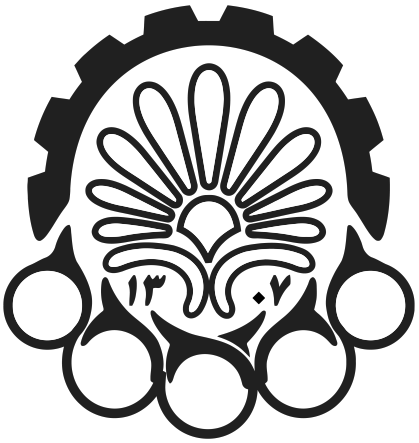
شبکه‌های عصبی و یادگیری عمیق
دکتر صفابخش



دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

تمرین چهارم
شبکه CNN

۲۵ اردیبهشت ۱۴۰۳

شبکه‌های عصبی و یادگیری عمیق

تمرین چهارم

رضا آدینه پور ۴۰۲۱۳۱۰۵۵

دانشکده مهندسی کامپیوتر

سوال اول - نظری

نحوه اشتراک گذاری پارمترها در لایه های کانولوشنی باعث ویژگی Equivariance نسبت به Translation می شود. این ویژگی را شرح دهید و کاربرد آن‌را توضیح دهید.

پاسخ

شبکه‌های CNN دارای ویژگی Equivariance هستند. یعنی با اعمال تبدیلاتی (مانند جابه‌جایی) در ورودی شبکه، تبدیل‌هایی متناظری را در خروجی ایجاد می‌کند. تاکو کوهن در [۱] به عنوان اولین نفر این به این موضوع پرداخت.
اگر تعریف کانولوشن به صورت زیر باشد:

$$(f \star \Psi)(x) \quad = \quad \sum_{y \in \mathbf{Z^2}} \sum_{k=1}^{K} g_k(y) \Psi_k(y - x)$$

در اینجا Ψ و f هردو دارای کانال k هستند. که در این مقاله $k = 1$ درنظر گرفته شده است.
ما در اینجا یک تصویر f داریم که می‌خواهیم آن را با یک کرنل Ψ کانوالو کنیم تا Feature map های تصویر را به‌دست آوریم. سپس می‌خواهیم بدانیم که برای هر تبدیل t آیا دو مورد زیر یکسان است یا خیر:

۱. تبدیل تصویر f با t و کانولوشن حاصل تبدیل با کرنل Ψ

۲. کانولوشن تصویر f با Ψ و سپس تبدیل حاصل با t

بنابر می‌بایست رابطه زیر را اثبات کنیم:

$$(L_t f) \star \Psi \quad = \quad L_t(f \star \Psi)$$

برای اثبات یک تغیر متغیر به صورت $y \leftarrow x + y$ انجام می‌دهیم و رابطه کانولوشن را بازنویسی می‌کنیم:

$$\begin{aligned}(f \star \Psi)(x) \quad &= \quad \sum_{y \in \mathbf{Z^2}} f(y) \Psi(y - x) \\ &= \quad \sum_{y \in \mathbf{Z^2}} f(x + y) \Psi(y)\end{aligned}$$

دو طرف معادله را باتوجه به عبارتی که می‌خواهیم آن را اثبات کنیم بازنویسی می‌کنیم:

پاسخ

$$\begin{aligned}((L_t f) \star \Psi)(x) &= ((f \circ t^{-1}) \star \Psi)(x) \\ &= \sum_{y \in \mathbf{Z^2}} f(t^{-1}(x+y))\Psi(y) \\ &= \sum_{y \in \mathbf{Z^2}} f(x+y-t)\Psi(y)\end{aligned}$$

و $L_t(f \star \Psi)$ به صورت زیر تعریف می‌شود:

$$\begin{aligned}(L_t(f \star \Psi))(x) &= (f \star \Psi)(x-t) \\ &= \sum_{y \in \mathbf{Z^2}} f((x-t)+y)\Psi(y) \\ &= \sum_{y \in \mathbf{Z^2}} f(x+y-t)\Psi(y)\end{aligned}$$

و مشاهده می‌شود که دو طرف تساوی باهم برابر است.
همچنین از کاربردهای آن می‌توان به موارد زیر اشاره کرد:

۱. **Spatial Consistency**
تضمین می‌کند که الگوها یا ویژگی‌ها را می‌توان بدون توجه به موقعیت آنها در ورودی تشخیص داد و شبکه عصبی را در برابر تغییرات و Translation ها انعطاف‌پذیر می‌کند.

۲. **کاهش پیچیدگی**
از آنجایی که پارامترهای یکسان در کل فضای ورودی استفاده می‌شود، CNN ها پارامتر کمتری در مقایسه با شبکه‌های Fully connected با اندازه مشابه دارند.

۳. تعمیم یادگیری

*

References

[1] Cohen T, Welling M. Group equivariant convolutional networks. InInternational conference on machine learning 2016 Jun 11 (pp. 2990-2999). PMLR.

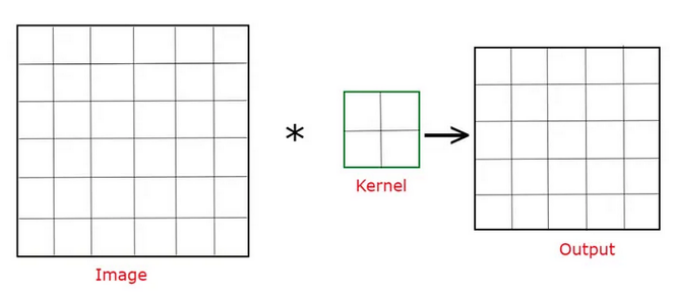
سوال دوم - نظری

شبکه‌های عمیق از عدم تفسیرپذیری رنج می‌برند. تلاش برای حل این مشکل، دو ایده Deconvolutional و Up-convolutional مطرح شده است. بررسی کنید و توضیح دهید هرکدام از دو روش، به چه صورت منجر به تفسیرپذیری می‌شوند؟

پاسخ

پیش از توضیح دادن این دو روش که چگونه به تفسیرپذیری کمک می‌کنند، ابتدا این دو روش را مختثرا توضیح می‌دهیم.

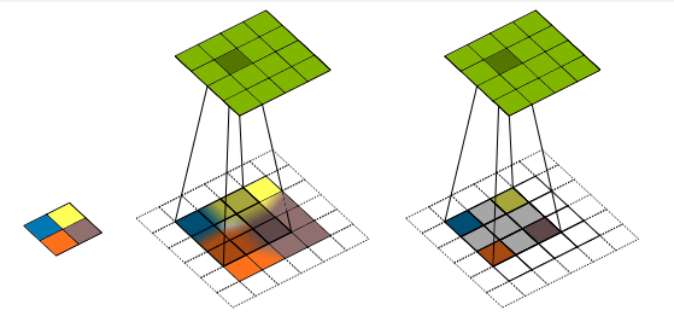
۱. **شبکه Deconvolutional یا Transposed convolutional layer:**
در لایه‌های کانولوشن ویژگی‌های مهم تصویر با استفاده از یک کرنل استخراج می‌شود و خروجی به عنوان Feature map شناخته می‌شود. ابعاد تصویر (ممکن است) کاهش یابد و اطلاعات مهم تصویر حفظ می‌شود.



شکل ۱: لایه کانولوشن

لایه Deconvolution دقیقا برعکس لایه‌های کانولوشنی عمل می‌کند. یعنی از یک Feature map می‌توان به تصویر رسید. الگ.ریتم Deconv با نگاشت نقشه‌های ویژگی به فضای ورودی، این امکان را فراهم می‌کند

۲. **Up-Convolution:** لایه Up-convolution نیز همانند Deconvolution ابعاد ورودی را زیاد می‌کند و هدف آن تولید یک تصویر بزرگتر از ورودی آن است.



شکل ۲: لایه Up-convolution

در بسیاری از مراجه این دو تکنیک را معادل‌با هم می‌گیرند چرا که در هر دو روش هدف افزایش ابعاد ورودی است و این کار دقیقا برعکس کانولوشن انجام می‌شود.

پاسخ

لایه Deconvolution و Up-convolutionبا نمایش نقشه‌های ویژگی به فضای ورودی، به ما امکان می‌دهد ببینیم چه نوع الگوهای ورودی نورون‌های خاصی را فعال می‌کنند. در Up-convolution

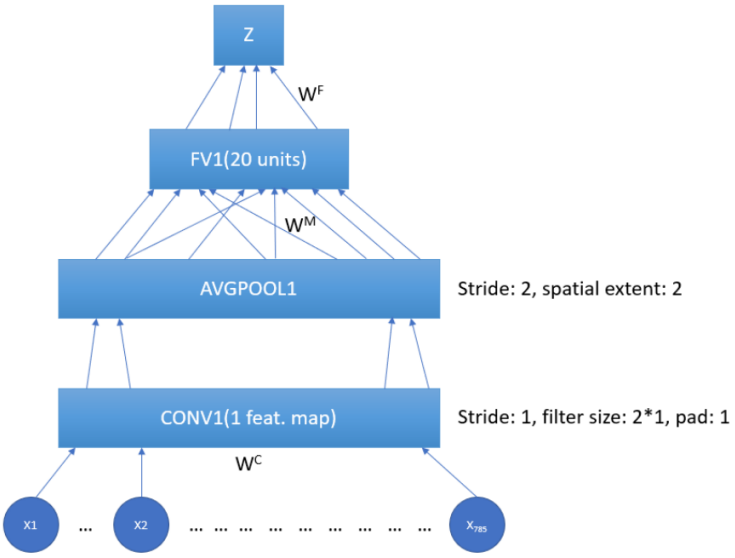
*

References

[1] Durall R, Keuper M, Keuper J. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020 (pp. 7890-7899).

سوال سوم - نظری

معماری شبکه کانولوشنی زیر را درنظر بگیرید:



شکل ۳: شبکه کانولوشنی مورد بررسی در سوال سوم

- ابعاد ورودی 1×785 و خروجی شبکه 1×1
- لایه ورودی X با Zero-padding با طول ۱
- لایه کانولوشنی یک‌بعدی Conv1 با یک کرنل 1×2 و تابع فعال‌سازی ReLU
- لایه Average-polling(AVGPOOL1)
- لایه تمام متصل FC1 با تابع فعال‌سازی ReLU
- لایه خروجی Z که به لایه FC1 کاملا متصل است و تابع فعال‌سازی Sigmoid

وزن لایه FC1 به Z را با W_i^F، بایاس Z را با b^F، وزن لایه AVGPOOL1 به FC1 را با W_{ij}^A، بایاس FC1 را با b_i^M، بردار W^C برابر $[W_1^C, W_2^C]$ و بایاس لایه کانولوشنی را با b^C نشان می‌دهیم. داده‌های مجموعه آموزش به‌صورت X^i و خروجی مورد انتظار به‌صورت Y^i است. همچنین خروجی‌های لایه‌های شبکه به‌ترتیب $c(X^i)$، $a(X^i)$، $f(X^i)$، $z(X^i)$ می‌نامیم. در این صورت، تابع هزینه به‌صورت زیر تعریف می‌شود:

$$cost(X,Y) = \sum_n cost(X^{(n)}, Y^{(n)}) = \sum_n (-Y^{(n)} log(z(X^{(n)})) - (1 - Y^{(n)}) log(1 - z(X^{(n)})))$$

باتوجه به مفروضات بالا، به پرسش‌های زیر پاسخ دهید:

۱. تعداد پارامترهای شبکه بالا را با ذکر جزئیات محاسبه کنید.

پاسخ

بر اساس اطلاعات داده شده برای لایه conv1، تعداد پارامترهای این لایه به‌صورت زیر محاسبه می‌شود:

(آ) وزن ها: ۲

(ب) بایاس: ۱(چون یک کرنل داریم)

بنابراین پارامترهای این لایه می‌شود:

$$\text{conv۱:} 2 + 1 = 3$$

لایه AVGPOOL1 پارامتری ندارد. زیرا تنها down-sampling انجام می‌دهد.
در لایه FC1 داریم:
با توجه به stride=2 ابعاد خروجی برابر است با:

$$\frac{785 + 2 * 0 - 2}{2} + 1 = 393 \rightarrow dim = [1 \times 393]$$

بنابراین تعداد پارامترهای این لایه می‌شود:

(آ) وزن‌ها: 20×393

(ب) بایاس: ۲۰ (تعداد واحدها)

$$(393 \times 20) + 20 = 7860 + 20 = 7880$$

برای لایه خروجی (Z) داریم:

$$20 + 1 = 21$$

و درمجموع تعداد کل پارامترهای شبکه برابر است با:

(آ) لایه کانولوشن: ۳

(ب) لایه پولینگ: ۰

(ج) لایه تمام متصل:۷۸۸۰

(د) لایه خروجی: ۲۱

$$\text{parameters: Total} 3 + 0 + 7880 + 21 = 7904$$

۲. برای فقط یک نمونه آموزشی، مقدار $\frac{\partial Cost}{\partial W_1^C}$ و $\frac{\partial Cost}{\partial W_{ji}^A}$ را با جزئیات محاسبه کنید.

سوال چهارم - نظری

کانولوشن متسع[۱] روشی برای افزایش میدان پذیرش (Receptive field) شبکه‌های کانولوشنی است که به‌صورت زیر تعریف می‌شود: (دقت شود خروجی تنها برای اندیس‌هایی از کرنل و تصویر همپوشانی کامل دارند، محاسبه می‌شود)

$$(k * I)(i, j) = \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} K(m, n) I(i + D_m, j + D_n)$$

۱. در یک شبکه کانولوشنی با یک لایه کانولوشن $K \times K$ با طول گام یک، عرض میدان پذیرش را بدست آورید.

۲. برای ورودی $I \in \mathbf{R}^{M \times N}$ و کرنل $K \in \mathbf{R}^{F \times F}$، نشان دهید خروجی عملگر متسع دارای ابعاد $(M - DF + D) \times (N - DF + D)$ است. متغیر D به معنی Dilation است.

۳. نشان دهید کانولوشن متسع معادل کانولوشن با کرنل متسع شده $K' = K \otimes A$ است. ماتریس A را مشخص کنید. (عملگر $\otimes$ به معنی Kronecker product است.)

[۱] Dilated convolution

سوال پنجم - عملی

شبکه‌های کانولوشنی با توجه به توانایی آن‌ها در استخراج و یادگیری خودکار ویژگی‌ها، مقاومت نسبت به تغییرات و کارایی آن‌ها در مقابل پیچیدگی‌های وظیفه‌ی بازشناسی چهره، یک عنصر اساسی در اکثر اسن سیستم‌ها هستند. در این تمرین قصد داریم که با استفاده از شبکه‌های عصبی کانولوشنی به تحلیل احساسات چهره[۲] و طبقه‌بندی آن‌ها از روی تصویر بپردازیم. مجموعه داده‌ی این تمرین شامل ۱۲۰۰ تصویر نمونه‌گیری شده از هر کلاس مجموعه AffectNet می‌باشد. مجموعه داده AffectNet شامل ۴۵۰ هزار تصویر چهره با ۸ حالت مختلف می‌باشد که شکل ۴ نمونه‌هایی از آن را نشان می‌دهد.

شکل ۴: نمونه‌هایی از مجموعه داده AffectNet

۱. **پیش‌پردازش و داده افزایی:** مجموعه داد را از این لینک دانلود کنید و از هر کلاس سه نمونه را نمایش دهید. برای افزایش سرعت آموزش، تمامی تصاویر را به بازه $[0, 1]$ نرمال‌سازی کنید. همچنین داده‌ها را با پردازش مناسب افزونه کنید. توضیح دهید که به‌نظر شما استفاده از چه پردازش‌هایی در این حالت مناسب است و چرا در این مسئله نیاز به داده‌افزایی وجود دارد؟ از هر کلاس سه نمونه‌ی افزونه شده را نمایش دهید و همچنین تعداد کل نمونه‌ها پیش و پس از داده‌افزایی را در گزارش خود بیاورید.

۲. **یادگیری انتقالی** یک رویکرد رایج در هوش مصنوعی است که از یک مدل از قبل آموزش دیده برای یک وظیفه متفاوت اما مرتبط استفاده می‌کند و آن را با وظایف جدید تطبیق می‌دهد. با استفاده از شبکه پیش آموزش دیده VGG16 وظیفه بازشناسی حالت چهره را بر روی مجموعه داده ارائه شده انجام دهید. برای فرآیند آموزش، از داده‌های موجود در پوشه Train استفاده کنید. نمودار خطا و دقت در فرآیند آموزش و نمودار ROC و ماتریس درهم‌ریختگی را برای داده‌های موجود در پوشه Validation گزارش کنید.

به‌کارگیری شبکه‌های ازپیش آموزش دیده به‌طور خاص در زمانی که داده‌ی کمی وجود دارد مزایای زیادی دارد اما این شبکه‌ها با توجه به معماری ازپیش تعریف شده و نسبتا سنگین آنها برای استفاده در ابزارهای کاربردی مانند تلفن‌همراه مناسب نیستند. مدل‌های موجود در تلفن های همراه باید نیازهای ذخیره‌سازی را به حداقل برسانند و درعین حال افت عملکرد قابل توجهی نداشته باشند. برای دستیابی به این امر، در این مقاله سه معماری سبک از سه شبکه کانولوشنی مطرح یعنی VGG، AlexNet و MobileNet مطرح شده است. نتایج به‌دست آمده نشان می‌دهد که این سه معماری عملکرد مشابهی نسبت به آخرین مدل‌های پیشرو در این زمینه دارند.

[۲] Facial expression recognition

۳. معماری مطرح شده برای شبکه VGG که جزئیات آن در شکل ۵ آمده است را پیاده‌سازی کنید. این مدل را بر روی مجموعه داده ارائه شده آموزش دهید و نمودار خطا و دقت آن را رسم کنید. همچنین با استفاده از داده موجود در پوشه Validation مدل را تست کنید و نمودار ROC و ماتریس درهم‌ریختگی آن را گزارش کنید. تعداد پارامترهای این مدل و عملکرد آن را با مدل قسما قبل مقایسه و تحلیل کنید.

Type	Shape	Output
2×Conv	$3 \times 3 \times 16$	$128 \times 128 \times 16$
MaxPool	2×2	$64 \times 64 \times 16$
2×Conv	$3 \times 3 \times 32$	$64 \times 64 \times 32$
MaxPool	2×2	$32 \times 32 \times 32$
2×Conv	$3 \times 3 \times 64$	$32 \times 32 \times 64$
MaxPool	2×2	$16 \times 16 \times 64$
2×Conv	$3 \times 3 \times 128$	$16 \times 16 \times 128$
MaxPool	2×2	$8 \times 8 \times 128$
2×Conv	$3 \times 3 \times 128$	$8 \times 8 \times 128$
MaxPool	2×2	$4 \times 4 \times 128$
Flatten	2048	–
2×Dense	1024	–
Dense	8 or 2	1 label or 2 floats

شکل ۵: معماری شبکه VGG ارائه شده در مقاله

۴. برای درک هرچه بهتر عملکرد شبکه‌های کانولوشنی ابزارهای متنوعی وجود دارد. یکی از این ابزارها نقشه‌ی فعال‌سازی کلاس[۳] یا به اختصار CAM است که یک نمونه از آن در شکل ؟؟ آمده است. بررسی کنید که استفاده از این ابزار چه پیش‌بینی برای بهبود شبکه‌های کانولوشنی فراهم می‌آورد. برای دو نمونه با اشتباه دسته بندی شده و دو نمونه به درستی دسته بندی شده‌ی به ازای هر کلاس در مدل سوال ۳ نقشه‌ی فعالسازی کلاس را به‌دست آورید و با تحلیل نتایج به‌دست آمده، رویکردی برای بهبود شبکه پیشنهادی سوال ۳ ارائه دهید.

[۳] Class activation map