به نام آن‌که جان را فکرت آموخت

تمرین شبکه های عصبی کانولوشنی (CNN)

سپهر ایلامی

۱. برای یک لایه کانولوشنی با ابعاد ورودی W x H، در صورتی که ابعاد کرنل 2K1 x K، تعداد کانال‌های ورودی C_in و تعداد فیلترهای کانولوشنی C_out باشد و پدینگ برابر P و استراید برابر S داشته باشیم (در هر دو جهت یکسان)، موارد زیر را حساب کنید:

الف) تعداد المان‌های ورودی

$$W \times H \times C_{in}$$

ب) سایز خروجی (عرض، طول، تعداد کانال)

$$\left\lfloor \frac{W - K_1 + 2P}{S} \right\rfloor + 1$$
$$\left\lfloor \frac{H - K_2 + 2P}{S} \right\rfloor + 1$$
$$C_{out}$$

پ) تعداد پارامترهایی که در شبکه یاد می‌گیریم

$$(K_1 \times K_2 \times C_{in} + 1) \times C_{out}$$

ت) تعداد ضرب‌هایی که برای محاسبه خروجی انجام می‌دهیم

$$\left(\left\lfloor \frac{W - K_1 + 2P}{S} \right\rfloor + 1\right) \times \left(\left\lfloor \frac{H - K_2 + 2P}{S} \right\rfloor + 1\right) \times C_{out} \times (K_1 \times K_2)$$

---

۲. همه موارد بالا را برای یک لایه Pooling با کرنل به ابعاد K، استراید 2S و پدینگ 2P محاسبه کنید.

الف) تعداد المان‌های ورودی

$$W \times H \times C_{in}$$

ب) سایز خروجی (عرض، طول، تعداد کانال)

$$\left\lfloor \frac{W - K + 2P_2}{S_2} \right\rfloor + 1$$
$$\left\lfloor \frac{W - K + 2P_2}{S_2} \right\rfloor + 1$$
$$C_{out}$$

پ) تعداد پارامترهایی که در شبکه یاد می‌گیریم

$$0$$

- **توضیح: ما در لایه ی Pooling هیچ پارامتری را یاد نمیگیریم. صرفا از بین تعدادی پیکسل ، با روش مشخصی یک خروجی بدست می آوریم. مثلا در max Pooling چند پیکسل را ورودی میگیریم و ماکسیمم آنها را به عنوان خروجی میدهیم.**

ت) تعداد ضرب‌هایی که برای محاسبه خروجی انجام می‌دهیم.

$$0$$

- **توضیح: در لایه ی Pooling هیچ ضربی انجام نمیشود. بلکه صرفا از تعدادی پیکسل ورودی (با توجه به روشمان) یک خروجی را انتخاب میکنیم و به لایه بعدی میدهیم.**

۳. همه موارد بالا را برای یک لایه FC که همان ورودی را به N نرون می‌برد حساب کنید.

الف) تعداد المان‌های ورودی

$$W \times H \times C_{in}$$

ب) سایز خروجی (عرض، طول، تعداد کانال)

$$N$$

- **هر نورون یک خروجی میدهد. پس برای n نورون ما N خروجی داریم.**

پ) تعداد پارامترهایی که در شبکه یاد می‌گیریم

$$(W \times H \times C_{in} + 1) \times N$$

ت) تعداد ضرب‌هایی که برای محاسبه خروجی انجام می‌دهیم.

$$(W \times H \times C_{in}) \times N$$

---

۴. با در نظر گیری مقادیر زیر، تعداد پارامترها، تعداد ضرب‌های لازم برای محاسبات و حجم خروجی را برای هر ۳ سوال قبل حساب کنید.

W = H = 256
K1 = K2 = 3
P = 1
S = 1
C_in = 64
C_out = 128
K = 2
S2 = 2
P2 = 0
N = 1000

FC Layer	Pooling Layer	Convolutional Layer	
4,194,304	4,194,304	4,194,304	تعداد المان‌های ورودی
1000	2,097,152	8,388,608	سایز خروجی (عرض، طول، تعداد کانال)
4,194,305,000	0	73,856	تعداد پارامترهایی که در شبکه یاد می‌گیریم
4,194,304,000	0	75,497,472	تعداد ضرب‌هایی که برای محاسبه خروجی انجام می‌دهیم.

۵. در صورت اجرای شبکه زیر روی GPU با سایز رم ۱۲ گیگ، ماکزیمم تعداد تصویر ورودی که می‌توان در یک بچ گذاشت و خروجی آن‌ها را با یک اجرا روی شبکه به دست آورد چند تا است؟ (برای این کار باید قسمتی از شبکه که بیشترین حجم مصرفی را دارد پیدا کنید.)

Input: 256 x 256	# B 1 256 256
[64] Conv 3 x 3, s=1, p=1	# B 64 256 256
[64] Conv 3 x 3, s=1, p=1	# B 64 256 256
Pool 2 x 2, s=2, p=0	# B 64 128 128
[128] Conv 3 x 3, s=1, p=1	# B 128 128 128
[128] Conv 3 x 3, s=1, p=1	# B 128 128 128
Pool 2 x 2, s=2, p=0	# B 128 64 64
[256] Conv 3 x 3, s=1, p=1	# B 256 64 64
[256] Conv 3 x 3, s=1, p=1	# B 256 64 64
Pool 2 x 2, s=2, p=0	# B 256 32 32
[512] Conv 3 x 3, s=1, p=1	# B 512 32 32
[512] Conv 3 x 3, s=1, p=1	# B 512 32 32
Pool 2 x 2, s=2, p=0	# B 512 16 16
[512] Conv 3 x 3, s=1, p=1	# B 512 16 16
[512] Conv 3 x 3, s=1, p=1	# B 512 16 16
Pool 2 x 2, s=2, p=0	# B 512 8 8
Flatten	# B × 512 × 8 × 8
FC (4096)	# 4096
FC (4096)	# 4096
FC (2)	# 4096

لایه ای بیشترین حجم مصرفی را از RAM میگیرد که حاصلضرب Batch Size در تعداد کانال ها در طول و عرض هر کانال به بیشینه مقدار خود برسد. این عدد در لایه های دوم و سوم ماکسیمم است.

و برابر است با: 4194304 × B که همان $B \times 2^{22}$ است.

اگر هر عدد را float32 در نظر بگیریم ، یعنی 4 بایت جا میخواهد. پس 12 گیگابایت رم حداکثر تعداد 3×2^{30} عدد را میتواند در خودش ذخیره کند.

با برابر قرار دادن دو عبارت بالا یعنی $B \times 2^{22}$ و3×2^{30} ، حداکثر مقدار Batch Size بدست خواهد آمد که این مقدار بیشینه 768 است. این بیشترین اندازه ی Batch Size است.

۶. محدوده Receptive Field یعنی بازه‌ای از تصویر ورودی را که در دید نرون i,j (نرون سطر i و ستون j) در آخرین لایه کانولوشنی بر حسب i و j به دست آورید.

S : Stride
P : Pooling
K : Kernel Size

فرض کنید پیکسل i و j در هرلایه ، متناظر با بازه ی f(i) و f(j) در لایه قبلی باشد.
به همین ترتیب (فارغ از اینکه لایه Conv است یا Pool ، برای هر دو ، فرمول زیر صدق میکند.) تا لایه اول که خود عکس باشد برمیگردیم.

$$f(i) = [i \times S - P, i \times S - P + Kernel - 1]$$
$$f(j) = [j \times S - P, j \times S - P + Kernel - 1]$$

بدین ترتیب اگر فرضا n لایه کانولوشنال و پولینگ در کل داشته باشیم ، نرون i , j در واقع دارد بازه ی زیر را در تصویر اصلی میبیند :

$$[\ (f(i)[0])^n\ , (f(j)[0])^n\ ]$$
