

به نام خدا



دانشگاه صنعتی امیرکبیر

Amirkabir University
of Technology

پروژه دوم درس یادگیری عمیق

آشنایی با شبکه های عصبی به عنوان تخمین گرهای عمومی، بهینه سازهای درجه
یک و درجه دو و مقدمات تعمیم پذیری در شبکه های عصبی

استاد درس: دکتر صفابخش

نام: زهرا اخلاقی

شماره دانشجویی: ۴۰۱۱۳۱۰۶۴

فهرست مطالب

2.....	بخش اول (شبکه‌های عصبی به عنوان تخمین‌گرهای عمومی)
2.....	(۱)
4.....	(۲)
5.....	(۳)
7.....	بخش دوم (مقایسه بهینه سازهای درجه اول و درجه دوم)
7.....	(۴)
8.....	(۵)
8.....	(۶)
11.....	بخش سوم (تعمیم پذیری شبکه‌های عصبی)
11.....	(۷)
12.....	(۸)
13.....	(۹)

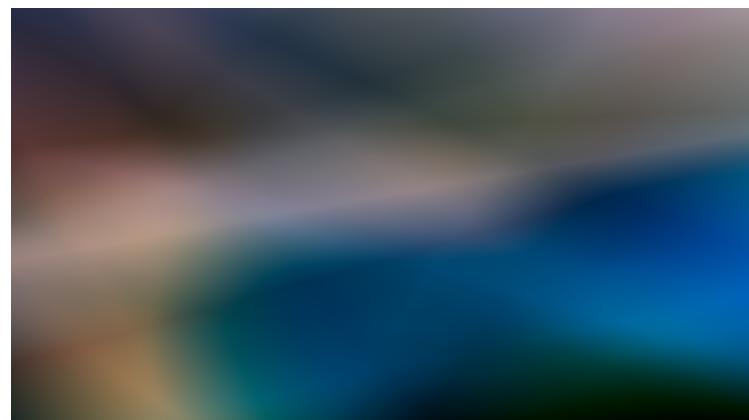
بخش اول(شبکه های عصبی به عنوان تخمین گرهای عمومی)

در نتایج ارائه شده 20 درصد دادهها برای تست و ۸۰ درصد برای آموزش استفاده شده اند و تابع بهینه سازی SGD است. کلاس ImagePixelDataset برای تغییر مقیاس طول و عرض تصویر به فاصله ۱-۰ استفاده شده و کلاس های multi layer perceptron MLP , MLP3,MLP4 برای پیاده سازی با تعداد لایه های مختلف استفاده شده است (در تمام این شبکه ها تابع فعالسازی RELU است).

(۱)

بررسی تعداد نورون و لایه در آموزش شبکه:

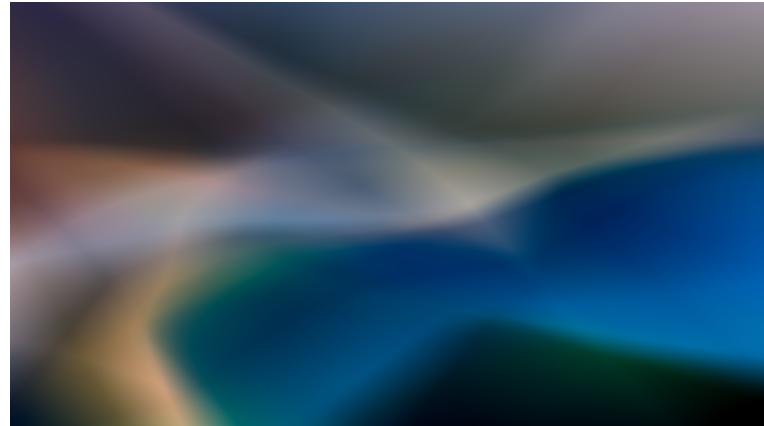
- شبکه با دو لایه و لایه پنهان با ۲۵۶ نورون، اندازه دسته: ۱۶



- شبکه با سه لایه و لایه پنهان با ۱۲۸ و ۲۵۶ نورون، اندازه دسته: ۱۶



- شبکه با چهار لایه و لایه پنهان با ۶۴ و ۱۲۸ و ۲۵۶ نورون، اندازه دسته: ۱۶



Test Loss: 0.0724

- شبکه با سه لایه و لایه پنهان با ۸۴ نورون، اندازه دسته ۱۶:



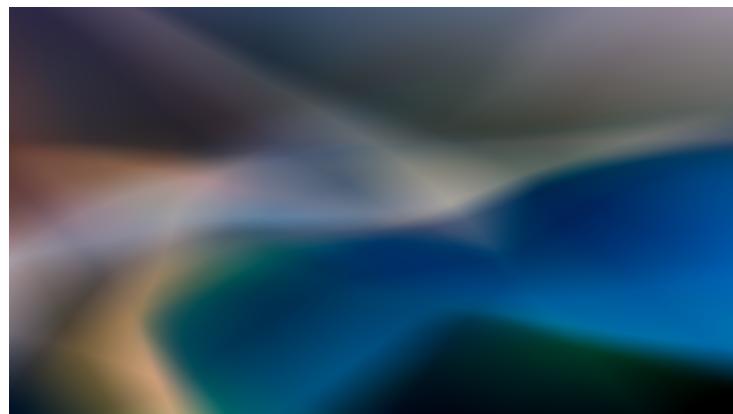
تعداد نورون‌ها و تعداد لایه‌ها تاثیر زیادی بر آموزش و عملکرد شبکه دارند، این تأثیرات به مسئله‌ی مورد نظر، حجم داده بستگی دارد. تعداد نورون‌ها و لایه‌ها به توانایی مدل در یادگیری و نمایش الگوهای پیچیده‌تر اطلاعات ارتباطی اجزا موجود در داده‌ها مرتبط است. شبکه‌های بزرگ‌تر ممکن است توانایی یادگیری اطلاعات پیچیده‌تری را داشته باشند ولی شبکه‌های بزرگ پیچیده‌تر هستند و زمان بیشتری برای آموزش آن‌ها مورد نیاز است و مشکل بیش برآذش در آن‌ها وجود دارد.

تعداد لایه‌ها و نورون‌ها همچنین بر انتخاب توابع فعال‌سازی تاثیر دارد. مثلاً برخی توابع فعال‌سازی ممکن است برای مسائل خاص مناسب‌تر باشند. انتخاب تعداد نورون‌ها و لایه‌ها باید با توجه به خصوصیات مسئله و محدودیت‌های موجود صورت گیرد. از آزمایش‌های مختلف و روش‌های ارزیابی مدل می‌توان برای تنظیم بهینه تعداد نورون‌ها و لایه‌ها استفاده کرد.

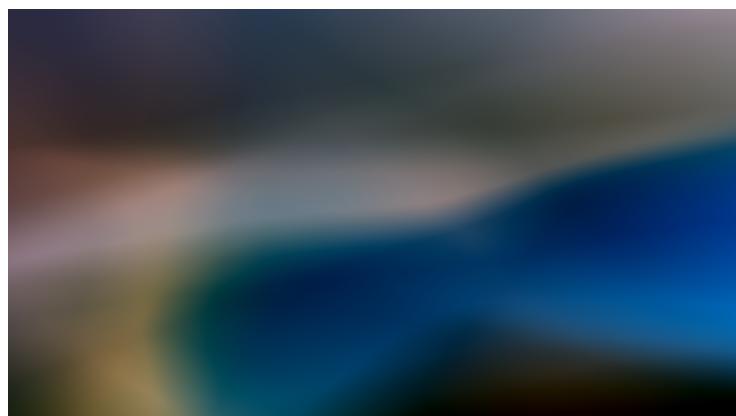
(۲)

این قسمت بر روی شبکه با ۴ لایه با اندازه ۲۵۶ و ۱۲۸ و ۶۴ بررسی شده

اندازه دسته: ۱۶ ●



اندازه دسته: ۶۴ ●



اندازه دسته: ۱۰۲۴ ●



اندازه دسته تعیین می‌کند چند داده به هر بار آموزش شبکه اعمال شود و وزنهای شبکه بعد از اعمال هر دسته به روز می‌شوند. اندازه دسته تاثیر زیادی بر سرعت آموزش دارد. دسته بزرگتر باعث افزایش سرعت آموزش می‌شوند ولی اندازه دسته بیش‌آموزی را کاهش داده و تخمین گرادیان‌ها را بهبود می‌بخشد به کاهش احتمال وقوع تغییرات بزرگ در تخمین گرادیان‌ها کمک می‌کند.

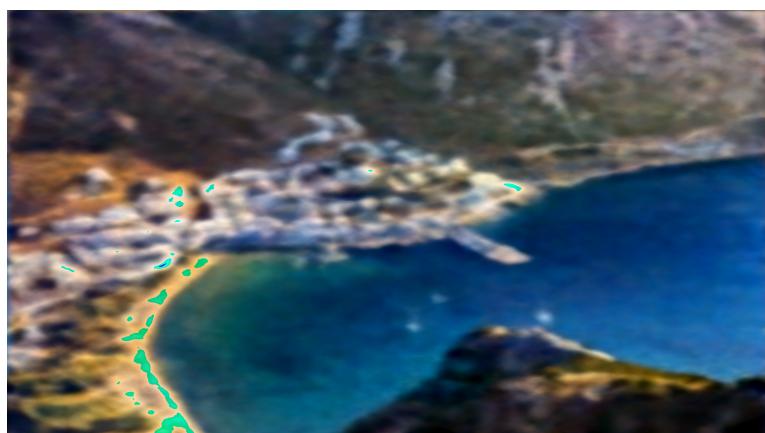
استفاده از اندازه دسته کوچکتر می‌تواند به انحراف استاندارد گرادیان کمک کند. این کاهش انحراف استاندارد می‌تواند باعث سرعت همگرایی بهتر و پایداری آموزش شود. در دسته‌های بزرگتر، تخمین‌های گرادیان به معنای آماری بهتری از گرادیان کل داده‌ها خواهند بود و باعث شود که بهینه‌سازی به سمتی هدایت شود که ممکن است در مسائل کلی تر منجر به بایاس کمتری شود.

همانطور که در شکل‌های بالا مشخص اس دقت تخمین در اندازه تصویر کوچکتر بهتر است ولی سرعت آموزش با افزایش اندازه تصویر بیشتر می‌شود.

(۳)

گزارش این قسمت بر روی شبکه ۴ لایه با اندازه ۳ لایه مخفی ۱۲۸، ۲۵۶ و ۶۴ می‌باشد.

L=4



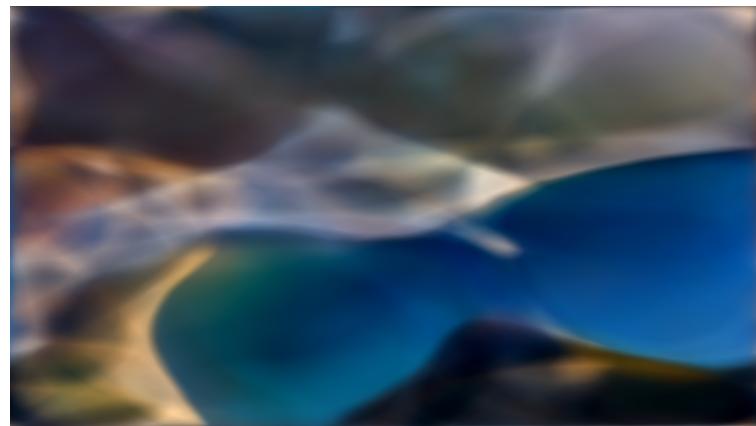
Test Loss: 0.0203

L=2



Test Loss: 0.0269

L=1



Test Loss: 0.0292

افزایش مقدار L باعث بهتر شدن دقت شبکه میشود و شبکه بهتر می تواند تصویر را آموزش بیند و تصویر خروجی را تخمین بزند ولی افزایش L باعث کاهش سرعت آموزش شبکه میشود و در آموزش مدلها در محیط کولب افزایش بیش از حد اندازه L مشکل کش حافظه وجود دارد.
کمترین مقدار تست شده برای L برابر با ۱ است ولی اگر نیاز به مشاهده دقت بیشتر و جزئیات بیشتری داریم باید مقدار L بزرگتر انتخاب سود.

بخش دوم (مقایسه بهینه سازهای درجه اول و درجه دوم)

(۴)

بهینه سازی درجه دوم به معنای استفاده از مشتق دوم در فرآیند بهینه سازی است. در حالت کلی، بهینه سازی درجه دوم دارای مزایایی زیر در بهینه سازی شبکه است:

- سرعت همگرایی بیشتر: استفاده از مشتق دوم می‌تواند باعث تسریع فرآیند همگرایی بهینه سازی شود. این به معنای کاهش تعداد مراحل یا تعداد دفعاتی است که به نقطه بهینه می‌رسد.
- استفاده بهتر از اطلاعات محلی: مشتق دوم می‌تواند اطلاعات بیشتری از خطا و تغییرات در محلی نقطه فعلی فضا را ارائه دهد. این اطلاعات به بهینه سازی کمک می‌کند تا بهبودهای دقیق‌تری در هر مرحله انجام شود.
- پایداری بالاتر: استفاده از مشتق دوم پایداری بیشتری نسبت به نویزهای محلی دارد. اما با وجود این مزایا، استفاده از بهینه سازی درجه دوم در شبکه‌های عصبی ممکن است با چندین چالش مواجه شود:
- محاسبات پیچیده‌تر: محاسبه مشتق دوم ممکن است محاسبات پیچیده‌تری نسبت به مشتق اول را ایجاد کند و این امر ممکن است باعث افزایش هزینه محاسباتی شود.

- بهینه سازی سخت تر است و زمان بیشتری برای بهینه سازی مورد نیاز است و این مدل ها بسیار پیچیده تر هستند به صورتی که توسعه سخت تری دارند، در حالی که استفاده از بهینه سازی درجه اول راحت تر است
- برای محاسبه مشتق دوم، نیاز به داشتن اطلاعات بیشتری از داده ها و شبکه می باشد که ممکن است در مواردی که داده ها محدود هستند، یک چالش باشد.
- استفاده از حافظه بیشتر برای انجام محاسبات پیچیده و هزینه سربار استفاده از حافظه استفاده از بهینه سازی درجه دوم در شبکه های عصبی، به دلیل چالش های محاسباتی و پیچیدگی مسائل، معمولاً محدود تر است و بسیاری از الگوریتم های بهینه سازی امروزی بر اساس مشتق اول عمل می کنند.

(۵)

- الگوریتم بهینه سازی L-BFGS یک الگوریتم بهینه سازی کمینه یابی غیرخطی است که براساس روش توسعه یافته است. L در نام L-BFGS به معنای "محدود" است، زیرا الگوریتم تعداد محدودی از آخرین تغییرات پارامترها در فرآیند بهینه سازی را نگهداری می کند.
- ویژگی ها و مزایای L-BFGS عبارتند از:

- فضای حافظه محدود: L-BFGS فضای حافظه ای محدود را برای ذخیره اطلاعات گذشته از تغییرات پارامترها استفاده می کند.
- با استفاده از ترکیبی از مشتق اول و یک ماتریس دودویی تقریبی (که از آن به نام Hessian استفاده می شود)، L-BFGS اطلاعات بهینه سازی را با حفظ کیفیت، به صورت فشرده نگه می دارد. این ویژگی کاربردی است زیرا در شبکه های عصبی عمیق با تعداد پارامترهای بالا، محاسبات Hessian کلان و پرهزینه ممکن است غیرممکن یا زمان بر باشد.

- سرعت همگرایی: L-BFGS به عنوان یک الگوریتم بهینه سازی کمینه یابی دسته ای (batch) باعث سرعت همگرایی بالا در مقایسه با الگوریتم های تکاملی می شود، به ویژه در مسائل بهینه سازی پیچیده.
- این الگوریتم این امکان را می دهد که بدون نیاز به محاسبه دقیق مشتقات دوم، بهینه سازی را انجام دهد. با توجه به این ویژگی ها، L-BFGS به عنوان یک الگوریتم بهینه سازی فشرده با حافظه محدود و قابل استفاده در مسائل بهینه سازی پیچیده، در بهبود عملکرد بهینه سازی در شبکه های عصبی عمیق مؤثر است.

(۶)

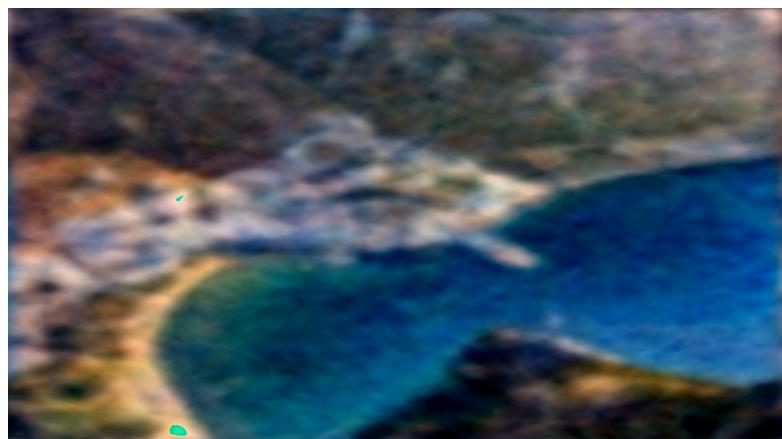
الگوریتم‌های بهینه‌سازی SGD و L-BFGS دو الگوریتم معروف در زمینه بهینه‌سازی هستند و از هر کدام در شرایط خاصی استفاده می‌شود. اما این دو الگوریتم تفاوت‌های مهمی در عملکرد و کاربرد دارند.

الگوریتم SGD معمولاً به تعداد تکرار کمتری نسبت به L-BFGS نیاز دارد. این به دلیل این است که در هر مرحله، SGD از یک نمونه تصادفی از داده‌ها برای بهروزرسانی وزن‌ها استفاده می‌کند، به همین دلیل زمانی که داده‌ها بزرگ و حافظه محدود است می‌تواند مفید باشد. L-BFGS ممکن است به تعداد تکرار بیشتری نیاز داشته باشد تا به همگایی برسد. این الگوریتم از حافظه محدود برای ذخیره اطلاعات پیشین و بهروزرسانی وزن‌ها استفاده می‌کند.

SGD سرعت همگایی بیشتری نسبت به L-BFGS دارد، اما این تفاوت به میزان تعداد تکرارها و سایر پارامترهای تنظیم شده برای هر الگوریتم بستگی دارد. با توجه به محاسبات پیچیده‌تر مشتق دوم در L-BFGS مخصوصاً در مسائل با ابعاد بالا ممکن است زمان بیشتری برای همگایی بهینه‌سازی نیاز باشد.

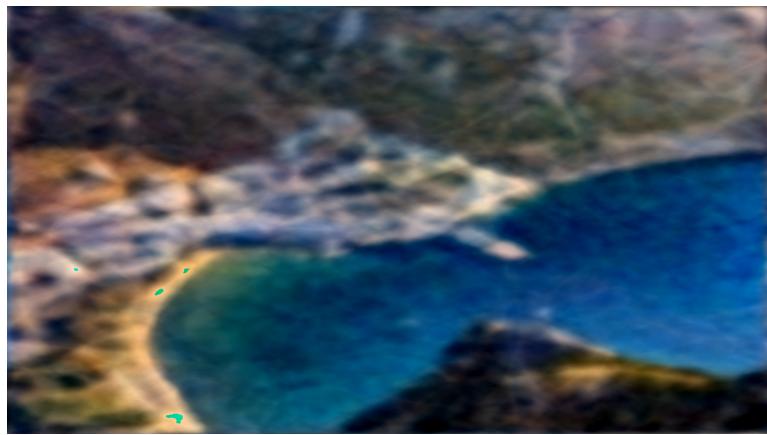
در شکل‌های زیر $L=3$ (از شبکه با ابعاد بالاتر سوال ۳ در این قسمت استفاده شده است) در نظر گرفته شده است. ارزیابی نامناسبی از تغییرات در اندازه تصاویر مواجه شود.

در شبکه با دو لایه و لایه پنهان ۲۵۶ L-BFGS



Test Loss: 0.0263 Duration: 1378.901659488678 seconds

L-BFGS در شبکه با سه لایه و لایه پنهان ۱۲۸ و ۲۵۶:



Duration: 2948.0328373908997 seconds Test Loss: 0.0252

در شبکه با چهار لایه و لایه پنهان ۲۵۶ و ۱۲۸ و ۶۴ L-BFGS:



Duration: 3322.069444656372 seconds Test Loss: 0.0249

در استفاده از الگوریتم L-BFGS پارامترها نقش زیادی در بهینه سازی مدل دارند و تعیین نامناسب پارامترها باعث میشود مدل به درستی آموزش نبیند و الگوریتم L-BFGS دقیق بالاتری نسبت به SGD دارد (test_loss کوچکتر) و زمان بیشتری برای آموزش آن لازم است.

بخش سوم (تعمیم پذیری شبکه‌های عصبی)

تعمیم‌پذیری در زمینه شبکه‌های عصبی عمیق به این معنا است که مدل برای داده‌هایی که در آموزش شرکت نکرده است نیز عملکرد قابل قبولی ارائه می‌دهد. به عبارت دیگر، تعمیم‌پذیری به امکان استفاده از مدل بر روی داده‌های جدید و متفاوت از داده‌های آموزش شده اشاره دارد.

(۷)

در یادگیری ماشین، دو مجموعه ارزیابی و تست به منظور ارزیابی عملکرد مدل‌ها استفاده می‌شوند. این دو مجموعه از اهمیت بسیاری برخوردارند و در مراحل مختلف توسعه مدل‌ها به کار می‌روند.

مجموعه ارزیابی (Validation Set):

مجموعه ارزیابی در مرحله آموزش مدل برای تنظیم پارامترهای مدل به کار می‌رود. در هر مرحله از آموزش، مدل با استفاده از مجموعه ارزیابی، تنظیمات مختلفی از پارامترهای خود را آزمایش کرده و از طریق ارزیابی عملکرد در این مجموعه، تصمیم می‌گیرد که کدام پارامترها بهترین عملکرد را ارائه می‌دهند و هدف اصلی از مجموعه ارزیابی، انتخاب بهترین مدل و بهترین تنظیمات پارامتری در مراحل آموزش مدل است.

مجموعه تست (Test Set):

مجموعه تست، عملکرد نهایی مدل را بر روی داده‌های جدید و ناشناخته ارزیابی کند. این مجموعه به عنوان یک بررسی نهایی در مرحله نهایی یا حتی بعد از آموزش، برای اطمینان از قابلیت تعمیم مدل به داده‌های جدید به کار می‌رود و هدف اصلی از مجموعه تست، ارزیابی نهایی و عملکرد مدل در مقابل داده‌های جدید و ناشناخته است. این مجموعه به ما اطلاع می‌دهد که مدل آموزش یافته قابلیت تعمیم به داده‌های خارج از مجموعه آموزش را دارد یا خیر.

مجموعه تست از داده‌های تشکیل شده است که مدل هنگام آموزش ندیده است. این اطلاعات جدید به ما کمک می‌کنند تا از عملکرد مدل در مقابل داده‌های جدید واقعیت به دست آوریم و از ویژگی‌های تعمیم‌پذیر مدل اطمینان حاصل کنیم.

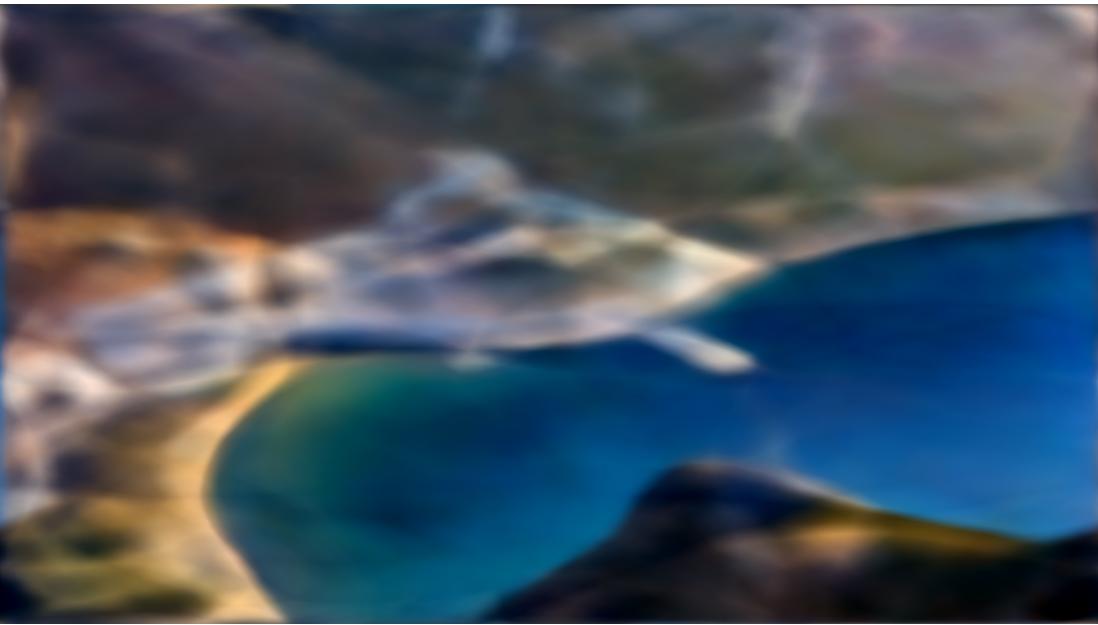
[OBJ]

(۸)

تصویر با رزولوشن دو برابر:



تصویر حاصل از آموزش شبکه با $L=2$ و شبکه ۴ لایه با اندازه لایه‌های پنهان ۲۵۶ و ۱۲۸ و ۶۴ نورون (در استفاده از L با اندازه بیشتر مشکل crash حافظه وجود داشت).



استفاده از تصاویر با رزولوشن بالاتر نسبت به تصاویری که شبکه در مرحله آموزش با آنها روبرو شده است، یک مثال مناسب برای ارزیابی تعمیم‌پذیری شبکه است و به ما این امکان را می‌دهد که ببینیم آیا شبکه قادر به تعمیم‌پذیری اطلاعات و ویژگی‌ها به داده‌های با رزولوشن بالاتر نیز است یا خیر.

تصاویر با رزولوشن بالاتر اطلاعات جزئی‌تری از صحنه را ارائه می‌دهند. این اطلاعات جزئی‌تر می‌توانند در ارزیابی شبکه برای تشخیص و تعیین ویژگی‌های کمک کنند. این روش به ما اجازه می‌دهد تا به دقت بسنجیم که مدل در مقابل داده‌های جدید با ویژگی‌های رزولوشن بالاتر چه عملکردی از خود نشان می‌دهد.

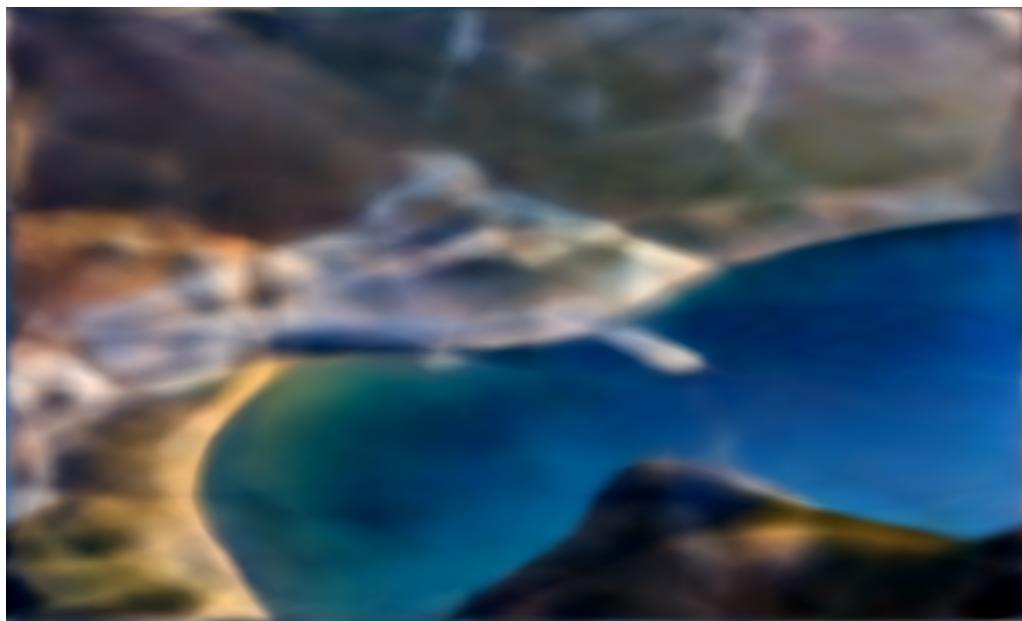
ارزیابی مدل با استفاده از داده‌های تست با رزولوشن بالاتر یکی از راه‌هایی است که به ما کمک می‌کند اطمینان حاصل کنیم که مدل به درستی یاد گرفته و به ویژگی‌ها و الگوهای جدیدی که در داده‌های با رزولوشن بالاتر وجود دارند، تعمیم پذیر است و دو برابر کردن رزولوشن تصویر با توجه به اینکه ورودی‌های شبکه مختصات تصویر هستند و مقیاس شبکه تغییری نکرده (کشیده تر یا پهن تر) می‌تواند مثال خوبی برای تست تعمیم پذیری شبکه باشد.

(۹)

افزودن ۲۰۰ پیکسل به اطراف تصویر:



تصویر حاصل از آموزش شبکه با $L=2$ و شبکه ۴ لایه با اندازه لایه‌های پنهان ۲۵۶ و ۱۲۸ و ۶۴ نورون (در استفاده از L با اندازه بیشتر مشکل crash حافظه وجود داشت)



افزودن حاشیه ۲۰۰ پیکسل به اطراف تصویر نمی‌تواند به صورت عمومی به عنوان یک مورد ارزیابی قابل اعتماد مورد استفاده قرار گیرد.

- اطلاعات مصنوعی: حاشیه افزودنی به تصویر اطلاعات مصنوعی افزوده می‌شود که در داده‌های واقعی و ناشناخته وجود ندارد بنابراین نمونه مناسبی برای تست ارزیابی شبکه نیست.
- تغییر در توزیع داده: افزودن حاشیه به اطراف تصویر باعث تغییر در توزیع داده شود. این تغییر ممکن است و با توجه به اینکه ورودی شبکه مختصات تصویر هستند افزودن ۲۰۰ پیکسل مثال نامناسبی از تعمیم پذیری شبکه باشد.
- اگر یک شبکه با تغییرات در اندازه تصاویر مواجه نشده باشد، افزودن حاشیه ممکن است به شبکه این انعطاف را ندهد. به عبارت دیگر، این کار مثال نامناسبی از تعمیم پذیری شبکه باشد.
بهتر است برای ارزیابی تعمیم‌پذیری، از داده‌های تستی که از پیش با توجه به توزیع داده‌های واقعی و مورد نظر انتخاب شده‌اند، استفاده شود. در صورتی که نیاز به ارزیابی تعمیم‌پذیری در شرایط خاص و یا با تغییرات مشخصی باشد، بهتر است از روش‌های ارزیابی معتبر و منطقی استفاده شود که با اهداف اصلی ارزیابی مدل هماهنگ باشند.