بسم خدا

پاسخ تمرینات تئوری سری سوم یادگیری عمیق

حسین آرمان‌مهر - 401706442

---

1) تعداد پارامترها: لایه FC با ورودی Flatten: $10 \times 10 \times 5 \rightarrow 100$

لایه Conv با فیلتر 10×10: $10 \times 10 \times 5$

با توجه به برابر بودن تعداد پارامترها بنظر می‌رسد هردو انتخاب خوبی باشند. روش اول یک مزیت محاسباتی دارد (به دلیل flatten کردن تصویر). سایر این تفاوت چندانی ندارند.

---

2) ورودی: 128×128

تعداد فیلتر: 16

سایز kernel: 5×5

stride: 1

Padding: 2

$$\text{خروجی هر لایه} : \frac{n - k + 2p}{s} + 1 = \frac{128 - 5 + 2\times 2}{1} + 1 = 128 \rightarrow \text{خروجی کل} : 128 \times 128 \times 16$$

(input size, kernel size, Padding, Stride)

* هر لایه: conv $\rightarrow 5\times5\times16$, Stride=1, Padding=2
  max Pooling $\rightarrow 2\times2$, Stride=2
  ReLU

$$\text{output size} = \frac{\frac{n-5+2\times2}{1}+1-2}{2}+1 = \frac{n-2}{2}+1 = \frac{n}{2}$$

(-1 و -1 در صورت کسر)

خروجی 3 لایه: $n \times \left(\frac{1}{2}\right)^3 = \frac{n}{8} = \frac{128}{8} = 16 \rightarrow$ خروجی کل 3 لایه: $16 \times 16 \times 16$

(عمق = تعداد فیلتر، سایز تصویر)

تعداد پارامترها لایه‌های conv: $(3\times3\times2+1)\times16 = 448$

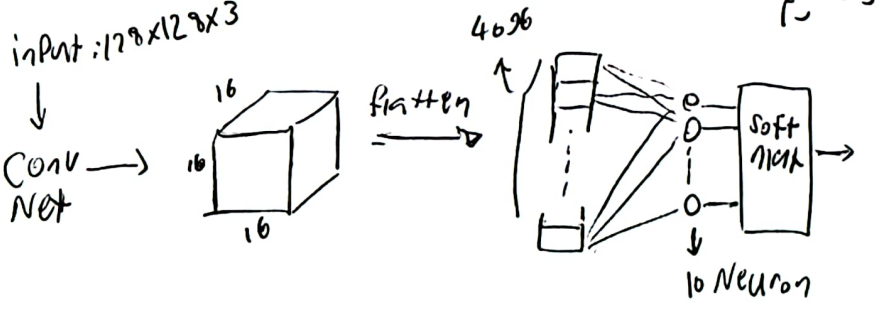
(27 = RGB × kernel size)

تعداد پارامترها در لایه‌های pooling: $\left[(3\times3\times16+1)\times16\right]\times2 = 4640$

$$\#\text{Parameters} = 4640 + 448 = 4640$$

نکته: لایه max Pooling پارامتر ندارد.

* برای حل مسئله Classification تا خروجی لایه کانولوشنال سه‌بعدی را توسط flatten کنیم و وارد یک شبکه FC یا لایه‌های سالم‌سازی مانند Softmax یا Sigmoid کنیم:



$$\Rightarrow \#\text{Parameters} = 4640 + 4096\times10+1 = 45601$$

(4640: conv layers — 40961: FC layer)

* محاسبه Receptive field در شبکه‌ها از روابط بازگشتی زیر حاصل می‌شود:

$$R'_{l} = K_{l-1} + (R'_{l-1} - 1) \times S_{l-1}$$

(Receptive field لایه قبل، لایه فعلی، Stride)

$$R_l = (n_{Pooling}) \times R'_l \quad \text{(در صورت وجود Pooling)} \quad ; \quad R_l = R'_l \quad \text{(در صورت عدم وجود Pooling)}$$

$$\Rightarrow R_3 = 2\left(K_2 + (R_2 - 1) \times 1\right) = 2(R_2 - 1) \overset{(*1)}{=} 2(536 + 4) = 1080 \rightarrow \text{Receptive field}$$

(Pooling = 2، kernel = 5، stride)

$$R_2 = 2\left(K_2 + (R_1 - 1) \times 1\right) = 2(R_1 + 4) \overset{(*1)}{=} 2(264 + 4) = 536 \quad (*2)$$

($K_2 = 5$)

$$R_1 = 2\left(K_1 + (R_0 - 1) \times 1\right) = 2(132) = 264 \quad (*1)$$

($K_1 = 5$، $R_0 = 128$)

---

* سوالات مفهومی U-net

3)

الف) ویژگی اصلی شبکه U-net نسبت به شبکه‌های کانولوشنی عادی، تعداد بلوک‌های کانولوشنی در دو سمت downsampling و upsampling است و وجود skip connection بین بلوک‌های کانولوشنی متناظر در هر دو سمت برای حفظ ویژگی‌های اصلی تصویر است. ساختار U شکل آن نیز به این امر کمک شده است.

ب) وجود skip connection در این شبکه‌ها باعث حفظ ویژگی‌های فضایی (Spatial) عکس از لایه‌های اصلی سمت downsampling شده و همچنین فرآیند error back Propagation را در فرآیند آپدیت وزن‌ها و پارامترهای شبکه، سریع‌تر کرده و مانع پدیده Gradient Vanishing می‌شود.

ج) وجود skip connection با توجه به اینکه باعث حفظ اطلاعات مکانی تصویر می‌شوند، به دلیل کوچک بودن دیتاست تصاویر پزشکی، توانایی این نوع شبکه‌ها می‌تواند در کاربرد Semantic segmentation از اهمیت بالایی برخوردار باشد. در تصاویر پزشکی به دلیل شباهت زیاد بافت (texture) در background و foreground، وجود skip connection می‌تواند با ترکیب ویژگی‌های سطح بالا و سطح پایین تصویر از بافت short skip connection ~ segment کردن عکس کمک کند.

* سوالات مفهومی ResNet و DenseNet:

الف) در شبکه ResNet در بخش‌هایی از شبکه خروجی لایه کانولوشنی با خروجی لایه کانولوشنی دیگری از لایه‌های قبل جمع می‌شود یعنی:

$$x_l = conv(x_{l-1}) + x_{l-1}$$

ادامه حل سوالات مفهومی (بخش الف) DenseNet

ادار شبکه Densenet لایه کانولوشنی از مجموعه‌ای از ویژگی های لایه قبل که به یکدیگر concatenate شده استفاده می‌کنند:

$$x_\ell = conv([concat(x_0, \dots, x_{\ell-1})])$$

تعداد کانال های لایه ℓ

ب) این شبکه با ایجاد Skip connection از لایه های قبلی باعث می‌شود گرادیان از بیشتر مسیر عبور کند و مانع vanishing gradient شود. باتوجه به این نکته می‌توانیم شبکه را در ازای کاهش تعداد لایه ویژگی ها عمیق‌تر کرده و تعداد پارامترهای شبکه را کاهش داد که یکی دیگر از مزیت های این شبکه است.

* سوالات محاسبه‌ای مربوط به U-net

الف) باتوجه به اینکه در شبکه U-net در هر مرحله downSampling سایز تصویر از 4 برابر کوچک می‌شود، پس سایز تصویر (فضای ویژگی) در عمیق‌ترین لایه این شبکه برابر می‌شود با:

$$16 \times 16 \longleftarrow \frac{256}{24} = \frac{256}{16} = 16$$

ب)

$$128 \times (3 \times 3 \times 64 + 1) = 73856$$

(577)

تعداد پارامترهای کانولوشن لایه

* سوالات محاسبه‌ای DenseNet:

الف) باتوجه به اینکه شبکه Densenet تمام فیچرمپ های لایه های قبل از خود را دریافت می‌کند پس 192 = 64 + 128 فیچرمپ از لایه های قبل خود دریافت کرده و با 256 فیچرمپ خود concatinate میکند پس مجموعاً 448 = 256 + 192 فیچرمپ خواهیم داشت

ب)

تعداد کانال خروجی شبکه

$$\begin{cases} \text{خروجی لایه 1} : 32 + k \\ \text{خروجی لایه 2} : 32 + 32 + k + k = 64 + 2k \\ \text{خروجی لایه 3} : \underbrace{32}_{\text{ورودی}} + \underbrace{(32+k)}_{\text{لایه 1}} + \underbrace{(64+2k)}_{\text{لایه 2}} + k = 128 + 4k = 224 \end{cases}$$

پاسخ سوالات تئوری مربوط به تمرین عملی ۲ :

الف) در شبکه های کانولوشن معمولی عملیات grid sampling در کرنل آن ثابت است در حالی که در کانولوشن های از الگوی متفاوتی همراه با offset های یادگیری توسط شبکه استفاده می شود. این بدان معناست که شبکه براساس task مشخص شده، بهترین محل انتخاب sample ها را برای kernel کانولوشن به منظور استخراج بهترین ویژگی ها و افزایش Receptive Field شبکه یاد می گیرد. بنابراین هر فیلتر sample های کرنل خود را از بخش های مهم تصویر برمی دارد.

ب) شبکه های deformable به دلیل اینکه محل انتخاب sample ها را در فرآیند یادگیری فرا می گیرند، انعطاف پذیری بیشتری را نسبت به شبکه های کانولوشنی عادی با grid sampling عادی دارند. زیرا می توانند ویژگی های مختص به هریک از تبدیل های هندسی (geometric Transformations) را با تغییر شکل هندسی کرنل خود استخراج کنند.

ج) شبکه های کانولوشن معمولی به دلیل ثابت بودن ساختار کرنل کانولوشنی خود حساسیت بالایی نسبت به تغییرات هندسی عکس ورودی دارند و به همین خاطر در صورتی که عکس ورودی چرخش زیادی داشته باشد، شبکه قادر به تشخیص و کلاس بندی اشیاء داخل تصویر نیست. به عبارتی شبکه های کانولوشنی معمولی Shift Invariant هستند.