تمرین تئوری سری اول مبانی یادگیری عمیق

کورش تقی پور پاسدار 400521207

۱۹ آذر ۱۴۰۳

فهرست مطالب

۱ سوال اول

۱.۱ الف

۱.۱.۱ فرمول کلی

در لایه‌های کانولوشنی، تعداد پارامترها از فرمول زیر بدست می‌آید.

$$P = (Cin \times H \times W + 1) \times Cout$$

لایه‌های MaxPolling، AvgPolling، Dropout پراامتری ندارند. تعداد پارامترهای لایه خطی هم از فرمول زیر بدست می‌آید.

$$P = (Nin + 1) \times Nout$$

که Nin به تعداد ویژگی‌های ورودی و Nout تعداد ویژگی‌های خروجی اشاره دارد.

Layer 1 ۲.۱.۱

در این لایه اندازه خروجی برابر با $128 \times 128 \times 32$ می‌باشد. همچنین تعداد پارامترهای آن برابر با 4736 می‌باشد.میدان تاثیر این لایه برابر با 49 می‌باشد.

Layer 2 ۳.۱.۱

در این لایه اندازه خروجی برابر با $62 \times 62 \times 64$ می‌باشد. تعداد پارامترهای آن نیز برابر با 51264 می‌باشد. همچنین میدان تاثیر آن نیز برابر با 1225 پیکسل از ابعاد تصویر ورودی می‌باشد.

Layer 3 ۴.۱.۱

در این لایه، اندازه خروجی برابر با $31 \times 31 \times 64$ می‌باشد. این لایه پارامتری ندارد و میدان تاثیر لایه برابر با 4900 پیکسل از تصویر ورودی می‌باشد.

Layer 4 ۵.۱.۱

در این لایه، اندازه خروجی برابر با $27 \times 27 \times 128$ می‌باشد. این لایه 73856 پارامتر دارد.

Layer 5 ۶.۱.۱

در این لایه، اندازه خروجی برابر با $25 \times 25 \times 128$ می‌باشد. این لایه 147584 پارامتر دارد.

Layer 6 ۷.۱.۱

در این لایه، اندازه خروجی برابر با $12 \times 12 \times 128$ می‌باشد. این لایه پارامتری ندارد.

Layer 7 ۸.۱.۱

در این لایه، اندازه خروجی برابر با $10 \times 10 \times 256$ می‌باشد. این لایه 295168 پارامتر دارد.

Layer 8 ۹.۱.۱

در این لایه، اندازه خروجی برابر با $5 \times 5 \times 256$ می‌باشد. این لایه پارامتری ندارد.

Layer 9 ۱۰.۱.۱

در این لایه (فرض می‌شود یک لایه Flatten قبل آن اعمال شده است) اندازه خروجی برابر با 1024 بوده و تعداد پارامترهای آن 6554624 می‌باشد.

Layer 10 ۱۱.۱.۱

در این لایه اندازه خروجی تغییری نکرده و پارامتری هم ندارد.

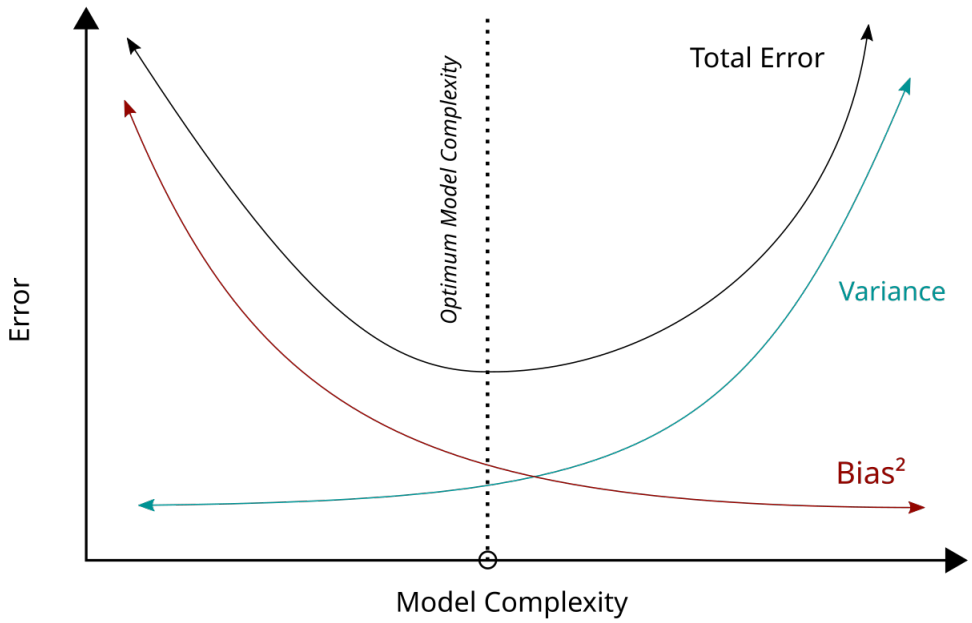
Layer 11 ۱۲.۱.۱

در این لایه تعداد پارامترها برابر با 10250 می‌باشد و اندازه خروجی هم برابر با 10 می‌باشد.

۲ سوال دوم

۱.۲ الف

بایاس به طور خلاصه به میزان خطای ناشی از ساده‌سازی بیش از حد مدل و واریانس به حساسیت بیش از حد مدل به داده‌های آموزشی اشاره دارد. این دو مورد در تضاد با یکدیگر بوده و افزایش قدرت مدل به افزایش واریانس و کاهش بایاس منجر می‌شود و برعکس. در حالت‌هایی که مدل ساده بوده،مدل توانایی یادگیری کافی داده‌ها را ندارد. به همین دلیل در این حالت بایاس مدل زیاد بوده و واریانس آن کم می‌باشد. در حالتی که مدل پیچیده است، مدل می‌تواند داده‌ها را یادبگیرد و بایاس آن کم است، ولی ممکن است حساسیت بیش از حد به داده‌ها پیدا کند و واریانس آن زیاد شود. در تصویر زیر می‌توان به درک شماتیکی از این دو مفهوم رسید.



شکل ۱: Trade-off Bias-Variance

۲.۲ ب

درصورتی‌که عملکرد مدل عصبی برروی داده‌های آموزشی خوب باشد ولی برروی داده‌های تست پایین باشد، مدل عصبی دچار Overfitting شده است. Overfitting زمانی رخ می‌دهد که مدل بجای الگوهای کلی بین داده‌ها، نویزها را نیز یاد بگیرد یا به عبارت دیگر، داده‌ها را حفظ کند. از راه‌کارهای مقابله با آن می‌توان به استفاده از dropout، کاهش قدرت مدل عصبی یا کاهش تعداد epoch ها اشاره کرد.

۳.۲ ج

مشکل ناپدید شدن گرادیان، به کاهش بسیار زیاد گرادیان به خصوص برای پارامترهای لایه‌های ابتدایی‌تر گفته می‌شود که سبب می‌شود تا بدلیل کاهش بسیار زیاد گرادیان محاسبه شده و نزدیک صفر بودن آن، پارامترهای آن لایه‌ها در عمل تغییر چندانی نکرده و آموزش نبینند. از طرفی، انفجار گرادیان به مقداردهی بسیار زیاد گرادیان گفته می‌شود که سبب می‌شود تا پارامتر مربوط به آن گرادیان به نقطه نامعلومی رفته و تابع مدل تغییر بسیار زیادی بکند.
ناپدید شدن گرادیان عمدتا در توابع فعالسازی مانند Sigmoid که مشتق آنها معمولا کوچک است رخ می‌دهد و زمانی که از این توابع در لایه‌های میانی استفاده کنیم که سبب می‌شود با ضرب چندباره گرادیان در این مقادیر، گرادیان بسیار کوچک شود.
انفجار گرادیان هم در توابعی مانند ReLU رخ می‌دهد که خروجی آنها به یک بازه محدود نیست.

۴.۲ د

عمیق‌تر کردن شبکه MLP سبب می‌شود تا تعداد نورون‌های لازم به طور نمایی کاهش یابد. به عبارت دیگر، یک شبکه MLP سطحی [۱] نسبت به یک شبکه عمیق [۲] که هردو عملکرد مشابهی دارند، تعداد نورون‌های بیشتری (از مرتبه نمایی) دارد. درمورد بررسی نتایج شبکه‌های عمیق و سطحی، باید گفت که درصورتی که در شبکه سطحی از نورون‌های کافی استفاده شود، می‌تواند همانند شبکه عمیق و یا حتی بهتر عمل کند. به عبارت دیگر، عمیق بودن شبکه لزوما به معنی کسب نتایج بهتر نیست. البته لازم به ذکر است که شبکه‌های عمیق به دلیل تعداد کمتر نورون‌ها، نیاز به بهینه‌کردن نورون‌های کمتری بوده، در نتیجه آموزش سریعتر و در نتیجه همگرایی [۳] سریعتری دارند.

۳ سوال سوم

با توجه به عملیات کانولوشن، حاصل اعمال فیلتر برروی ورودی بصورت زیر خواهد بود.

7	-18
10	4

حال با اعمال لایه ادغام حداکثر سراسری[۴]، خروجی برابر با 10 خواهد شد. حال خانه‌های خروجی را بصورت z_1, z_2, z_3, z_4 نمایش دهیم که به ترتیب از چپ به راست با خانه‌های بالا سمت چپ، بالا سمت راست، پایین سمت چپ، پایین سمت راست متناظر خواهند بود. همچنین خروجی لایه ادغام حداکثر سراسری را هم a در نظر می‌گیریم.
حال طبق بیان سوال گرادیان تابع ضرر نسبت به خروجی نهایی را داریم.

$$\frac{dLoss}{da} = 1$$

از طرفی با توجه به تابع حداکثری[۵]،مقدار گرادیان تنها به خانه‌ای با بیشترین مقدار انتقال خواهد یافت و گرادیان سایر خانه‌ها برابر با صفر خواهد بود.

$$\frac{da}{dz_1} = 0, \frac{da}{dz_2} = 0$$
$$\frac{da}{dz_3} = 1, \frac{da}{dz_4} = 0$$

از طرفی اگر وزن‌های فیلتر F را به ترتیب بالا سمت چپ، بالا سمت راست، پایین سمت چپ و پایین سمت راست متناظرا برابر با w_1, w_2, w_3, w_4 در نظر می‌گیریم، خروجی هر فیلتر برابر با $x_1w_{1i} + x_2w_{2i} + x_3w_{3i} + x_4w_{4i} + b = z_i$ خواهد

[1] Shallow
[2] Deep
[3] Convergence
[4] Global max pooling(GAP)
[5] Max

بود. با توجه به این فرمول، گرادیان هر وزن براساس z_i بصورت زیر خواهد بود.

$$\frac{dz_i}{dw_1} = x_1, \frac{dz_i}{dw_2} = x_2$$
$$\frac{dz_i}{dw_3} = x_2, \frac{dz_i}{dw_4} = x_4$$
$$\frac{dz_i}{db} = 1$$

با توجه به موارد بالا، گرادیان هر وزن بصورت زیر محاسبه خواهد شد.

$$\begin{aligned}
\frac{dLoss}{dw_1} &= \frac{dLoss}{da} \times \frac{da}{dz_3} \times \frac{dz_3}{dw_1} = -1 \\
\frac{dLoss}{dw_2} &= \frac{dLoss}{da} \times \frac{da}{dz_3} \times \frac{dz_3}{dw_2} = 5 \\
\frac{dLoss}{dw_3} &= \frac{dLoss}{da} \times \frac{da}{dz_3} \times \frac{dz_3}{dw_3} = 3 \\
\frac{dLoss}{dw_4} &= \frac{dLoss}{da} \times \frac{da}{dz_3} \times \frac{dz_3}{dw_4} = 0 \\
\frac{dLoss}{db} &= \frac{dLoss}{da} \times \frac{da}{dz_3} \times \frac{dz_3}{db} = 1
\end{aligned}$$

۴ سوال چهارم

در ابتدا داده اول را وارد شبکه می‌کنیم.

$$y = 1 \times 1^2 - 1 \times (-1)^2 - 1 \times 1 \times -1 + 1 = 2 \quad (۱)$$

سپس به محاسبه تک تک مشتق‌ها می‌پردازیم.

$$\begin{aligned}
\frac{dLoss}{dy} &= -2 \times (10 - 2) = -16 \\
\frac{dy}{da} &= x_1^2 = 1 \\
\frac{dy}{db} &= x_2^2 = 1 \\
\frac{dy}{dc} &= x_1 x_2 = -1 \\
\frac{dy}{dd} &= 1
\end{aligned}$$

حال به محاسبه مقادیر جدید پارامترها می‌پردازیم.

$$\begin{aligned}
\Delta a^1 &= \beta \Delta a^0 + \mu \nabla_a E = 0.9 \times 0 + 0.1 \times (-16 \times 1) = -1.6 \\
a^1 &= a^0 - \Delta a^1 = 1 - (-1.6) = 2.6 \\
\Delta b^1 &= \beta \Delta b^0 + \mu \nabla_b E = 0.9 \times 0 + 0.1 \times (-16 \times 1) = -1.6 \\
b^1 &= b^0 - \Delta b^1 = -1 - (-1.6) = 0.6 \\
\Delta c^1 &= \beta \Delta c^0 + \mu \nabla_c E = 0.9 \times 0 + 0.1 \times (-16 \times -1) = 1.6 \\
c^1 &= c^0 - \Delta c^1 = -1 - (1.6) = -2.6 \\
\Delta d^1 &= \beta \Delta d^0 + \mu \nabla_d E = 0.9 \times 0 + 0.1 \times (-16 \times 1) = -1.6 \\
d^1 &= d^0 - \Delta d^1 = 1 - (-1.6) = 2.6
\end{aligned}$$

حال به وارد کردن داده دوم به شبکه می‌پردازیم.

$$y = 2.6 \times 2^2 + 0.6 \times (0)^2 - 2.6 \times 2 \times 0 + 2.6 = 13 \qquad (۲)$$

سپس به محاسبه تک تک گرادیان‌ها می‌پردازیم.

$$\begin{aligned}
\frac{dLoss}{dy} &= -2 \times (13 - 13) = 0 \\
\frac{dy}{da} &= x_1^2 = 4 \\
\frac{dy}{db} &= x_2^2 = 0 \\
\frac{dy}{dc} &= x_1 x_2 = 0 \\
\frac{dy}{dd} &= 1
\end{aligned}$$

حال به محاسبه مقادیر جدید پارامترها می‌پردازیم.

$$\begin{aligned}
\Delta a^2 &= \beta \Delta a^1 + \mu \nabla_a E = 0.9 \times -1.6 + 0.1 \times (0 \times 4) = -1.44 \\
a^2 &= a^1 - \Delta a^2 = 2.6 - (-1.44) = 4.04 \\
\Delta b^2 &= \beta \Delta b^1 + \mu \nabla_b E = 0.9 \times -1.6 + 0.1 \times (0 \times 0) = -1.44 \\
b^2 &= b^1 - \Delta b^2 = 0.6 - (-1.44) = 2.04 \\
\Delta c^2 &= \beta \Delta c^1 + \mu \nabla_c E = 0.9 \times 1.6 + 0.1 \times (0 \times 0) = 1.44 \\
c^2 &= c^1 - \Delta c^2 = -2.6 - (1.44) = -4.04 \\
\Delta d^2 &= \beta \Delta d^1 + \mu \nabla_d E = 0.9 \times -1.6 + 0.1 \times (0 \times 1) = -1.44 \\
d^2 &= d^1 - \Delta d^2 = 2.6 - (-1.44) = 4.04
\end{aligned}$$

حال مقادیر نهایی پارامترها بصورت زیر می‌باشد.

$$a = 4.04 \qquad b = 2.04 \qquad c = -4.04 \qquad d = 4.04$$