبکه یک بلاک وکسل[۲۴] با ابعاد 132x132x166 از تصویر(حجم) ورودی با ۳ کانال و خروجی آن 44x44x28 وکسل، به ترتیب در سه جهت x، y و z می‌باشد؛ بنابراین مانند شبکه یو–نت استاندارد تصویر ورودی با سایز دلخواه به چند بلاک تقسیم شده و هر بلاک مجزا پردازش می‌شود.

---

[۲۲] Bottleneck– در شبکه‌های کانولوشنی به لایه‌ای گفته می‌شود که حجم کانال‌های ویژگی آن نسبت به لایه قبل به شدت کمتر است.

[۲۳] Batch normalization

[۲۴] voxel tile– وکسل کوچکترین عنصر سازنده یک حجم سه بعدی است(مشابه پیکسل در تصاویر دوبعدی)



شکل ۵- ساختار شبکه یو-نت سه‌بعدی[۴]

۳ -۲ - آموزش شبکه و نتایج تجربی

پس از اعمال سافت-ماکس روی خروجی شبکه، از تابع آنتروپی متقابل وزن‌دار به عنوان تابع انرژی شبکه استفاده می‌شود. همانطور که در مقدمه این بخش اشاره شد، شبکه از روی تعداد کمی از برش‌های دوبعدی(که به صورت دستی قطعه‌بندی شده‌اند) آموزش می‌بیند؛ بنابراین برش‌هایی که قطعه‌بندی نشده‌اند، وزن صفر خواهند داشت. به منظور ایجاد مقاومت نسبت به تغییر شکل، داده های آموزشی جدید با ایجاد تغییر شکل در داده‌های فعلی ایجاد می‌شوند و درنهایت، از روش گرادیان نزولی تصادفی برای آموزش شبکه استفاده می‌شود.

در زمان انتشار مقاله چیچک و همکاران[۴] کارهای اندکی به قطعه‌بندی تصاویر حجمی پزشکی پرداخته‌بودند و دوکاری که چیچک و همکاران[۴] در مقاله خود به آن‌ها اشاره کردند، به دلایلی که در اینجا به آن نمی‌پردازیم، روی مجموعه دادگان آن‌ها قابل استفاده نبودند. به همین دلیل متاسفانه نمی‌توان نتایج شبکه یو-نت سه‌بعدی را به صورت دقیق بررسی کرد. مجموعه دادگان مورد بررسی تنها شامل سه تصویر حجمی بوده که هر بار یکی از آنها به عنوان داده تست استفاده می‌شود. نتایج در شکل ۶ برای دو حالت نیمه اتوماتیک و تمام اتوماتیک ارائه شده‌است. در مجموع نتایج از نظر معیار IoU در محدوده مناسبی قرار دارد که نشان دهنده عملکرد مطلوب این شبکه است. نکته جالب این نتایج این است که انجام عمل نرمال‌سازی دسته‌ای در برخی موارد باعث تضعیف عملکرد شبکه شده‌است.

Table 1: Cross validation results for semi-automated segmentation (IoU)

test slices	3D w/o BN	3D with BN	2D with BN
subset 1	0.822	0.855	0.785
subset 2	0.857	0.871	0.820
subset 3	0.846	0.863	0.782
average	0.842	0.863	0.796

Table 2: Effect of # of slices for semi-automated segmentation (IoU)

GT slices	GT voxels	IoU S1	IoU S2	IoU S3
1,1,1	2.5%	0.331	0.483	0.475
2,2,1	3.3%	0.676	0.579	0.738
3,3,2	5.7%	0.761	0.808	0.835
5,5,3	8.9%	0.856	0.849	0.872

Table 3: Cross validation results for fully-automated segmentation (IoU)

test volume	3D w/o BN	3D with BN	2D with BN
1	0.655	0.761	0.619
2	0.734	0.798	0.698
3	0.779	0.554	0.325
average	0.723	0.704	0.547

شکل ۶- نتایج تجربی شبکه یو-نت سه‌بعدی در دو حالت نیمه اتوماتیک و تمام اتوماتیک[۴]

۴- شبکه یو-نت باقی‌مانده‌ای

اگرچه تاثیر افزایش عمق در بهبود عملکرد شبکه‌های کانولوشنی عمیق مشهود است، اما یک مشکل شبکه‌های بسیار عمیق مشکل گرادیان محوشونده/انفجاری[۲۵] در آموزش آن‌هاست. یادگیری باقی‌مانده‌ای عمیق[۲۶] [۱۰] به عنوان روشی برای حل این مشکل، پس از معرفی مورد توجهات فراوان قرار گرفت. استفاده از بلاک‌های باقی‌مانده‌ای که دارای نگاشت همانی[۲۷] از ورودی به خروجی هستند، یادگیری شبکه‌های بسیار عمیق را ممکن کرده و باعث افزایش عمق شبکه‌های عمیق و درنتیجه بهبود عملکرد آن‌ها شد.

به نظر می‌رسد اولین بار ژانگ و همکاران[۵] از بلاک‌های باقی‌مانده‌ای در شبکه یو-نت استفاده کردند. ایده آن‌ها این بود که این کار باعث می‌شود آموزش شبکه یو-نت، به عنوان یک شبکه عمیق، بهود یابد. ژانگ و همکاران از این شبکه برای مسئله استخراج راه‌ها از تصاویر هوایی، که به نوعی یک مسئله قطعه‌بندی است، استفاده کردند.

۴-۱- ساختار شبکه

ایده ژانگ و همکاران[۵] برای شبکه یو-نت باقی‌مانده‌ای بسیار ساده است؛ استفاده از یک بلاک باقی‌مانده‌ای (شکل ۷-الف) به جای بلاک‌های عادی شبکه یو-نت(شکل ۷-ب). این بلاک‌ها علاوه بر نگاشت همانی، شامل دو لایه نرمال‌سازی دسته‌ای(BN) نیز می‌باشند. ساختار کل شبکه در شکل ۸ نمایش‌داده شده‌است. در اینجا بخش میانی شبکه به علت ساختار متفاوت، پل[۲۸] نامیده می‌شود؛ بنابراین شبکه به ترتیب از بخش‌های کدگذار، پل و کدگشا تشکیل شده‌است. به جای استفاده از لایه‌ادغام برای کاهش ابعاد کانال‌های ویژگی، گام اولین لایه کانولوشنی در هربلاک از بخش کدگذار ۲ در نظر گرفته شده‌است. مشابه شبکه یو-نت قبل از هر بلاک از بخش کدگشا، یک عمل نمونه افزایی و متعاقب آن، عمل الحاق کانال‌های ویژگی فعلی با کانال‌های ویژگی متناظر از بخش کدگذار صورت می‌گیرد. همچنین در این شبکه برخلاف شبکه یو-نت استاندارد، عمل گسترش مرز با صفر صورت می‌گیرد و بنابراین برش کانال‌های ویژگی کدگذار نیاز نیست. در مجموعه این شبکه دارای ۱۵ لایه کانولوشنی است که در مقایسه با ۲۳ لایه کانولوشنی شبکه یو-نت استاندارد، کمتر است(در این شبکه تعداد گام های کدگذار و کدگشا به سه کاهش یافته است).

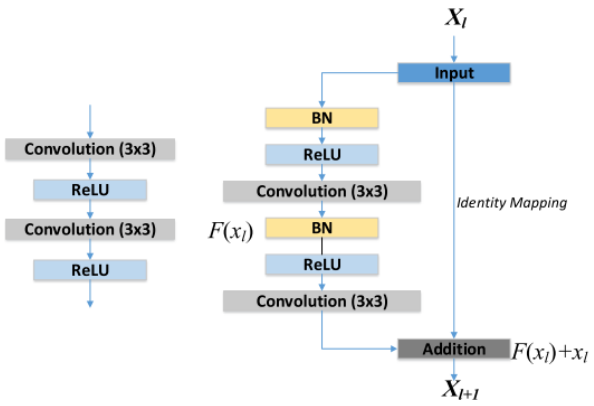
---

[۲۵] Vanishing/Exploding Gradient

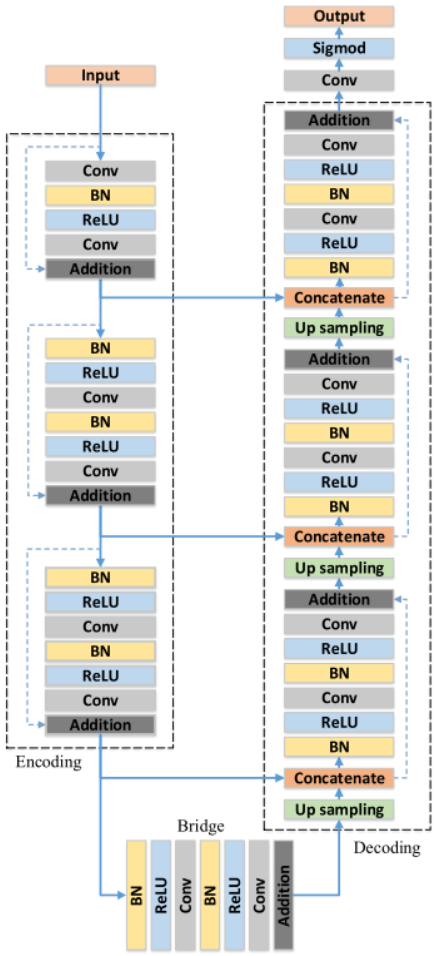
[۲۶] Deep Residual Learning

[۲۷] Identity mapping

[۲۸] Bridge



شکل ۷ – الف) بلاک سازنده شبکه یو-نت باقی‌مانده‌ای. ب) بلاک سازنده شبکه یو-نت[۵]



شکل ۸- ساختار شبکه یو-نت باقی‌مانده‌ای[۵]

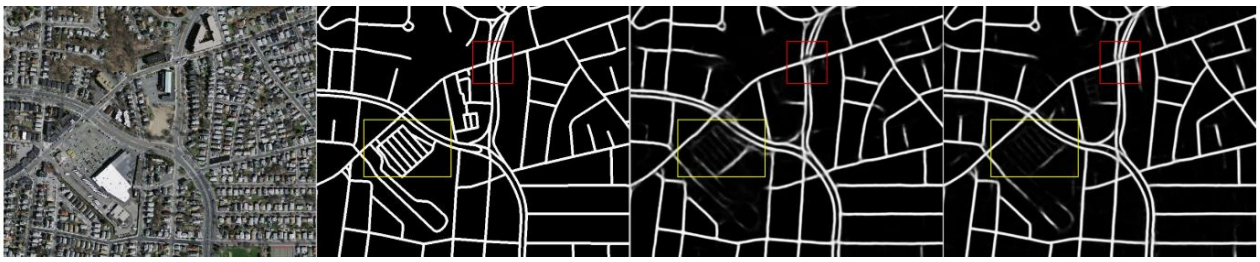
۴ -۲ - آموزش شبکه و نتایج تجربی

در این شبکه، از خطای میانگین مربعات به عنوان تابع انرژی استفاده می‌شود، اما استفاده از سایر توابع مانند آنتروپی متقابل نیز ممکن است. برای آموزش، از روش گرادیان نزولی تصادفی استفاده می‌شود. مشابه شبکه یو-نت و یو-نت سه بعدی، در اینجا نیز تصاویر ورودی به بلاک‌های همپوشان تقسیم شده و خروجی نهایی از ترکیب آن‌ها بوجود می‌آید. این کار باعث می‌شود شبکه یو-نت باقی‌مانده‌ای قابلیت پردازش تصویر با هر ابعادی را داشته‌باشد. با توجه به کاربرد این شبکه(استخراج راه‌ها از تصاویر هوایی)، افزونگی داده‌ها انجام نمی‌شود و برخلاف شبکه‌های پیشین، هر دسته شامل ۸ تصویر آموزشی می‌باشد.

نتایج ارزیابی شبکه روی مجموعه دادگان راه‌های ماساچوست[۱۱] با معیار نقطه سربه‌سر[۲۹] در شکل ۹ مشاهده می‌شود. اگرچه تعداد پارامترهای شبکه یو-نت باقی‌مانده‌ای حدود یک چهارم شبکه یو-نت استاندارد می‌باشد؛ اما مطابق این شکل، یونت باقی‌مانده‌ای از یو-نت و سایر شبکه‌ها عملکرد بهتری داشته‌است.

Model	Breakeven point
Mnih-CNN [2]	0.8873
Mnih-CNN+CRF [2]	0.8904
Mnih-CNN+Post-Processing [2]	0.9006
Saito-CNN [5]	0.9047
U-Net [24]	0.9053
ResUnet	**0.9187**

شکل ۹- ارزیابی شبکه‌های مختلف روی مجموعه دادگان راه‌های ماساچوست[۵]



شکل ۱۰- نمونه نتایج داده آزمون در دیتاست ماسوچوست – از چپ به راست: تصویر ورودی، برچسب‌های واقعی(حقیقت مبنا[۳۰])، خروجی یو-نت، خروجی یونت-باقی‌مانده‌ای[۵]

۵- شبکه یو-نت++

ایده اصلی یو-نت[۳] استفاده از اتصالات پرشی به منظور ترکیب کانال‌های ویژگی بخش کدگذار و کدگشا بود. همانطور که پیش از این اشاره شد، فرض آن بود که اتصالات پرشی باعث بازیابی اطلاعات فضایی می‌شود که توسط لایه‌های ادغام از بین می‌رود. این ایده به ظاهر ساده موجب عملکرد بسیار مطلوب یو-نت و افزایش استفاده از آن در زمینه‌های مختلف شد. به طور خاص کاربرد یو-نت در پزشکی، کاربردی که مخترعان این شبکه به دنبال آن بودند، نقش بسیاری در پیشرفت‌های اولیه شبکه یو-نت داشت. به طور کلی قطعه‌بندی در تصاویر پزشکی نسبت به تصاویر عادی نیاز به حساسیت بیشتری دارد و بویژه در سیستم‌های پزشکی تمام خودکار، هر خطای کوچک ممکن است هزینه بسیاری داشته‌باشد. ژو و همکاران[۶] به همین منظور به دنبال بهبود شبکه یو-نت، شبکه یو-نت++ را معرفی کردند. نظر آن‌ها این بود که انتقال مستقیم کانال‌های ویژگی از بخش کدگذار به بخش کدگشا، به دلیل اینکه کانال‌های ویژگی این دوبخش از نظر معنایی با هم تفاوت دارند، احتمالا آموزش و کارآیی شبکه را با مشکل مواجه می‌کند؛ در نتیجه اتصالات پرشی را به گونه‌ای که در ادامه خواهیم دید ارتقا دادند.

۵ -۱ - ساختار شبکه

در شکل ۱۱ شمای کلی ساختار شبکه یو-نت++ ارائه شده‌است. چیزی که شبکه یو-نت++ را از شبکه یو-نت(اجزای مشکی در شکل ۱۱) متمایز می‌کند، اتصالات پرشی ارتقا یافته(اجزای آبی و سبز در شکل ۱۱) و استفاده از نظارت عمیق[۱۲](اجزای قرمز در شکل ۱۱) می‌باشد. در این شبکه اتصالات پرشی ساده

---

[۲۹] Break-even point

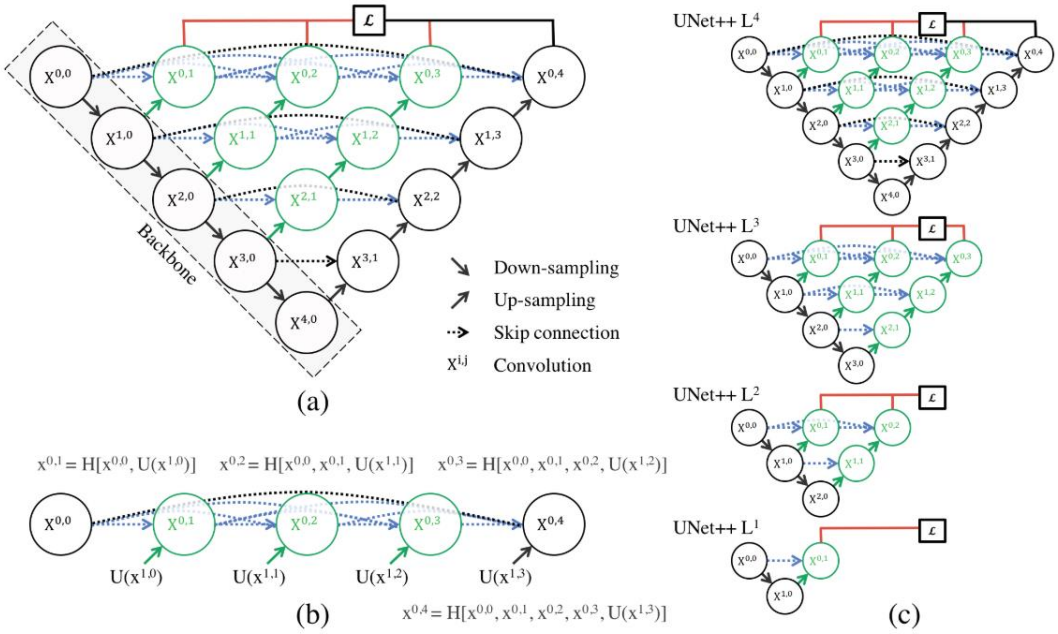
[۳۰] Ground truth

با ساختاری شبیه به یک بلاک کانولوشنی متراکم[۳۱][۱۳] جایگزین شده‌اند. تعداد لایه‌های کانولوشنی در هر اتصال به سطح آن در ساختار شبکه وابسته است. به عنوان مثال اتصال بین $X^{1,0}$ و $X^{1,3}$ شامل ۲ لایه کانولوشنی می‌باشد که ورودی هر یک از الحاق خروجی تمام لایه‌های قبلی همان سطح و خروجی نمونه‌افزایی شده از لایه‌متناظر یک سطح پایین‌تر، تشکیل می‌شود. به طور رسمی اگر $x^{i,j}$ خروجی گره $X^{i,j}$ باشد، که i سطح گره در طول لایه کدگذار و j شماره گره در طول اتصال پرشی است، تعریف می‌کنیم:

$$x^{i,j} = \begin{cases} H(x^{i-1,j}) & if\ j = 0 \\ H([x^{i,k}]_{k=0}^{j-1}, U(x^{i+1,j-1})) & if\ j > 0 \end{cases} \quad (1)$$

در معادله ۱، H نشان دهنده عمل کانولوشن و سپس تابع فعالسازی، U نشان‌دهنده نمونه‌افزایی و [] نشان دهنده الحاق می‌باشد. ایده مخترعان این بوده‌است که چنین ساختاری برای اتصالات پرشی، کانال‌های ویژگی کدگذار و کدگشا را از نظر معنایی به هم نزدیک می‌کند و در نتیجه، آموزش شبکه آسان‌تر خواهد شد. شکل ۱۱–b اتصال پرشی در بالاترین سطح شبکه را با جزئیات نشان می‌دهد.

شبکه یو–نت++ با بهره‌گیری از نظارت عمیق[۱۲] می‌تواند در دو حالت بکار گرفته شود: ۱) حالت دقیق که در آن خروجی از میانگین همه شاخه‌های قطعه‌بندی بدست می‌آید. ۲) حالت سریع که در آن خروجی تنها از یکی از شاخه‌های قطعه‌بندی بدست می‌آید. انتخاب شاخه قطعه‌بندی در حالت ۲، پیچیدگی و سرعت شبکه را مشخص می‌کند. شکل ۱۱–c نشان می‌دهد که چگونه انتخاب شاخه قطعه‌بندی موجب ایجاد شبکه‌هایی با ساختار مختلف می‌شود.



شکل ۱۱– ساختار شبکه یو–نت++ a) ساختار کلی شبکه b) جزئیات یک نمونه از مسیرهای پرشی ارتقا یافته c) سطوح مختلف استفاده از نظارت عمیق [۶]

۵ –۲ – آموزش شبکه و نتایج تجربی

با توجه به ساختار تودرتوی شبکه یو–نت++، کانال‌های ویژگی با رزولوشن کامل در سطح‌های معنایی مختلفی ساخته می‌شود($x^{0,j}$,j=1,...,4) که در نظارت عمیق با هم ترکیب می‌شوند. برای هریک از این سطح‌های معنایی، از ترکیب آنتروپی متقابل دودویی و ضریب تشابه سرنسن[۳۲] به عنوان تابع انرژی استفاده می‌شود که به صورت زیر تعریف می‌شود:

---

[۳۱] Dense convolution block

[۳۲] Srensen-Dice similarity coefficient

$$L(Y,Y') = -\frac{1}{N}\sum_{b=1}^{N}\left(\frac{1}{2}.Y_b log Y'_b + \frac{2.Y_b.Y'_b}{Y_b + Y'_b}\right) \qquad (2)$$

در معادله ۲، Y_b و Y'_b به ترتیب مقادیر پیش‌بینی شده و مقادیر واقعی برای b–امین تصویر هستند که به بردار تبدیل شده‌اند و N نشان دهنده سایز دسته می‌باشد. درنهایت از روش آدام[۳۳] برای آموزش شبکه استفاده می‌شود.

ژو و همکاران[۶] شبکه ابداعی خود را روی چهاردیتاست مختلف ارزیابی کردند. در شکل ۱۲ نتایج این ارزیابی براساس معیار IoU با دو ساختار دیگر از شبکه یو–نت ارائه شده‌است. شبکه یو–نت++ در هر دوحالت با نظارت عمیق و بدون آن، از دو شبکه دیگر با اختلاف قابل قبولی بهتر عمل کرده‌است.

Architecture	Params	Dataset			
		Cell nuclei	Colon polyp	Liver	Lung nodule
U-Net [9]	7.76M	90.77	30.08	76.62	71.47
Wide U-Net	9.13M	90.92	30.14	76.58	73.38
UNet++ w/o DS	9.04M	**92.63**	**33.45**	79.70	76.44
UNet++ w/ DS	9.04M	92.52	32.12	**82.90**	**77.21**

شکل ۱۲ – مقایسه نتایج ارزیابی شبکه یو–نت++ و دو شبکه یو–نت دیگر بر اساس معیار IoU. شبکه یو–نت++ به دو صورت با نظارت عمیق و بدون آن ارزیابی شده‌است.

۶– شبکه یو–نت با مکانیزم توجه

اگرچه مکانیزم توجه اولین بار برای کاربردهای مرتبط با پردازش متن معرفی شد، اما به سرعت درسایر زمینه‌ها از جمله بینایی ماشین[۱۴] و به طور خاص مسئله قطعه‌بندی[۱۵] نیز از مفهوم آن استفاده شد. به طور کلی مکانیزم توجه در مدل‌های بینایی ماشین، با برجسته‌تر کردن بخش‌های مهم‌تر از کانال‌های ویژگی می‌توانند موجب بهبود عملکرد مدل شوند. اشلمپر و همکاران[۷] با ارائه سازوکاری تحت عنوان گیت توجه[۳۴] ،که قابلیت استفاده در همه شبکه‌های کانولوشنی را دارد، نوع ارتقا یافته‌ای از شبکه یو–نت با نام اتنشن یو–نت[۳۵] معرفی کردند. آن‌ها کاربرد گیت توجه را در مدل‌های دسته‌بندی نیز نشان دادند، اما در این گزارش صرفا کاربرد آن در شبکه یو–نت بررسی خواهد شد. ابتکار شبکه اتنشن یو–نت این است که نقشه‌های منتقل شده از طریق اتصالات پرشی، قبل از الحاق از گیت‌های توجه عبور کنند. استدلال آن‌ها این است که از این طریق بخش‌های مهم‌تر از این نقشه‌ها، به طریقی که خواهیم دید، برجسته شوند.

در بخش بعد ابتدا مدل گیت توجه را معرفی می‌کنیم و سپس نشان خواهیم داد که چگونه شبکه اتنشن یو–نت با استفاده از آن ساخته خواهد شد.

۶ –۱ – ساختار شبکه

ابتدا ساختار یک گیت توجه را تشریح می‌کنیم. این گیت قابلیت استفاده در شبکه‌های کانولوشنی سه‌بعدی را نیز دارد. فرض کنیم $x^l = \{x_i^l\}_{i=1}^n$ نقشه‌های ویژگی[۳۶] لایه دلخواه l و x_i^l بردار ویژگی برای پیکسل i باشد که اندازه آن F^l (تعداد کانال‌های ویژگی) است. برای هر گیت توجه، ضرایب توجه $\alpha^l = \{\alpha_i^l\}_{i=1}^n$ را تولید می‌کند به نحوی که $\alpha \in [0,1]$ باشد. خروجی گیت توجه $\hat{x}^l = \{\alpha_i^l x_i^l\}_{i=1}^n$ می‌باشد که در آن هر بردار ویژگی با توجه به ضریب توجه، مقیاس یافته‌است. در ادامه نحوه محاسبه ضرایب توجه را تشریح می‌کنیم. در شبکه‌های کانولوشنی، برای استخراج اطلاعات مفهومی، نقشه‌های ویژگی به طور متناوب نمونه‌کاهی می‌شوند. نقشه‌های ویژگی در لایه‌های انتهایی، اطلاعاتی مانند اشیا هدف و محل تقریبی آن‌ها را ارائه می‌کنند. فرض کنیم g

---

[۳۳] Adam

[۳۴] Attention gate

[۳۵] Attention U-Net – در اینجا کلمه اتنشن به علت تاکید بر نام خاص شبکه، ترجمه نمی‌شود.

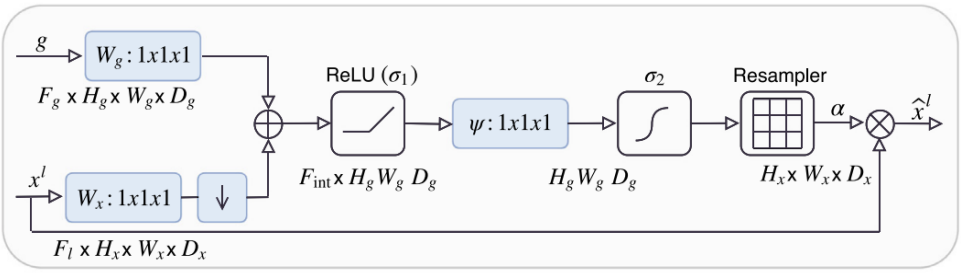
[۳۶] Feature maps

نقشه‌های ویژگی از یکی از لایه‌های انتهایی شبکه باشد که چنین اطلاعاتی را شامل می‌شود. گیت‌های توجه با در نظر گرفتن g و x_i^l، ضریب α_i^l را استخراج می‌کند. به بیان ساده‌تر، گیت توجه یاد میگیرید که کدام بخش از نقشه‌های ویژگی مهم‌تر است و آن را در ضریب بزرگتری ضرب می‌کند. به طور رسمی، ضرایب توجه به روش توجه جمعی[۳۷][۱۶] از معادلات زیر محاسبه می‌شود:

$$q_{att,i}^l = \psi^T \left(\sigma_1\left(W_x^T x_i^l + W_g^T g + b_{xg}\right)\right) + b_\psi \qquad (3)$$

$$\alpha^l = \sigma_2\left(q_{att}^l\left(x^l, g; \Theta_{att}\right)\right) \qquad (4)$$

در روابط فوق $\sigma_1(x)$ یک تابع فعالسازی مانند ReLU و $\sigma_2(x)$ یک تابع نرمالسازی می‌باشد(در اتنشن یو-نت از سیگموئید استفاده می‌شود). همانطور که در معادله ۴ نشان داده شده‌است. یک گیت توجه را می‌توان با مجموعه پارامترهایش(Θ_{att}) شامل تبدیلات خطی $W_x \in R^{F_l \times F_{int}}$، $W_g \in R^{F_g \times F_{int}}$ و بایاس‌های $b_\psi \in R$ و $b_{xg} \in R^{F_{int}}$، توصیف کرد. تبدیلات خطی به وسیله کانولوشن 1x1x1 انجام می‌شود. در شکل ۱۳ ساختار یک گیت توجه ارائه شده‌است. از آنجا که کانال‌های ویژگی در g و x^l دارای ابعاد متفاوتی هستند، x^l نمون‌کاهی شده و سپس ضرایب توجه α^l نمونه‌افزایی می‌شوند.

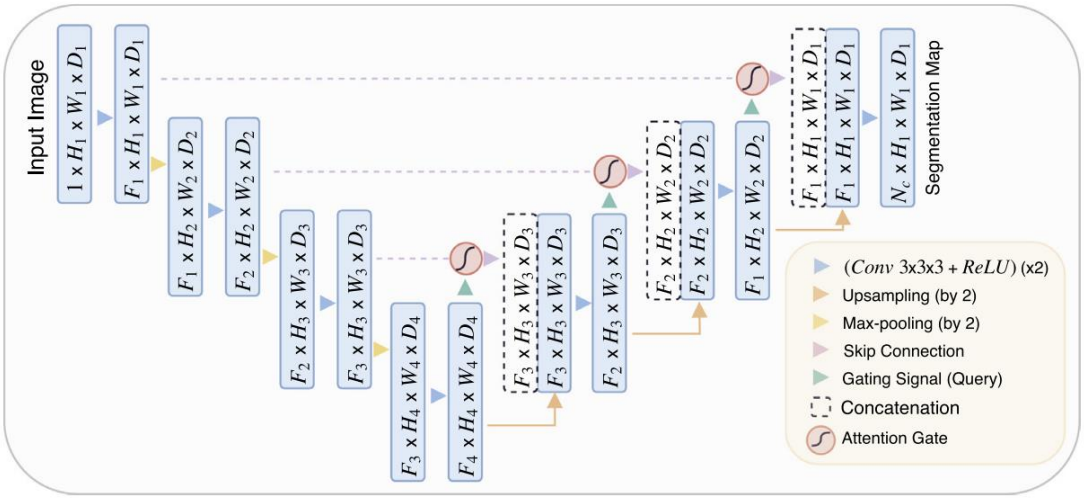


شکل ۱۳-ساختار گیت توجه[۷]

اشلمپر و همکاران[۷] گیت توجه را در ساختار یو-نت استاندارد سه‌بعدی استفاده‌کردند. نقشه‌های ویژگی بخش کدگذار که در ساختار استاندارد یو-نت به طور مستقیم به نقشه‌های ویژگی کدگشا الحاق می‌شد، در ساختار اتنشن یو-نت از گیت‌های توجه عبور کرده و در نتیجه بخش‌های مهم‌تر آن برجسته می‌شود. همانطور که در شکل ۱۴ مشاهده می‌شود، در این شبکه ورودی گیت توجه g، کانال‌های ویژگی قدم قبلی کدگشا است. همانطور که گفته شد این شبکه ارتقا یافته شبکه یو-نت سه‌بعدی است اما به سادگی می‌توان نوع دوبعدی آن‌را نیز ارائه داد.

---

[۳۷] Additive attention



شکل ۱۴ ساختار اتنشن یو-نت. N_c نشان دهنده تعداد کلاس‌های مسئله است[۷]

۶ -۲ - آموزش شبکه و نتایج تجربی

پارامترهای گیت توجه را می‌توان از روش‌های معمول انتشار به عقب[۳۸] آموزش داد. نکته مهم در مورد گیت‌های توجه این است که نه تنها در گذر جلورو[۳۹] بلکه در گذر پشت‌رو[۴۰] نیز نواحی مهم برجسته می‌شوند. برای روشن شدن این موضوع گرادیان خروجی گیت توجه را نسبت به پارامترهای لایه کانولوشنی l-1 (Φ^{l-1}) بدست می‌آوریم:

$$\frac{\partial \hat{x}_i^l}{\partial \Phi^{l-1}} = \frac{\partial(\alpha_i^l f(x_i^{l-1}; \Phi^{l-1}))}{\partial \Phi^{l-1}} = \alpha_i^l \frac{\partial(f(x_i^{l-1}; \Phi^{l-1}))}{\partial \Phi^{l-1}} + \frac{\partial \alpha_i^l}{\partial \Phi^{l-1}} x_i^l \qquad (5)$$

عبارت اول از سمت راست معادله فوق در ضریب α_i^l ضرب شده‌است. این بدین معناست که گرادیان بخش‌های مهم برجسته‌تر شده و درنتیجه وزن لایه‌های پیشین، بیشتر براساس نواحی مهم‌تر از تصویر بروزرسانی می‌شود. برای آموزش شبکه از روش آدام و سایز دسته از ۲ و ۴ استفاده شده‌است. تکنیک‌های افزونگی داده، مشابه یو-نت سه‌بعدی، در اینجا نیز انجام می‌شود و درنهایت تابع انرژی شبکه، سرنسن می‌باشد.

اشلمپر و همکاران[۷] شبکه اتنشن یو-نت را روی دو مجموعه دادگان سه‌بعدی مرتبط با کاربرد پزشکی ارزیابی کردند که در این گزارش یک مورد از آن ارائه می‌شود. همانطور که در شکل ۱۵ مشاهده می‌شود، شبکه اتنشن یو-نت در اکثر موارد با اختلاف قابل قبولی بهتر از شبکه یو-نت عمل کرده‌است؛ اگرچه تعداد پارامترهای آن نیز نسبت به یو-نت بیشتر است. به همین منظور اشلمپر و همکاران[۷] طی آزمایشات دیگر نشان دادند که با افزایش پارامترهای یو-نت، عملکرد آن بهبود چندانی نمی‌یابد و بنابراین عملکرد شبکه اتنشتن یو-نت، فراتر از صرفاً افزایش پارامتر هاست.

---

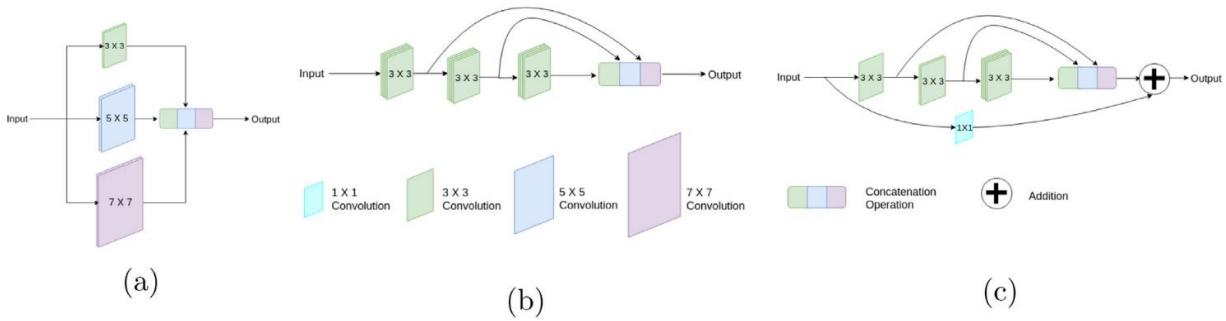
[۳۸] Back-propagation
[۳۹] Forward pass
[۴۰] Background

Method Train/Test split	U-Net 120/30	Att U-Net 120/30	U-Net 30/120	Att U-Net 30/120
Pancreas DSC	0.814 ± 0.116	**0.840 ± 0.087**	0.741 ± 0.137	**0.767 ± 0.132**
Pancreas precision	0.848 ± 0.110	0.849 ± 0.098	0.789 ± 0.176	**0.794 ± 0.150**
Pancreas recall	0.806 ± 0.126	**0.841 ± 0.092**	0.743 ± 0.179	**0.762 ± 0.145**
Pancreas S2S Dist (mm)	2.358 ± 1.464	**1.920 ± 1.284**	3.765 ± 3.452	3.507 ± 3.814
Spleen DSC	0.962 ± 0.013	0.965 ± 0.013	0.935 ± 0.095	**0.943 ± 0.092**
Kidney DSC	0.963 ± 0.013	0.964 ± 0.016	0.951 ± 0.019	0.954 ± 0.021
Number of params	5.88 M	6.40 M	5.88 M	6.40 M
Inference time	0.167 s	0.179 s	0.167 s	0.179 s

شکل ۱۵– مقایسه شبکه اتنشن یو–نت با شبکه یو–نت روی مجموعه دادگان 150 - CT[۷]

۷– شبکه مولتی‌رز یو–نت

یکی از چالش‌های مهم در اکثر مدل‌های بینایی ماشین، حضور سوژه‌ها با اندازه مختلف در تصاویر ورودی است که می‌تواند منجر به کاهش دقت مدل شود. یکی از اولین راه حل‌ها برای غلبه بر این مشکل، توسط سگدی و همکاران[۱۷]، استفاده از فیلتر با سایزهای مختلف به صورت موازی بود[۱۷](شکل ۱۶ –a). مشکل این کار این است که اعمال موازی فیلترهای با سایز مختلف، حافظه مورد نیاز را به شدت افزایش می‌دهد. راه بهتر، استفاده از این ایده است که فیلترهای 3x3 که به صورت پی‌درپی قرار دارند، یک فیلتر بزرگتر را شبیه‌سازی می‌کنند[۹] (شکل ۱۶–b). به عنوان مثال جفت لایه کانولونشی 3x3 که در هر گام از شبکه یو–نت استاندارد قرار دارد، در واقع همانند یک فیلتر 5x5 عمل می‌کنند. ابتهاج و رحمان[۸] با بهره‌گیری از این ایده، به نحوی که خواهیم دید، در هر گام از شبکه یو–نت، فیلترهایی با سایز 3x3، 5x5 و 7x7 اعمال می‌کنند تا از این طریق، عملکرد یو–نت برای تشخیص سوژه‌هایی با مقیاس مختلف، بهبود یابد. بهبود دیگر آن‌ها، کاهش اختلاف معنایی بخش‌های کدگذار و کدگشا(مشابه یو–نت++) به طریقی است که در ادامه خواهیم دید.

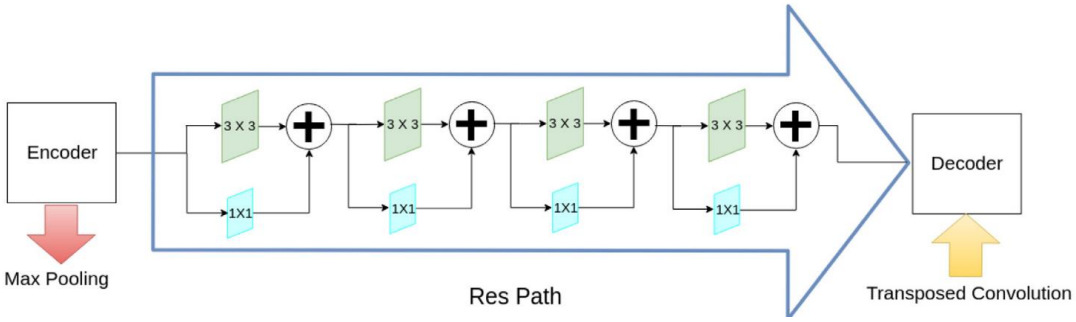


شکل ۱۶– راهکارهای مختلف پیشنهادی برای استخراج ویژگی با مقیاس‌های مختلف – a) اعمال موازی فیلتر با سایزهای مختلف مشابه پیشنهاد سگدی و همکاران[۹] b) اعمال متناوب فیلترهای 3x3 می‌تواند اعمال فیلترهای بزرگتر را شبیه‌سازی کند. c) شکل نهایی بلوک مولتی‌رز[۸]

۷ –۱ – ساختار شبکه

همانطور که در مقدمه این بخش گفته شد، به جای اعمال فیلتر با سایزهای مختلف به صورت موازی، می‌توان از ساختار متناوب مانند شکل ۱۶–b استفاده کرد. اگرچه این روش باعث کاهش میزان استفاده از حافظه می‌شود، اما همچنان حافظه زیادی مصرف می‌کند؛ زیرا در دو لایه کانولوشنی متناوب، مقدار حافظه مصرفی با توان دوم تعداد فیلترهای لایه اول متناسب است؛ بنابراین ابتهاج و رحمان[۸] برای لایه اول کانولوشنی کمترین تعداد فیلتر را درنظر گرفتند و سپس به ترتیب تعداد فیلترها را برای لایه‌های بعدی افزایش دادند تا از این طریق مصرف حافظه بهبود یابد. علاوه بر این، آن‌ها از یک اتصال باقی‌مانده‌ای مشابه بلاک‌های باقی‌مانده‌ای[۱۰]، که روی آن یک لایه کانولوشنی 1x1 اعمال می‌شود استفاده کردند. طرح نهایی بلاک مولتی‌رز، که به جای گام‌های عادی در شبکه یو–نت استفاده خواهد شد، در شکل ۱۶ –c ارائه شده‌است.

همانطور که در شبکه یو-نت++ بررسی شد، وجود تفاوت معنایی بین کانال‌های ویژگی که از طریق اتصالات پرشی با هم ترکیب می‌شوند، عملکرد یو-نت را تضعیف خواهد کرد. به طور مشابه، ابتهاج و رحمان[۸] در شبکه مولتی‌رز یو-نت با استفاده از اتصالات پرشی ارتقا یافته، سعی در بهبود عملکرد یو-نت دارند[۴۱]. به طور دقیق‌تر آن‌ها در هر اتصال پرشی مجموعه‌ای از لایه‌های کانولوشنی با اتصالات باقی‌مانده‌ای قرار دادند(شکل ۱۷) و آن را مسیر رز[۴۲] نامیدند. تعداد لایه‌های کانوشونی در هر مسیر رز بستگی به سطح آن در شبکه دارد. برای بالاترین مسیر رز از ۴ لایه کانولوشنی و برای سطوح پایین‌تر به ترتیب ۳، ۲ و ۱ لایه کانولوشنی استفاده شده‌است.

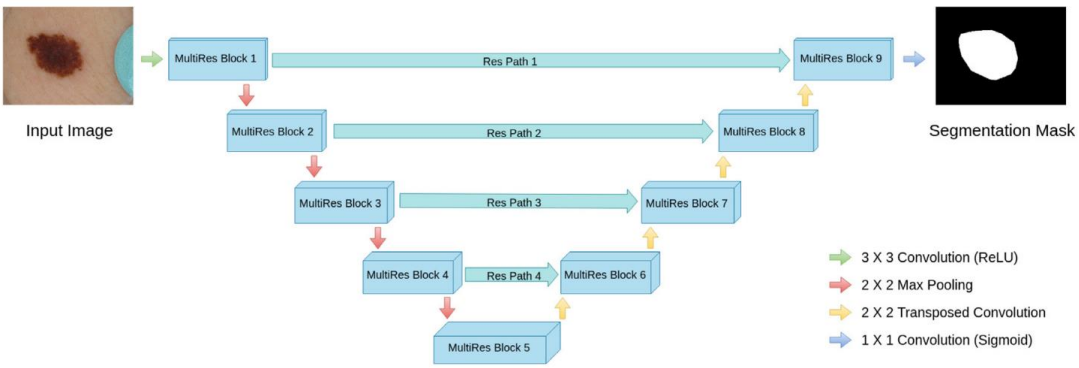


شکل ۱۷- یک نمونه از مسیر رز برای پر کردن فاصله معنایی بین کدگذار و کدگشا[۸]

در شبکه مولتی‌رز یو-نت، دو لایه کانوشونی در هرگام با بلاک‌های مولتی‌رز و اتصالات پرشی با مسیر رز جایگزین شده‌اند(شکل ۱۸). تعداد فیلترها در هر بلاک مولتی‌رز، توسط پارامتر W کنترل می‌شود. جهت مقایسه با شبکه یو-نت، مقدار W را به صورت ضریبی از تعداد فیلترهای شبکه یو-نت در گام متناظر با آن(U) به صورت زیر محاسبه می‌کنیم.

$$W = \alpha \times U \qquad (6)$$

ابتهاج و رحمان[۸] از مقدار ۱.۶۷ برای ضریب α استفاده می‌کنند. همانطور که پیش از این اشاره شد، در هر بلاک که دارای سه لایه کانولوشنی است، جهت کنترل حافظه مصرفی، تعداد فیلترها به تدریج افزایش می‌یابد. بنابراین برای لایه کانولوشنی اول تا سوم به ترتیب $[\frac{W}{6}]$، $[\frac{W}{3}]$ و $[\frac{W}{2}]$ فیلتر درنظر گرفته می‌شود.



شکل ۱۸-ساختار شبکه مولتی‌رز یو-نت

[۴۱] اگرچه ایده ارتقا اتصالات پرشی اولین بار در شبکه یو-نت++ مطرح شد، اما نویسندگان مقاله مولتی‌رز یو-نت هیچ اشاره‌ای به آن نکردند.

[۴۲] Res path

۷ -۲ - آموزش شبکه و نتایج تجربی

برای آموزش شبکه از تابع انرژی آنتروپی متقابل دودویی و روش آدام استفاده شده‌است. جهت مقایسه بهتر عملکرد شبکه پیشنهادی با شبکه یو-نت استاندارد، از افزونگی داده‌ها استفاده نمی‌شود. شبکه مولتی‌رز یو-نت با شبکه یو-نت در چندین مجموعه دادگان مختلف دو بعدی و سه بعدی مقایسه شده‌اند. در شکل ۲۰ نتایج ارزیابی این دو شبکه برای پنج مجموعه دادگان ارائه شده‌است. همانطور که مشاهده می‌شود شبکه مولتی‌رز نت در همه موارد با اختلاف قابل قبولی از شبکه یو-نت عملکرد بهتری داشته است؛ در حالی که تعداد پارامترهای آن‌ها نسبتا یکسان است(شکل ۱۹).

2D		3D	
Model	Parameters	Model	Parameters
U-Net (baseline)	7,759,521	3D U-Net (baseline)	19,078,593
MultiResUNet (proposed)	7,262,750	MultiResUNet 3D (proposed)	18,657,689

شکل ۱۹- تعداد پارامترهای مدل‌های یو-نت و مولتی‌رز یو-نت[۸]

Modality	MultiResUNet (%)	U-Net (%)	Relative improvement (%)
Dermoscopy	80.2988 ± 0.3717	76.4277 ± 4.5183	5.065
Endoscopy	82.0574 ± 1.5953	74.4984 ± 1.4704	10.1465
Fluorescence microscopy	91.6537 ± 0.9563	89.3027 ± 2.1950	2.6326
Electron microscopy	87.9477 ± 0.7741	87.4092 ± 0.7071	0.6161
MRI	78.1936 ± 0.7868	77.1061 ± 0.7768	1.4104

شکل ۲۰-مقایسه شبکه یو-نت و مولتی‌رز یو-نت در مجموعه دادگان‌های مختلف[۸]

۸- جمع‌بندی

در این گزارش به صورت خلاصه به بررسی شبکه یو-نت و چندین نسخه بهبود یافته آن پرداخته‌شد. شبکه یو-نت جز اولین شبکه‌های تمام کانولوشنی بود که برای مسئله قطعه‌بندی همه‌نما در کاربرد پزشکی معرفی شد. این شبکه با ساختار متقارن کدگذار-کدگشا و اتصالات پرشی توانست دو ایراد اصلی شبکه‌های تمام کانولوشنی، نیاز به داده آموزشی زیاد و مصالحه بین اطلاعات معنایی و فضایی را تا حد زیادی برطرف کند. شبکه یو-نت پس از معرفی در بسیاری از کاربردها مورد استفاده قرار گرفت، اما بخش زیادی از پیشرفت‌ها و محبوبیت خود را مدیون محققین حوزه پزشکی است. سال‌ها پس از معرفی شبکه یو-نت، تعداد بسیار زیادی شبکه براساس آن توسعه یافته که بررسی همه آن‌ها در این مقال نمی‌گنجد. اگرچه هم‌اکنون در مجموعه دادگان‌های عمومی، شبکه یو-نت جایی در میان برترین‌ها ندارد[۲, ۱۸]، اما نسخه‌های جدیدتر آن همچنان در حوزه پزشکی درمیان برترین مدل‌ها هستند[۱۹]. یکی از دلایل این موضوع این است که معمولا مجموعه دادگان‌های پزشکی داده‌های بسیار کمتری را شامل می‌شوند و یکی از ویژگی‌های شبکه‌های مبتنی بر یو-نت، توانایی یادگیری با داده‌های بسیار کم است.

٩- مراجع

[١] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440 .

[٢] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2021.

[٣] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, (Lecture Notes in Computer Science, 2015, ch. Chapter 28, pp. 234-241.

[٤] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, 2016: Springer, pp. 424-432 .

[٥] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters,* vol. 15, no. 5, pp. 749-753, 2018.

[٦] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*: Springer, 2018, pp. 3-11.

[٧] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis,* vol. 53, pp. 197-207, 2019.

[٨] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks,* vol. 121, pp. 74-87, 2020.

[٩] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826 .

[١٠] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778 .

[١١] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *European Conference on Computer Vision*, 2010: Springer, pp. 210-223 .

[١٢] C.-Y. Lee ,S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*, 2015: PMLR, pp. 562-570 .

[١٣] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708 .

[١٤] J. Sun, J. Jiang, and Y. Liu, "An Introductory Survey on Attention Mechanisms in Computer Vision Problems," in *2020 6th International Conference on Big Data and Information Analytics (BigDIA)*, 2020: IEEE, pp. 295-300 .

[١٥] M. Ren and R. S. Zemel, "End-to-end instance segmentation with recurrent attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6656-6664 .

[١٦] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[١٧] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9 .

[١٨] P. w. Code. "Semantic Segmentation." https://paperswithcode.com/task/semantic-segmentation/latest (accessed.

[١٩] H. Huang *et al.*, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 1055-1059 .