Receptive Field Encoding Model for Dynamic Natural Vision

Fatemeh Kamali[1], Amir Abolfazl Suratgar[1*], Mohamad Bagher Menhaj[1], Reza Abbasi Asl[2]

[1]Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran
[2]Department of Neurology, University of California, San Francisco, CA, USA

Article Info:

Received: **13 Feb 2019** ***Revised:*** **5 Jun 2019** ***Accepted*****: 18 Aug 2019**

ABSTRACT

Introduction: Encoding models are used to predict human brain activity in response to sensory stimuli. The purpose of these models is to explain how sensory information represent in the brain. Convolutional neural networks trained by images are capable of encoding magnetic resonance imaging data of humans viewing natural images. Considering the hemodynamic response function, these networks are capable of estimating the blood oxygen level dependence of subject viewing videos without any recurrence or feedback mechanism. For this purpose, feature map extracted from the convolutional neural network and the concept of receptive field has been used for the encoding model. The main assumption of this model is that activity in each voxel encodes a spatially localized region across multiple feature maps and for each voxel and this area are fixed for all feature maps. Contribution of each feature map in the activity of each voxel is determined by the corresponding weight. ***Materials and Methods:*** In this study, three healthy volunteers watching a set of videos. This collection contains images that represent real-life visual experience. MRI and fMRI data are acquired on a 3 tesla MRI system phase-array surface coil. ***Results:*** Data revealed that human visual cortex has hierarchical structure. Earlier visual areas have a smaller receptive field size in and response to simple feature like edge, whereas higher visual areas have a larger receptive field size and response to more complex features, such as pattern. ***Conclusion:*** This model of video stimuli has a higher interpretation capacity than the previous models.

Key words:

1. Magnetic Resonance Imaging
2. Visual Cortex
3. Brain

***Corresponding Author:** Amir Abolfazl Suratgar

E-mail: *a-suratgar@aut.ac.ir*

[DOI: 10.29252/shefa.7.4.1]

مدل رمزگذاری میدان تأثیر برای بینایی طبیعی دینامیکی

فاطمه کمالی[۱]، امیرابوالفضل صورتگر[۱*]، محمدباقر منهاج[۱]، رضا عباسی اصل[۲]

[۱]دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر، تهران، ایران

[۲]دانشکده نورولوژی، دانشگاه کالیفرنیا، سانفرانسیسکو، ایالات متحده آمریکا

اطلاعات مقاله:

دریافت: ۲۴ بهمن ۱۳۹۷ **اصلاحیه:** ۱۵ خرداد ۱۳۹۸ **پذیرش:** ۲۷ مرداد ۱۳۹۸

چـــــکیده

مقدمه: مدل‌های رمزگذاری برای پیش‌بینی فعالیت مغز انسان در پاسخ به محرک‌های حسی مورد استفاده قرار می‌گیرند. هدف این مدل‌ها توضیح دادن نحوهٔ ارائهٔ اطلاعات حسی در مغز است. شبکه‌های عصبی کانولوشنی که به وسیلهٔ تصاویر آموزش دیده‌اند قادر به رمزگشایی داده‌های تصویربرداری رزونانس مغناطیس عملکردی از انسان‌ها در حال مشاهدهٔ تصاویر طبیعی هستند. با در نظر گرفتن تابع پاسخ همودینامیک، این شبکه‌ها بدون داشتن هیچ مکانیسم بازگشتی یا پسخور قادر به تخمین میزان اکسیژن خون وابسته به سطح برای تصاویر ویدیویی نیز هستند. برای این منظور از نقشه‌های ویژگی استخراج شده از شبکهٔ عصبی کانولوشن و مفهوم میدان تأثیر در مدل رمزگذاری استفاده شده است. فرض اصلی در این مدل این است که برای هر واکسل یک منطقهٔ مکانی در نقشه ویژگی کدگذاری می‌شود و این مناطق برای همهٔ نقشه‌های ویژگی ثابت است. سهم هر نقشه ویژگی در فعالیت واکسل از طریق وزن مربوطه مشخص می‌شود. ***مواد و روش‌ها:*** در این پژوهش سه داوطلب سالم در حال تماشای مجموعه‌ای از تصاویر ویدیویی هستند. این مجموعه حاوی تصاویری است که نمایانگر بینایی طبیعی در زندگی واقعی است. داده‌های ام آر آی و اف ام آر آی با استفاده از کویل‌های سطحی ارایه فازی ۳ تسلا سیستم ام آر آی گرفته شده است. ***یافته‌ها:*** داده‌ها نشان داد که قشر بینایی انسان دارای ساختاری سلسله مراتبی است. نواحی دیداری اولیه دارای میدان تأثیر کوچک‌تری هستند و به ویژگی‌های ساده مثل لبه پاسخ می‌دهند، در حالی که نواحی دیداری سطح بالاتر دارای میدان تأثیر بزرگ‌تری بوده و به ویژگی‌های پیچیده‌تر مانند الگو پاسخ می‌دهند. ***نتیجه‌گیری:*** این مدل برای تصاویر ویدیویی ظرفیت تفسیرپذیری بالاتری را نسبت به مدل‌های پیشین دارد.

کلید واژه‌ها:

۱. تصویربـرداری رزونانس مغنا طیسی
۲. قشر بینایی
۳. مغز

* **نویسنده مسئول:** امیرابوالفضل صورتگر

آدرس الکترونیکی: *a-suratgar@aut.ac.ir*

[DOI: 10.29252/shefa.7.4.1]

مقدمه

برای سال‌ها دانشمندان درصدد درک نحوهٔ فعالیت مغز انسان بوده‌اند. مدل‌های رمزگذاری[1] پاسخ مغز به محرک‌های خارجی را تخمین می‌زنند. در سال‌های اخیر شبکه‌های عصبی مصنوعی برای انجام وظایف بینایی آموزش دیده‌اند. نتایج نشان می‌دهد که میزان دقت این شبکه‌ها برای انجام وظایفی نظیر طبقه‌بندی تصاویر یا برچسب زدن به تصاویر تا حد بسیار خوبی قابل قبول می‌باشد. در کارهای گذشته نشان داده شده است که پردازش‌های انجام شده توسط این مدل‌ها مشابه سلسله مراتب بینایی انسان می‌باشد (۱). به همین دلیل بازنمودهای داخلی که توسط این شبکه‌ها ارائه می‌شود می‌تواند ویژگی‌های بصری کدگذاری شده توسط مغز را فراهم کند. پس از استخراج ویژگی توسط این شبکه‌ها به هر نقشه ویژگی یک وزن اختصاص می‌یابد. ویژگی‌های مهم به طور معمول دارای وزن‌های بزرگ هستند در حالی که ویژگی‌های با اهمیت کمتر وزن کمتری دارند. وزن برای ویژگی‌های بصری، از یک مجموعه دادهٔ آموزشی با استفاده از الگوریتم بهینه‌سازی مناسب به دست می‌آید. پس از تعیین وزن‌های مدل، توانایی مدل در تخمین فعالیت مغز توسط یک مجموعه تست ارزیابی می‌شود. این مجموعه متفاوت با مجموعهٔ آموزشی می‌باشد.

در روش جمعیت میدان تأثیر[2] (۲) ویژگی بصری، یک نقشه باینری از پیکسل‌هایی است که توسط یک محرک با کنتراست بالا مثل حلقه یا نوار به دست می‌آید. سپس برای هر واکسل، این نقشهٔ ویژگی با یک منطقهٔ مکانی ادغام می‌شود به نحوی که بهترین پاسخ را برای واکسل ارائه کند. همچنین فیلترهای ویولت گابور[3] (۳) مدل انرژی حرکتی[4] (۴)، مدل‌های معناگرا[5] (۵) و مدل‌هایی که از دیگر روش‌های یادگیری ماشین استفاده می‌کنند (۶) برای استخراج ویژگی استفاده شده است. پژوهش‌های اخیر نشان می‌دهد که یادگیری عمیق، جامع‌ترین مدل محاسباتی را برای رمزگذاری و استخراج ویژگی‌های سازماندهی شدهٔ سلسله مراتبی از تصاویر را فراهم می‌کند (۷). از شبکه‌های کانولوشنی برای استخراج ویژگی‌های سطح پایین و سطح بالا برای مسیر ونترال[6] (۸، ۹) و درسال[7] (۱۰) مسیر بینایی انسان استفاده شده است. مدل میدان تأثیر با ویژگی وزن داده شده می‌تواند مکان و شعاع میدان تلفیق ویژگی را برای تصاویر بیابد (۱۱). در این مقاله نشان داده می‌شود که این شبکه‌ها می‌توانند داده‌های تصویربرداری رزونانس مغناطیسی عملکردی[8] ناشی از تماشای فیلم‌های طبیعی را، حتی با عدم وجود دینامیک زمانی یا فیدبک مدل کنند.

مواد و روش‌ها

مجموعه داده‌ها

سه داوطلب سالم در این مطالعه شرکت کرده‌اند. هر داوطلب مجموعه‌ای از تصاویر طبیعی را در قالب ویدیو تماشا کرده است. در کل، دادهٔ آموزشی شامل ۳۷۴ کلیپ با نرخ ۳۰ فریم در ثانیه در مدت زمان ۲ ساعت و ۲۴ دقیقه است که به ۱۸ قسمت ۸ دقیقه‌ای تقسیم شده است. دادهٔ آزمون شامل ۵۹۸ کلیپ متفاوت در مدت زمان ۴۰ دقیقه است که به ۵ قسمت ۸ دقیقه‌ای تقسیم شده است. داده‌های آزمون از داده‌های آموزشی کاملاً متفاوت بوده و کلیپ‌ها از اینترنت (www.youTube.com, www.videoblocks.com) گرفته شده که نمایانگر تجربیات بینایی طبیعی انسان است. به‌عنوان مثال این کلیپ‌ها شامل انسان‌ها، حرکت حیوانات و مناظر طبیعی است. هر داوطلب دادهٔ آموزشی را ۲ مرتبه و دادهٔ آزمون را ۱۰ مرتبه و در روزهای مختلف مشاهده نموده است. هر دور آزمایش شامل چند بخش ۸ دقیقه و ۲۴ ثانیه‌ای است. در هر بخش قبل از شروع پخش فیلم فریم اول به صورت تصویر ثابت به مدت ۱۲ ثانیه نشان داده می‌شود و در انتهای بخش نیز فریم آخر به همین صورت به مدت ۱۲ ثانیه نشان داده می‌شود. برای آموزش و آزمون مدل رمزگذاری ۲ تکرار دادهٔ آموزشی و ۱۰ تکرار دادهٔ آزمون میانگین گرفته می‌شود (۱۲).

داده‌های T1 و T2 وزن‌دار شدهٔ ام آر آی و اف آم آر آی در یک سیستم ام آر آی ۳ تسلا گرفته شده است. برای گرفتن این تصاویر از کویل‌های سطحی، آرایهٔ فازی ۱۶ کاناله استفاده شده است. داده‌های اف ام آر آی با رزولوشن فضایی ۳/۵ میلی‌متر و رزولوشن زمانی ۲ ثانیه گرفته شده است. داده‌های اف ام آر آی پردازش شده روی

تصویر ۱– همبستگی متقابل بین سیگنال‌های اف ام آر آی به دست آمده در طی دو مرحله تکرار فیلم آموزشی (۱۲).

[1] Encoding model
[2] Receptive field
[3] Gabor wavelet
[4] Motion energy model
[5] Semantic model
[6] Ventral
[7] Dorsal
[8] Functional magnetic resonance imaging

[DOI: 10.29252/shefa.7.4.1]

سطح قشر فردی بر اساس الگوی چگالی میلین ثبت گردیده‌اند. پیش پردازش و ثبت با دقت بالا توسط پروژهٔ اتصالات مغزی انسانی[9] انجام شده است.

مروری برشبکه‌های کانولوشنی

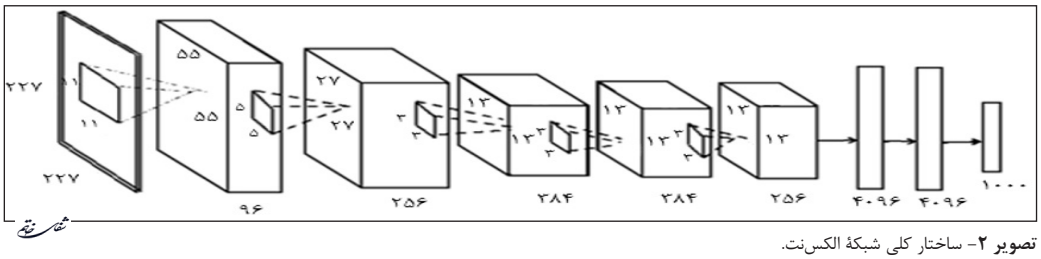
ما از شبکهٔ کانولوشنی از پیش آموزش دیده موسوم به الکس‌نت[10] برای استخراج ویژگی تصاویر ورودی، جهت تخمین سری زمانی ویژگی استفاده کردیم (۱۳). آموزش این مدل بر روی داده‌های چالش تشخیص بصری مقیاس بزرگ[11] صورت گرفته است (۱۴). این شبکه شامل ۸ لایه است که ۵ لایهٔ اول کانولوشنی و ۳ لایهٔ آخر متصل کامل هستند و برای کلاس‌بندی کردن اجسام آموزش دیده است. سایز تصویر ورودی این شبکه ۲۲۷ در ۲۲۷ می‌باشد. تصویر ورودی به لایهٔ اول داده شده؛ خروجی لایهٔ اول به لایهٔ دوم داده شده و به همین ترتیب ادامه می‌یابد. هر لایهٔ کانولوشن شامل تعدادی فیلتر است که تابع فعالسازی آن‌ها یکسوساز خطی واحد می‌باشد. خروجی لایهٔ اول نقشهٔ ویژگی با سایز ۵۵ در ۵۵ در ۹۶ و به ترتیب در لایه‌های بعد خروجی دوم برابر با ۲۵۶ نقشهٔ ویژگی ۲۷ در ۲۷، در لایهٔ سوم ۳۸۴ نقشهٔ ویژگی ۱۳ در ۱۳، در لایهٔ چهارم ۳۸۴ نقشهٔ ویژگی ۱۳ در ۱۳ و در لایهٔ پنجم ۲۵۶ نقشهٔ ویژگی ۱۳ در ۱۳ است. پس از لایه‌های کانولوشن سه لایهٔ متصل کامل با سایز ۴۰۹۶، ۴۰۹۶ و ۱۰۰۰ قرار دارد.

مروری بر تابع پاسخ همودینامیک

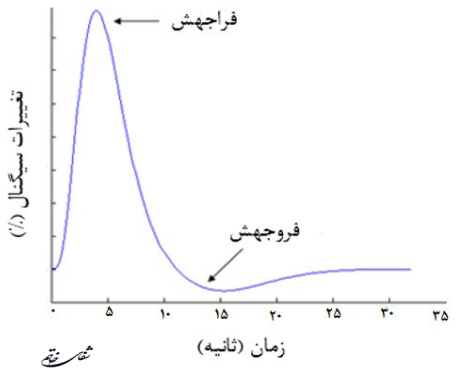
الگوی تغییرات ایجاد شده در میزان اکسیژن خون و متعاقب آن شدت سیگنال تشدید مغناطیسی در پاسخ به یک محرک را پاسخ همودینامیک[12] می‌نامند. پاسخ همودینامیک را می‌توان به صورت پاسخ ضربهٔ یک سیستم خطی در نظر گرفت که الگوی زمانی فعالیت عصبی مغز را به الگوی زمانی تغییر در شدت سیگنال تشدید مغناطیسی تبدیل می‌کند. یک پاسخ معمول میزان اکسیژن خون وابسته به سطح[13] به محرک در نمودار ۱ نشان داده شده است. پاسخ تقریباً حدود ۵ ثانیه بعد از تحریک به حداکثر مقدار خود می‌رسد و پس از آن با یک فروجهش تقریباً ۳۰ ثانیه‌ای همراه است (۱۵). در مطالعات قبلی معمولاً از زمان‌های طولانی بین محرک‌ها استفاده می‌کردند که اجازه دهند پاسخ به حالت معمول بازگردد. با این وجود، اگر فاصلهٔ بین وقوع محرک‌ها کم باشد پاسخ‌ها با هم همپوشانی خواهند داشت. این همپوشانی می‌تواند از طریق کانوالو کردن تابع پاسخ همودینامیکی مدل شود. نمودار ۱ پاسخ ضربهٔ میزان اکسیژن خون وابسته به سطح را نشان می‌دهد.

مدل رمزگذاری میدان تأثیر

میدان تأثیر یک نورون، یک ناحیهٔ خاص از فضای حسی است که حضور محرک در آن باعث شلیک سلسلهٔ اسپایک در آن نورون خواهد شد. مدل میدان تأثیر از سه قسمت اصلی شامل مجموعه‌ای از نقشه‌های ویژگی، یک بردار وزن و میدان ادغام ویژگی درست شده است. فرضیهٔ اصلی در این مدل این است که در فعالیت‌های اندازه‌گیری شده برای هر واکسل، هرچه پیکسل‌های نقشهٔ ویژگی به مرکز ادغام نزدیک‌تر باشند میزان تأثیر آن‌ها در فعالیت مغز بیش‌تر خواهد بود. همچنین فرض می‌شود برای یک واکسل مرکز و شعاع میدان ادغام برای همهٔ نقشه‌های ویژگی ثابت باقی می‌ماند.



تصویر ۲- ساختار کلی شبکهٔ الکس‌نت.



نمودار ۱- پاسخ ضربهٔ میزان اکسیژن خون وابسته به سطح کانونی.

[9] Human connectome project
[10] Alex net
[11] Large scale visual recognition challenge
[12] Hemodynamic response function
[13] Blood oxygen level dependence

[DOI: 10.29252/shefa.7.4.1]

بنابراین میدان ادغام ویژگی را می‌توان مانند یک پنجره بر روی نقشهٔ ویژگی در نظر گرفت که از یک نقشه به نقشهٔ بعد تغییر نمی‌کند. در این روش میدان ادغام ویژگی،گوسی دو بعدی به صورت زیر در نظر گرفته می‌شود:

$$g(x,y;\mu_x,\mu_y,\sigma_g)=\frac{1}{\sqrt{2\pi\sigma_g}}\exp[-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma_g{}^2}]$$

که در فرمول فوق، (μ_x,μ_y) مرکز میدان ادغام و σ_g شعاع میدان ادغام می‌باشد. برای تخمین پاسخ واکسل به تصویر محرک S_t، ابتدا نقشهٔ ویژگی $\phi^k(s_t)$ محاسبه کرده و پاسخ حاصل را با پاسخ هودینامیک کانوالو کرده تا تأخیر زمانی حاصل شود، سپس میدان ادغام ویژگی روی هر نقشه اعمال می‌شود، در مرحلهٔ آخر بردار وزن به این خروجی‌ها اعمال می‌شود. در واقع داریم:

$$\hat{r}_t=\sum_k^K w_k\int_{-D/2}^{D/2}\int_{-D/2}^{D/2}g(x,y;\mu_x,\mu_y,\sigma_g)(h(t)*\phi^k_{i(x)j(y)}(s_t))dxdy$$

که در این فرمول، $\hat{r}_t$ پاسخ تخمین زده شده در پاسخ به محرک s_t است. زمانی که مدل را برای یک واکسل خاص بهینه می‌کنیم در واقع به دنبال تعیین ضرایب μ_x,μ_y,σ_g و ضرایب وزن $w=(w_1,...,w_k)$ برای بهترین تخمین در پاسخ به هر فریم ویدیو دلخواه هستیم.

پارامترهای بهینهٔ مدل از طریق حداقل کردن مربعات خطا برای هر واکسل به صورت زیر به دست می‌آید:

$$E(\theta)=\sum_t(r_t-\hat{r}_t(\theta))^2$$

در فرمول بالا، r_t پاسخ اندازه‌گیری شده به تصویر S_t است و $\hat{r}_t$ پاسخ تخمین زده شده توسط مدل می‌باشد. برای تخمین وزن‌ها ابتدا، مجموعه‌ای از سایز و مکان میدان ادغام ویژگی را در نظر می‌گیریم. ساختار کلی این روش در تصویر ۳ نشان داده شده است.

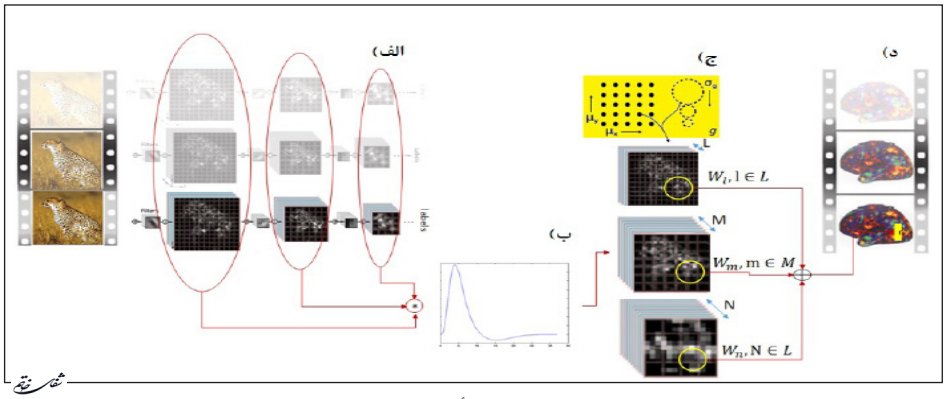
برای هر یک از این نامزدها در مجموعهٔ در نظر گرفته شده، گرادیان کاهشی را برای به دست آوردن وزن‌های بهینه اعمال می‌کنیم. برای این کار گرادیان $L(\theta)$ متناظر با w را برای ۸۰ درصد دادهٔ آموزشی اعمال خواهد شد. گرادیان کاهشی برای هر یک از نامزدها برای ۵۰ تکرار انجام خواهد شد و پس از اتمام آن میزان دقت بر روی ۲۰ درصد دادهٔ باقیمانده بررسی خواهد شد. پس از بهینه کردن وزن‌ها، نامزدی که منجر به رسیدن به کمترین میزان خطا شود به‌عنوان پارامتر بهینه برای μ_x,μ_y,σ_g انتخاب می‌شود. سپس مدل با پارامترهای بهینهٔ انتخاب شده برای ۱۰۰ درصد دادهٔ آموزشی مجدداً آموزش دیده که پارامترهای w بهینه انتخاب شوند. از w انتخاب شده برای تخمین بر روی مجموعهٔ آزمون استفاده می‌کنیم.

یافته‌ها

مدل میدان تأثیر وزن‌دار شده بر روی مجموعهٔ داده که به صورت عمومی منتشر شده ارزیابی شده است. مجموعهٔ داده شامل میزان اکسیژن خون وابسته به سطح در پاسخ به عکس‌های طبیعی برای واکسل‌های نواحی V1، V2، V3A،V3d ، V3B، V4، LO[14]، MT[15] از قشر بینایی مغز می‌باشد.

دقت تخمین پاسخ از طریق ضریب همبستگی بین پاسخ مشاهده شده و پاسخ تخمین زده شده بر روی مجموعهٔ آزمون اندازه‌گیری می‌شود. برای مجموعه‌ای از واکسل‌ها، میانگین میزان دقت برای بیان دقت تخمین استفاده می‌شود. در نمودار ۲ میزان دقت تخمین برای نواحی مختلف نشان داده شده است.

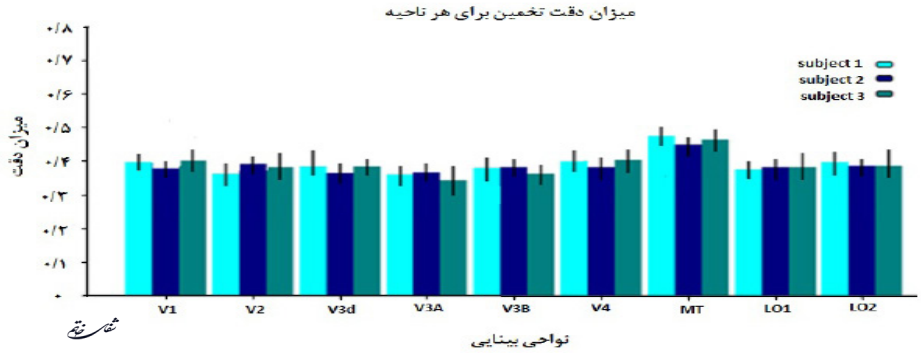
می‌دانیم که هرچه اهمیت یک نقشهٔ ویژگی بیش‌تر



تصویر ۳- ساختار کلی مدل میدان رمزگذاری با مفهوم میدان تأثیر. الف) نقشه‌های ویژگی با استفاده از شبکهٔ کانولوشنی برای هر فریم ویدیو استخراج می‌گردد. ب) نقشه‌های استخراج شده با تابع هودینامیک کانوالو می‌شود. ج) یک مجموعه از سایز و مکان میدان ادغام در نظر گرفته می‌شود. د) خروجی ناشی از اعمال میدان ادغام ویژگی بر روی نقشه‌های ویژگی را وزن‌دار کرده که این وزن‌ها از طریق گرادیان کاهشی بهینه می‌گردند.

[14] Lateral occipital
[15] Middle temporal

[DOI: 10.29252/shefa.7.4.1]



نمودار ۲- میزان دقت مدل در نواحی مختلف برای سه شخص مورد آزمایش. برای هر ناحیه میزان دقت به صورت (میانگین ± واریانس) برای تمام واکسل‌های ناحیهٔ مورد نظر نشان داده شده است.

باشد وزن بزرگتری به آن اختصاص می‌یابد اما در عین حال اندازهٔ وزن‌ها به ضرایب نقشهٔ ویژگی نیز وابسته است. بنابراین برای تعیین میزان مشارکت هر لایه در تخمین پاسخ بجای اندازهٔ وزن‌ها از معیار زیر استفاده می‌کنیم.

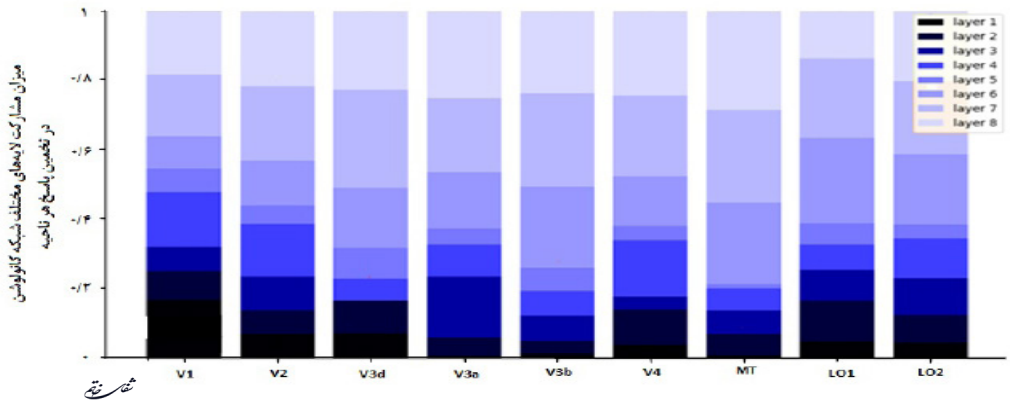
$$\rho_1 = \frac{cov(\tilde{r}_l, r)_t}{\sqrt{var(\tilde{r})_t var(r)_t}}$$

که در فرمول فوق مشابه اما برای نقشه‌های ویژگی لایهٔ ام محاسبه می‌شود. این میزان مشارکت برای نواحی مختلف در نمودار ۳ نشان داده شده است.

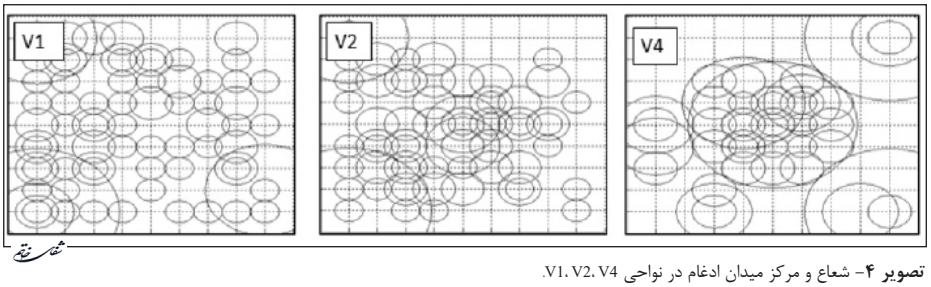
همانطور که در نمودار ۳ دیده می‌شود در نواحی سطح پایین میزان تأثیر لایه‌های اولیهٔ کانولوشن بیش‌تر است و هرچه در سلسله مراتب بینایی به سمت بالاتر می‌رویم میزان تأثیر لایه‌های پایین کمتر شده و تأثیر لایه‌های بالاتر بیش‌تر می‌گردد. در شبکه‌های کانولوشنی لایه‌های پایین‌تر به ویژگی‌های سطح پایین مانند لبه و کنتراست پاسخ می‌دهند در حالی که لایه‌های بالاتر به ویژگی‌های پیچیده‌تر مانند بافت‌های نامنظم، قسمتی از شی یا کل شی پاسخ می‌دهند. بنابراین می‌توان نتیجه گرفت که در سلسله مراتب بینایی نواحی اولیه به ویژگی‌های سطح پایین پاسخ داده و هرچه در سلسله مراتب بینایی به سمت بالاتر می‌رویم این ویژگی‌ها پیچیده‌تر می‌گردد.

بررسی شعاع و مرکز میدان‌های ادغام در نواحی مختلف می‌تواند اطلاعاتی در مورد سایز میدان تأثیر در نواحی مختلف ارائه کند. مرکز و شعاع میدان ادغام برای ۱۰۰ واکسل در نواحی V1، V2، V4 که دارای بیش‌ترین میزان دقت است در تصویر ۴ نشان داده شده است.

همانطور که دیده می‌شود هرچه در سلسله مراتب بینایی به سمت بالا حرکت می‌کنیم سایز میدان تأثیر بزرگتر شده و مراکز متمرکزتر می‌گردند.



نمودار ۳- میزان مشارکت لایه‌های مختلف شبکهٔ کانولوشن در تخمین پاسخ نواحی مختلف بینایی.



تصویر ۴- شعاع و مرکز میدان ادغام در نواحی V1, V2, V4.

[DOI: 10.29252/shefa.7.4.1]

بحث و نتیجه‌گیری

ما از مدل میدان تأثیر برای ساخت مدل رمزگذاری برای تخمین میزان اکسیژن خون وابسته به سطح برای تصاویر ویدیویی طبیعی استفاده کردیم. همانطور که دیده شد این مدل بدون داشتن هر نوع ساختاری برای استخراج ویژگی‌های زمانی برای تصاویر ویدیویی به خوبی پاسخ می‌دهد.

استفاده از روش میدان تأثیر در مقایسه با روش‌هایی که به هر پیکسل نقشهٔ ویژگی وزن می‌دهند باعث ایجاد مدلی ساده‌تر و قابل تفسیرتر می‌شود. در واقع در روش‌های پیشین برای ساخت مدل رمزگذاری برای تصاویر ویدیویی به هر پیکسل در نقشهٔ ویژگی یک وزن اختصاص داده می‌شد که به محاسبات بسیار می‌انجامید. هرچند استفاده از روش‌های کاهش ابعاد مانند تحلیل مؤلفه‌های اساسی تا حدی تعداد پارامترها را کم می‌کند اما تعداد پارامترها همچنان بسیار زیاد است. تعداد پارامترها در لایهٔ اول تا هشتم به ترتیب برابر است با ۲۹۰۴۰۰، ۱۸۶۶۲۴، ۶۴۸۹۶، ۶۴۸۹۶، ۴۳۲۶۴، ۴۰۹۶، ۴۰۹۶ و ۱۰۰۰ است. در صورت استفاده از تحلیل مؤلفه‌های اساسی تعداد رگرسورها تا ۴۰ برابر کاهش می‌یابد. اما در صورت استفاده از مدل میدان تأثیر تعداد پارامترهای رگرسور در هر لایه به ترتیب برابر خواهد شد با ۹۶، ۲۵۶، ۳۸۴، ۳۸۴، ۲۵۶، ۴۰۹۶، ۴۰۹۶ و ۱۰۰۰. همانطور که دیده می‌شود در این روش تعدا پارامترها بسیار کمتر شده و تفسیر ساده‌تر می‌گردد.

بررسی نتایج مدل میدان تأثیر نشان می‌دهد که هرچه در سلسله مراتب بینایی به سمت بالا حرکت می‌کنیم میدان تأثیر بزرگ‌تر شده و متراکم‌تر می‌گردد و به ویژگی‌هایی پیچیده‌تر پاسخ می‌دهد. در واقع می‌توان نتیجه گرفت که هرچه در سلسله مراتب بینایی به سمت بالا حرکت می‌کنیم واکسل‌ها از انتخاب ویژگی در محیط حسی به سمت انتخاب شی پیش می‌روند.

منابع

1. Güçlü U, van Gerven MA. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci. 2015; 35(27): 10005-14.

2. Dumoulin SO, Wandell BA. Population receptive field estimates in human visual cortex. Neuroimage. 2008; 39(2): 647-60.

3. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452(7185): 352-5.

4. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol. 2011; 21(19): 1641-6.

5. Naselaris T, Stansbury DE, Gallant JL. Cortical representation of animate and inanimate objects in complex natural scenes. J Physiol Paris. 2012; 106(5-6): 239-49.

6. Stansbury DE, Naselaris T, Gallant JL. Natural scene statistics account for the representation of scene categories in human visual cortex. Neuron. 2013; 79(5): 1025-34.

7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436.

8. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. arXiv preprint arXiv:1601.02970. 2016.

9. Agrawal P, Stansbury D, Malik J, Gallant JL. Pixels to voxels: modeling visual representation in the human brain. arXiv preprint arXiv:1407.5104. 2014.

10. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 2014; 10(11): e1003915. doi: 10.1371/journal.pcbi.1003915.

11. St-Yves G, Naselaris T. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. NeuroImage. 2018; 180: 188-202.

12. Wen H, Shi J, Zhang Y, Lu KH, Cao J, Liu Z. Neural encoding and decoding with deep learning for dynamic natural vision. Cerebral Cortex. 2017; 28(12): 4136-60.

13. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012; 25(2): 1-9.

14. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision. 2015; 115(3): 211-52.

15. Henson R, Friston K. Convolution models for fMRI. statistical parametric mapping: The analysis of functional brain images. 2007: 178-92.

[DOI: 10.29252/shefa.7.4.1]