

یادگیری عمیق

پاییز ۱۴۰۱

استاد: دکتر فاطمیزاده

سید رکم سلی ۹۸۱۰۷۰۷۷

گردآورندها: سعید رضوی، هلیا حاج‌کاظمی، ارشیا همت



دانشگاه صنعتی شریف
دانشکده‌ی مهندسی برق

مهلت ارسال: جمعه ۱۷ آر

شبکه‌های عمیق کانولوشنی

تمرین سوم

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۶ روز و در مجموع ۲۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال شده پذیرفته نخواهد بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- هم‌کاری و همفکری شما در انجام تمرین مانع ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
(دقت کنید در صورت تشخیص مشابه غیر عادی برخورد جدی صورت خواهد گرفت.)
- در صورت همفکری و یا استفاده از هر منابع خارج درسی، نام همفکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- نتایج و پاسخ‌های خود را در یک فایل با فرمت zip به نام HW۳-Name-StudentNumber در سایت Quera قرار دهید. برای بخش عملی تمرین نیز لینک گیت‌هاب که تمرین و نتایج را در آن آپلود کرده‌اید قرار بدهید. دقต کنید هر سه فایل نوتبوک تکمیل شده بخش عملی را در گیت‌هاب قرار دهید.
- لطفا تمامی سوالات خود را از طریق کوئیرای درس مطرح بکنید (برای اینکه تمامی دانشجویان به پاسخ‌های مطرح شده به سوالات دسترسی داشته باشند و جلوی سوالات تکراری گرفته شود، به سوالات در بسترهای دیگر پاسخ داده نخواهد شد).
- دقت کنید کدهای شما باید قابلیت اجرای دوباره داشته باشند، در صورت دادن خطای هنگام اجرای کدن، حتی اگه خطای اشتباه تایپی باشد، نمره صفر به آن بخش تعلق خواهد گرفت.

سوالات نظری (۳۰۰ نمره)

۱. (۲۵ نمره)

یک مسئله دسته‌بندی با ۵ دسته، که دیتاست، شامل تصاویری به اندازه 10×10 پیکسل می‌باشد، داریم. دو شبکه عصبی یک لایه را به صورت زیر در نظر بگیرید. توضیح دهید کدام یک انتخاب بهتری می‌باشد؟

- یک لایه fully connected که ورودی آن، flatten (بردار شده) تصاویر دیتاست می‌باشد.
- یک لایه کانولوشن که در آن ۵ فیلتر به اندازه 10×10 داریم.

۲. (۲۵ نمره)

فرض کنید دیتاستی داریم شامل تصاویر رنگی به اندازه 128×128 . میخواهیم یک شبکه عصبی کانولوشن برای آن طراحی کنیم.

- اندازه خروجی و تعداد پارامترهای لایه اول کانولوشن را محاسبه کنید اگر ۱۶ فیلتر 5×5 با $1 = padding$ و $2 = stride$ داشته باشیم.

- فرض کنید هر لایه، ۳ قسمت شامل : کانولوشن، max pooling و تابع فعالسازی (Relu) را دارا می باشد، که لایه های کانولوشنی، هر کدام شامل ۱۶ فیلتر 5×5 با $stride = 1$ و $padding = 2$ و لایه های max pooling همگی 2×2 با 2 با $stride = 2$ هستند. ۳ لایه با این مشخصات را پشت سر هم در نظر بگیرید. اندازه تنسور در لایه خروجی نهایی و تعداد پارامترهای این ۳ لایه را حساب کنید.
- فرض کنید هدف، حل یک مسئله classification، که شامل ۱۰ دسته هست، می باشد. تعداد کل پارامترهای شبکه را در این حالت حساب کنید.
- در این قسمت، با یک مفهوم مهم آشنا می شویم. Receptive field بیانگر این است که نورون خروجی، تحت تاثیر چه مقدار از نورون های ورودی می باشد. در حقیقت تعیین می کند هر نورون خروجی از چه ناحیه ای با چه اندازه ای از تنسور ورودی تاثیر می پذیرد. حال Receptive field را برای یک نورون خروجی لایه سوم (قبل از لایه fully connected) بررسی کنید. (برای فهم بهتر این مفهوم میتوانید به این لینک مراجعه کنید)

۳. (۲۵ نمره)

در این تمرین قصد داریم به بررسی دو شبکه معروف یعنی Densely Connected Convolutional Networks و U-Net بپردازیم. برای بررسی هرچه بهتر این دو شبکه بهتر است به لینک های زیر مراجعه نمایید.

- <https://arxiv.org/pdf/1505.04597.pdf>
- <https://arxiv.org/pdf/1608.06993.pdf>

در این تمرین دو نوع سوال وجود دارد، یکی از این سوالات، سوالات مفهومیست که میزان تسلط شما بر روی شبکه های موجود را بررسی می کند و دوم سوالات محاسبه ای می باشد.

۱ - سوالات مفهومی مربوط به U-Net :

- ویژگی اصلی شبکه U-Net که آن را از یک شبکه کانولوشنی عادی متمایز می دارد چه می باشد و دلیل اینکه ما شاهد یک ساختار U شکل هستیم چه می باشد؟
- می دانیم که Skip connection ها نقشی پررنگ در این شبکه ها دارند، دلیل حضور این مورد را در شبکه U-Net بیان کنید.
- سوال اضافه (دارای نمره اضافه جزئی) : چرا این نوع از اتصالات در تصاویر پزشکی دارای اهمیت بیشتر می باشد و چه کمکی به ما در دامنه تشخیص موارد پزشکی می کنند؟

۲ - سوالات محاسبه ای مربوط به U-Net :

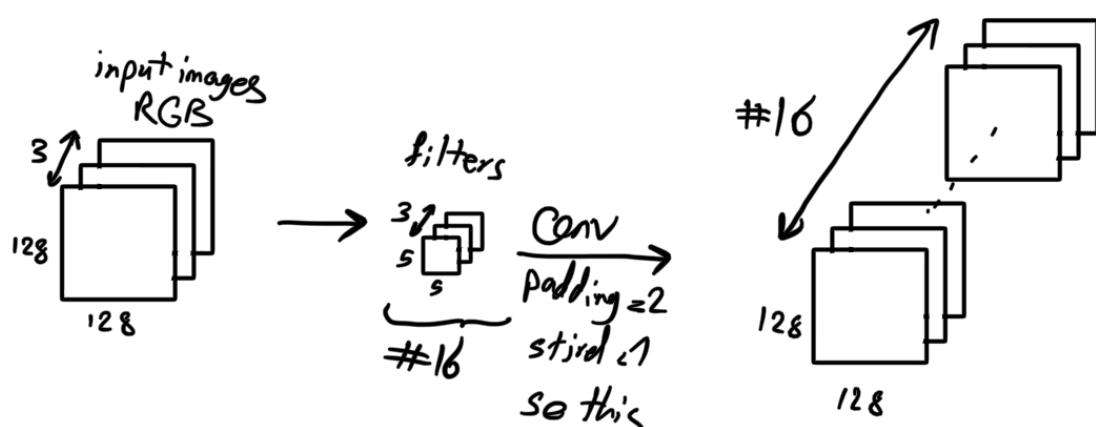
- تصویر کنید که ابعاد تصویر ورودی ما برای این شبکه 256×256 می باشد. حال فرض می شود که در این معماری هر لایه در انکدر ابعاد را به نصف کاهش می دهد و در دیکدر دو برابر می کند. در پایین ترین لایه (عمیق ترین لایه) این معماری، فضای ویژگی ما چند پیکسل خواهد داشت؟
- در U-Net، فرض کنید انکودر دارای لایه هایی با $64, 128, 256$ و 512 فیلتر است. اگر هر لایه کانولوشن از کرنال های 3×3 استفاده بکند، تعداد پارامترهای لایه کانولوشن دوم انکدر را محاسبه کنید.

۳ - سوالات مفهومی DenseNet :

- تفاوت های اصلی DenseNet's dense connections و ResNet's residual connections را بیان کنید. در مورد هر کدام از موارد گفته نیز، توضیح مختصری بدهید.
- بیان کنید که DenseNet چگونه مشکل vanishing gradient را کاهش می دهد و مزیت محاسباتی آن چه می باشد؟

Q2

Part 1 :



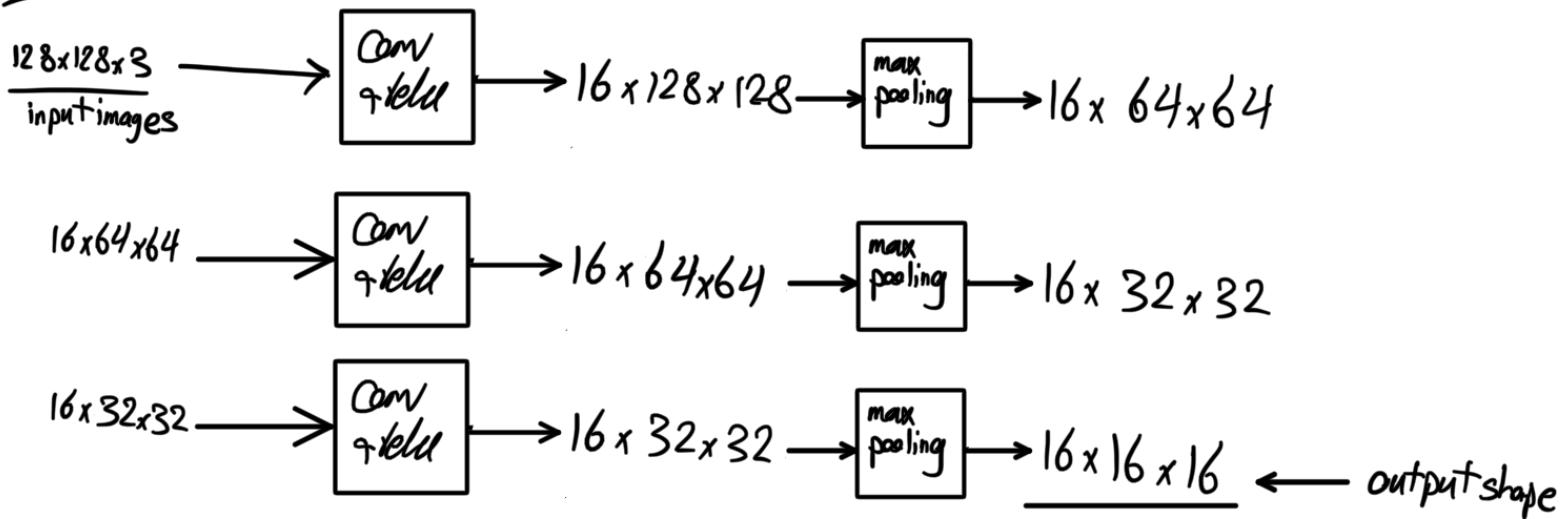
number of parameters: 1216

output shape: $16 \times 128 \times 128$

$$\frac{16}{\text{number of filters}} \times \frac{5}{w} \times \frac{5}{h} \times \frac{3}{d} + 16 = 1200 + 16 = \underline{1216}$$

weights biases

Q2 Part 2:



number of parameters: 14048

1st layer: $16 \times 5 \times 5 \times 3 + 16 = 1200 + 16 = 1216$

2nd layer: $16 \times 5 \times 5 \times 16 + 16 = 6400 + 16 = 6416$

3rd layer: $16 \times 5 \times 5 \times 16 + 16 = 6400 + 16 = 6416$ sum $\rightarrow 14048$

Q2 part 3

$16 \times 16 \times 16 \rightarrow$ flattening $\rightarrow 4096 \times 1 \xrightarrow{10 \text{ classes}} 10$ ← output
number of parameters:

$$\begin{array}{l} \text{FC} \\ \text{layer} \end{array} \rightarrow \underbrace{4096 \times 10}_{\text{weights}} + \underbrace{10}_{\text{biases}} = 40960 + 10 = 40970$$

Car backbone $\rightarrow 14048$

$$\Rightarrow \text{total parameters: } 14048 + 40970 = 55018$$

Q2 part 4

رده، maxpooling و conv می‌باشد
 $k=3, 4, 5$. که receptive field را کم کنند

پس از maxpooling Conv دارند که receptive field را کم کنند

$$R_{\text{conv}} = (R_{\text{previous}} - 1) \times \text{stride} + \text{filter size}$$

$$R_{\text{pool}} = R_{\text{previous}} \times \text{stride} \quad \leftarrow \text{پس از maxpooling}$$

$$R_{\text{conv}1} = (1 - 1) \times 1 + 5 = 5 \rightarrow R_{\text{pool}1} = 5 \times 2 = 10$$

$$R_{\text{conv}2} = (10 - 1) \times 1 + 5 = 14 \rightarrow R_{\text{pool}2} = 14 \times 2 = 28$$

$$R_{\text{conv}3} = (28 - 1) \times 1 + 5 = 32 \rightarrow R_{\text{pool}3} = 32 \times 2 = 64$$

بنابراین در خروجی receptive field نمودار داریم که همچنانکه اینها را کم کنند
maxpooling و conv می‌باشد

Q3

1:

Decoder (Contrating path) of Encoder (Expansive path)
و Decoder (Contracting path) of Encoder (Expansive path)

up convolution j'elbow Decoder و در این مرحله (32x32 pixels) (یعنی
پیکسل) and Segment هر کدام را در مکانی upSampling بزرگ کنیم.

۱۷

(c)

Q3

2: $256 \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \Rightarrow 16 \times 16 = 256$ pixels (a)

$$\frac{128}{\text{number of filters}} \times \frac{3 \times 3}{w/h} \times \frac{64}{\text{depth}} + \frac{128}{\text{biases}} = 73728 + 128 = \underline{\underline{73856}} \quad (\text{b})$$

Q3

ConCat photo = (photo1 + photo2 + photo3 + photo4) Desert, 2016

$$\text{Resnet: } x_\ell = H_\ell(x_{\ell-1}) + g_\ell$$

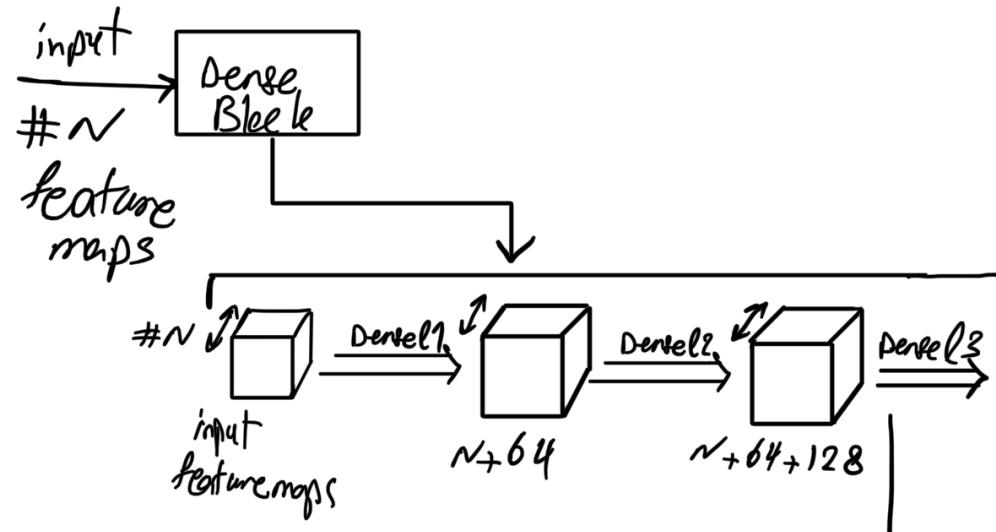
Dense net: $x_\ell = H_\ell([x_0, x_1, \dots, x_{\ell-1}])$

الآن نحن في المثلث (1 shell) concent for ملادي

جیھر لئے جو اسکے مطابق gradient ہو جائے تو اس کا خود ہے gradient vanishing point ہے۔

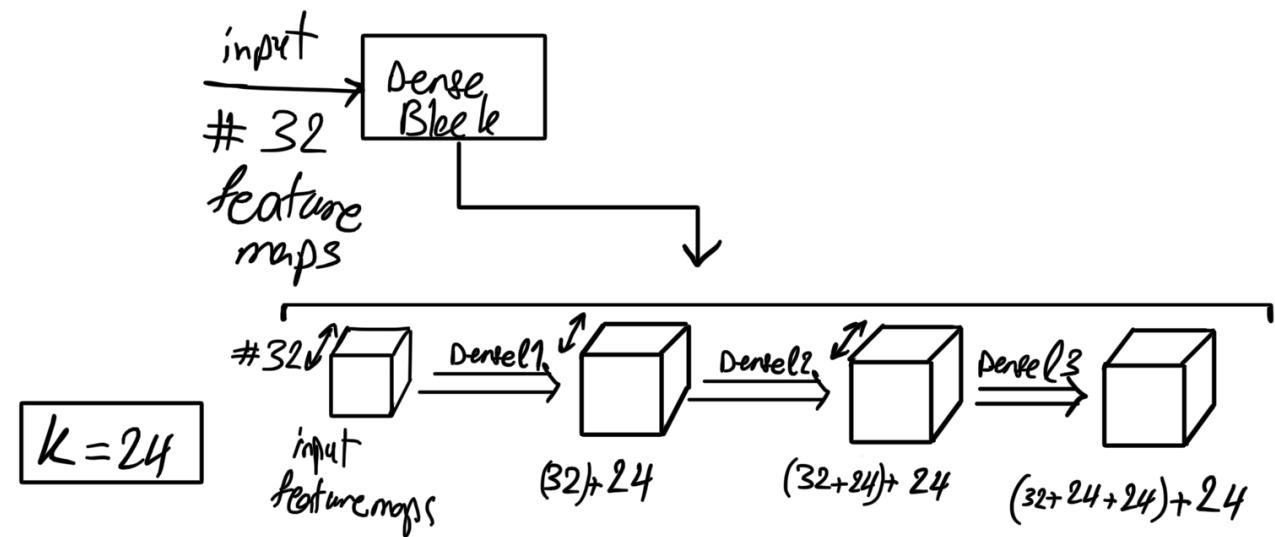
Q3

4.



$$\text{number of input feature maps of Dense layer 3} = N + 64 + 128 = N + 192$$

like part a we have :



$$\text{number of output channels at Dense layer 3} = 104$$

۴ - سوالات محاسبه‌ای DenseNet

- در یک DenseNet با سه لایه در یک Dense Block اگر لایه اول ۶۴ فیچر مپ تولید کند، لایه دوم ۱۲۸ فیچر مپ و لایه سوم ۲۵۶ فیچر مپ تولید کند، لایه سوم چند فیچر مپ ورودی را دریافت خواهد کرد؟
- با در نظر گرفتن نرخ رشد k در DenseNet، اگر هر لایه k فیچر مپ جدید تولید کند و ورودی یک dense block دارای ۳۲ کanal باشد، اگر $k = 24$ باشد لایه سوم در بلوک چند کanal خروجی خواهد داشت؟

سوالات عملی (۳۰۰ نمره)

۱. (۱۰۰ نمره) همانطور که می‌دانید، در بسیاری از مدل‌های شبکه عمیق، از یکتابع ضرر (Loss function) برای آموزش مدل استفاده می‌کنند. نکته‌ای که وجود دارد این است که لزوماً این تابع ضرر، مختص به لایه آخر شبکه عصبی نیست و می‌شود از آن در لایه‌های میانی نیز استفاده کنیم. همچنین، در بسیاری از موارد برای بهبود آموزش، می‌توانیم از جمع وزن دار چند تابع ضرر به صورت همزمان استفاده کنیم. در ادامه سوال، با این موارد بیشتر آشنا می‌شویم.

در این سوال، از معماری یکی از شبکه‌های معروف کانولوشنی (مانند Alexnet، Resnet50 و...) که بر روی imagenet ترین شده به دلخواه استفاده می‌کنیم. هدف ما آموزش یک classifier برای دو کلاس هواپیما (airplane) و ماشین (automobile) از دیتابست cifar10 است. همانطور که میدانیم، این نوع شبکه‌ها، شامل یک لایه fully connected هستند که ورودی آن برداری به اندازه feature vector است خروج شده و خروجی آن به به اندازه تعداد کلاس‌ها است. مثلاً برای Alexnet، خروجی لایه‌ی یکی مانده به آخر، یک بردار ۴۰۹۶ تایی است که از طریق یک لایه fully connected به یک بردار ۱۰۰۰ تایی برای مسئله دسته‌بندی مپ شده است. از آنجایی که تعریف‌های ما برای استخراج بردار ویژگی همیشه ممکن است دارای نقص‌هایی باشند و برخی موارد در نظر گرفته نشده باشند، چنانچه داده‌های بسیار زیاد و متنوعی به شبکه‌هایی مانند AlexNet نشان داده شود چه بسا بردار ویژگی که به دست می‌آورند بهتر از آنها می‌باشد که متخصصین تعریف می‌کنند. در مسائل مختلف دیده شده است که این بردار ویژگی‌ها حتی برای مسائلی که دسته‌بندی نیستند و یا دسته‌ها متفاوت از دسته‌های ImageNet هستند هم با معنی بوده و به نتایج خوب منجر می‌شوند. در ادامه شما ملزم به انجام موارد زیر هستید:

- یک شبکه کانولوشنی معروف به دلخواه انتخاب کنید، و با تغییر دادن لایه fully connected شبکه را برای دسته‌بندی یک مسئله دو کلاسه، آماده کنید. (دقت کنید این دو دسته، دو کلاس هواپیما (airplane) و ماشین (automobile) از دیتابست cifar10 هستند).

(آ) شبکه کانولوشنی خود را با استفاده از وزن‌های از قبل آموزش داده شده و cross entropy loss آموزش دهید (فقط وزن‌های مربوط به لایه fully connected را مقداردهی اولیه کنید). نمودار دقت و loss را به ازای epoch های مختلف رسم کنید. دقت کنید این دقت و loss بر روی دیتابست ترین باید صورت بگیرد.

(ب) در نهایت، دقت مدل ترین شده بالا را بر روی داده تست حساب کنید.