



یادگیری عمیق

نیم سال دوم ۰۴-۰۵
مدرس: مهدیه سلیمانی

ددلاین تمرین: ۸ خرداد

تمرین دوم

- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. مهلت تاخیر (مجاز و غیر مجاز) برای این تمرین ۷ روز است (یعنی حداکثر تاریخ ارسال تمرین ۱۵ خرداد است)
- در هر کدام از سوالات، اگر از منابع خارجی استفاده کرده‌اید باید آن را ذکر کنید. در صورت همفکری با افراد دیگر هم باید نام ایشان را در سوال مورد نظر ذکر نمایید.
- پاسخ تمرین باید ماحصل دانسته‌های خود شما باشد. در صورت رعایت این موضوع، استفاده از ابزارهای هوش مصنوعی با ذکر نحوه و مصداق استفاده بلامانع است.
- پاسخ ارسالی واضح و خوانا باشد. در غیر این صورت ممکن است منجر به از دست دادن نمره شود.
- از تحویل تمرین نظری به شکل تایپ شده (فایل Word-Latex یا هر نوع تایپ شده دیگری) بپرهیزید و حتما پاسخ دست نویس تحویل دهید.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد. به اسکرین‌شات از منابع یا پاسخ افراد دیگر نمره‌ای تعلق نمی‌گیرد.
- در صورتی که بخشی از سوال‌ها را جای دیگری آپلود کرده و لینک آن را قرار داده باشید، حتما باید تاریخ آپلود مشخص و قابل اتکا باشد.
- محل بارگذاری سوالات نظری و عملی در هر تمرین مجزا خواهد بود. به منظور بارگذاری بایستی تمرین تئوری در یک فایل pdf با نام `HW1_[First-Name]_[Last-Name]_[Student-Id].pdf` بارگذاری شوند.
- در صورت وجود هرگونه ابهام یا مشکل، در کوثرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.
- طراحان این تمرین: علیرضا علایی، مبین شریفی‌راد، عرفان صبحائی، شایگان ادیم، امیرحسین ایزدی، امیرحسین کوچکیان، محمد امانلو، عرفان یگانگی دستگردی، حسین شاه‌آبادی

بخش نظری (۱۰۰+۱۰ نمره)

پرسش ۱. عملگر کانولوش (۱۵ نمره)

فرض کنید $I \in \mathbb{R}^{H \times W}$ یک تصویر و عملگر انتقال^۱ را به صورت زیر تعریف می‌کنیم:

$$(T_{a,b}I)(i,j) = I(i-a, j-b)$$

همچنین یک عملگر $\mathcal{L}$ را هموردا نسبت به انتقال^۲ گویند اگر برای هر انتقال (a,b) داشته باشیم:

$$\mathcal{L}(T_{a,b}I) = T_{a,b}(\mathcal{L}(I)).$$

^۱ Translation Operator
^۲ Translation Equivariant

همچنین کانولوشن با هسته $K \in \mathbb{R}^{k \times k}$ به صورت زیر تعریف می‌شود:

$$(K * I)(i, j) = \sum_{m, n} K(m, n) I(i - m, j - n).$$

(آ) نشان دهید کانولوشن نسبت به انتقال هموردا است.

(ب) ثابت کنید هر عملگر خطی که نسبت به انتقال هموردا است، یک کانولوشن است. توضیح دهید این نتیجه چه تاثیری بر انتخاب شبکه‌های کانولوشنی و طراحی آن‌ها دارد.

(ج) فرض کنید تصاویر قبل ورود به شبکه به بردار $x \in \mathbb{R}^{HW}$ تبدیل می‌شوند.

ابتدا نشان دهید که ماتریس W متناظر با تبدیل خطی که نسبت به انتقال هموردا است، یک ماتریس Toeplitz است. سپس توضیح دهید چرا یک لایه خطی موجود در MLP به طور کلی نمی‌تواند نسبت به انتقال هموردا باشد.

(د) فرض کنید از یک شبکه عصبی دو لایه برای تشخیص یک الگو در یک تصویر استفاده می‌کنیم. حداقل تعداد پارامتر قابل یادگیری مورد نیاز در لایه اول برای تشخیص یک الگوی $k \times k$ را به ازای لایه اول کانولوشنی و خطی کاملاً متصل محاسبه و مقایسه کنید.

تعریف: ماتریس Toeplitz، ماتریسی است که درایه‌های هر قطر موازی با قطر اصلی، با هم برابر باشند. به عبارت دیگر مقدار درایه‌ها تابعی از اختلاف شماره سطر و ستون آن است. برای مثال ماتریس T که به صورت زیر تعریف شده، یک ماتریس Toeplitz است.

$$T = \begin{bmatrix} T_0 & T_{-1} & T_{-2} & T_{-3} \\ T_1 & T_0 & T_{-1} & T_{-2} \\ T_2 & T_1 & T_0 & T_{-1} \\ T_3 & T_2 & T_1 & T_0 \end{bmatrix}$$

پرسش ۲. محاسبات Forward و Backward در شبکه‌های کانولوشنی (۱۵ نمره)

فرض کنید ماتریس ورودی X و کرنل W به صورت زیر داده شده‌اند.

$$X = \begin{bmatrix} 1 & 2 & 0 & 3 & 1 \\ 2 & 1 & 3 & 0 & 2 \\ 0 & 1 & 2 & 1 & 1 \\ 3 & 0 & 1 & 2 & 0 \\ 1 & 2 & 1 & 0 & 3 \end{bmatrix} \quad W = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 1 \\ -1 & 1 & 0 \end{bmatrix}$$

داده شده‌اند. در این مسئله، کانولوشن با گام (Stride) برابر ۲ و پدینگ (Padding) برابر ۱ انجام می‌شود.
(آ) خروجی کانولوشن $Y = X * W$ را محاسبه کنید. ابعاد خروجی را مشخص کرده و مقدار هر درایه را بنویسید.
(ب) اگر تابع هزینه به صورت

$$L = \frac{1}{2} \sum_{i,j} (Y_{ij} - T_{ij})^2$$

تعریف شود، مشتقات زیر را استخراج کنید:

$$\frac{\partial L}{\partial Y}, \quad \frac{\partial L}{\partial W}, \quad \frac{\partial L}{\partial X}.$$

نتیجه را به صورت ماتریسی بنویسید و توضیح دهید که پدینگ و گام چگونه بر ابعاد و مقادیر مشتقات اثر می‌گذارند.

پرسش ۳. Efficient CNNs (۲۰ نمره)

در سال‌های اخیر، روند طراحی شبکه‌های عصبی کانولوشنی (CNN) به سمت مدل‌های عمیق‌تر و بزرگ‌تر برای دستیابی به دقت بالاتر حرکت کرده است. با این حال، در بسیاری از کاربردهای عملی مانند تلفن‌های هوشمند، منابع محاسباتی بسیار محدود بوده و بودجه محاسباتی معمولاً در حد ده‌ها تا چند صد مگافلاپس (MFLOPs) است. از این رو، تلاش بسیاری برای طراحی معماری‌های کارآمد انجام شده است که بتوانند در چنین محدودیت‌هایی عملکرد مناسبی ارائه دهند. در این سوال به بررسی ایده‌های معماری ShuffleNet می‌پردازیم که منجر به عملکرد مناسب به ازای بودجه محاسباتی محدود می‌شود. این معماری با کاهش هزینه محاسباتی اجزای اصلی شبکه، موفق به ارائه شبکه‌های بزرگ اما کارآمد شده است.

۱ کانولوشن عمقی قابل تفکیک

(آ) هزینه محاسباتی کانولوشن استاندارد برای اندازه ورودی $H \times W$ ، تعداد کانال ورودی C_{in} ، تعداد کانال خروجی C_{out} و اندازه فیلتر K را بر حسب FLOPs محاسبه کنید. فرض کنید طول گام برابر یک است و با استفاده از Padding ابعاد مکانی حفظ می‌شود.

(ب) کانولوشن عمقی قابل تفکیک^۳ روشی برای کاهش هزینه محاسباتی کانولوشن استاندارد می‌باشد. این روش را به طور کامل توضیح دهید.

(ج) مشابه بخش الف، هزینه محاسباتی کانولوشن عمقی قابل تفکیک را بدست آورید و با کانولوشن استاندارد مقایسه کنید.

(د) بررسی‌های عملی نشان می‌دهند کانولوشن عمقی قابل تفکیک جایگزین مناسبی برای کانولوشن است. فرض کنید ماتریس کانولوشن به فرم $W \in \mathbb{R}^{C_{out} \times k^2 C_{in}}$ مورد استفاده قرار می‌گیرد. بنابراین برای هر قطعه^۴ $k \times k$ که به فرم $x \in \mathbb{R}^{k^2 C_{in}}$ تبدیل می‌شود می‌توان حاصل عملیات لایه کانولوشنی را به صورت ضرب ماتریس نمایش داد. ثابت کنید

(آ) کانولوشن عمقی قابل تفکیک ماتریس کانولوشن W را به شکل ضرب دو ماتریس نمایش می‌دهد.

(ب) این ضرب ماتریسی به ازای ماتریس دلخواه W دقیق نخواهد بود.

(ج) با فرض برابر بودن رتبه^۵ ضرب این دو ماتریس با r ، این ضرب ماتریسی در حالت کلی بهترین تقریب رنک r از ماتریس W در نرم طیفی^۶ نخواهد بود.

۲ کانولوشن گروهی

در بسیاری از شبکه‌های کانولوشنی، بخش قابل توجهی از هزینه محاسباتی شبکه مربوط به کانولوشن‌های 1×1 ^۷ است. همچنین در بخش قبل کاربرد این نوع کانولوشن را دیدیم.

(آ) توضیح دهید که کانولوشن 1×1 به چه منظور در معماری‌هایی مانند GoogLeNet و ResNet استفاده می‌شود.

(ب) معماری ShuffleNet برای کاهش سربار محاسباتی کانولوشن 1×1 از ایده‌ی کانولوشن گروهی^۸ استفاده می‌کند. این روش را توضیح دهید.

Depthwise Separable Convolution^۳

Patch^۴

Rank^۵

Spectral Norm^۶

Pointwise Convolution^۷

Group Convolution^۸

ج) هزینه محاسباتی کانولوشن گروهی را برای اندازه ورودی $H \times W$ ، تعداد کانال ورودی C_{in} ، تعداد کانال خروجی C_{out} ، تعداد گروه g و اندازه فیلتر K را بر حسب FLOPs محاسبه کنید. فرض کنید طول گام برابر یک است و با استفاده از padding ابعاد مکانی حفظ می‌شود. آیا این روش منجر به بهبود هزینه محاسباتی در کانولوشن 1×1 می‌شود؟

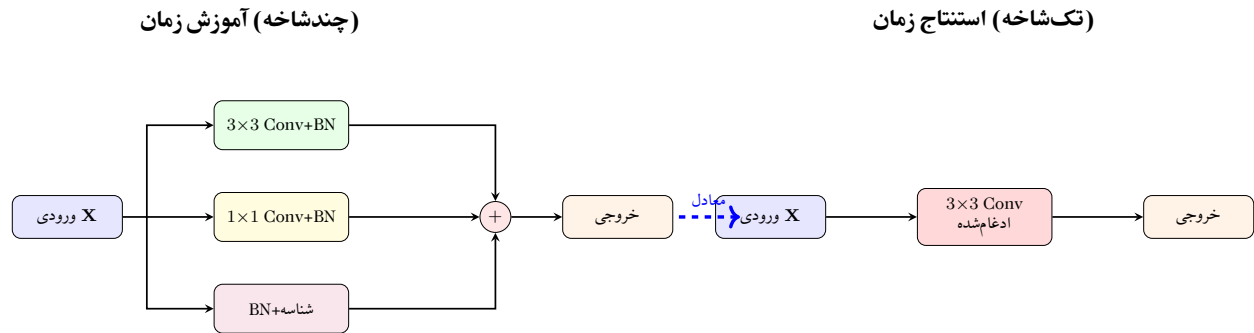
۳ شافل کردن کانال‌ها

آ) فرض کنید تعدادی لایه با کانولوشن گروهی پشت هم قرار بگیرند. در این صورت چه مشکلی برای جریان اطلاعات و استخراج ویژگی می‌تواند ایجاد شود.

ب) معماری ShuffleNet از ایده شافل کردن کانال‌ها بعد از اعمال کانولوشن گروهی بهره می‌برد. این ایده و نحوه حل کردن مشکل بخش الف را توضیح دهید. همچنین توضیح دهید به چه شکل این روش مشکلی برای جریان گرادینان در شبکه ایجاد نمی‌کند.

پرسش ۴. بازپارامتری سازی ساختاری: از چندشاخه به تک شاخه (۲۰ نمره)

پیش زمینه: معماری هایی مثل ResNet و GoogLeNet از ساختارهای چندشاخه ای برای بهبود جریان گرادین استفاده می کنند. اما در زمان استنتاج (Inference)، هر شاخه ای اضافی نیاز به حافظه و محاسبات مستقل دارد. در سال ۲۰۲۱، ایده ی بازپارامتری سازی ساختاری (Structural Re-parameterization) در معماری RepVGG مطرح شد: در زمان آموزش از چند شاخه استفاده کنیم و در زمان استنتاج، همه ی شاخه ها را در یک کانولوشن واحد ادغام (Fuse) کنیم. مزیت: هم از مزایای آموزش چندشاخه ای بهره می بریم و هم سرعت استنتاج تک شاخه ای را داریم.



شکل ۱: بازپارامتری سازی ساختاری: بلوک چندشاخه ای آموزش (چپ) به بلوک تک شاخه ای استنتاج (راست) تبدیل می شود. همه ی شاخه ها در یک کانولوشن 3×3 ادغام می شوند.

بخش اول: ادغام کانولوشن و نرمال سازی دسته ای

برای ورودی $X \in \mathbb{R}^{C_{in} \times H \times W}$ و کرنل $W \in \mathbb{R}^{C_{out} \times C_{in} \times K \times K}$ ، کانولوشن تعریف می شود:

$$(W * X)_{c,i,j} = \sum_{c'=1}^{C_{in}} \sum_{p=0}^{K-1} \sum_{q=0}^{K-1} W_{c,c',p,q} \cdot X_{c',i+p,j+q}$$

نرمال سازی دسته ای (BN) برای بردار ویژگی $y \in \mathbb{R}^C$

$$BN(y; \gamma, \beta, \mu, \sigma^2) = \gamma \odot \frac{y - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta$$

که $\gamma, \beta \in \mathbb{R}^C$ پارامترهای آموخته شده، $\mu, \sigma^2 \in \mathbb{R}^C$ آمارهای اجرایی، و $\epsilon > 0$ ثابت پایداری است.

(آ) ادغام کانولوشن و BN. نشان دهید که ترکیب:

$$y_c = \gamma_c \cdot \frac{(W_c * X + b_c - \mu_c)}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c$$

می تواند به صورت یک کانولوشن معادل نوشته شود. فرمول های دقیق وزن معادل W'_c و بایاس معادل b'_c را بر حسب $W_c, b_c, \gamma_c, \beta_c, \mu_c, \sigma_c^2, \epsilon$ استخراج کنید.

(ب) تبدیل کانولوشن 1×1 به 3×3 . یک کانولوشن 1×1 با وزن $W_{1 \times 1} \in \mathbb{R}^{C_{out} \times C_{in} \times 1 \times 1}$ داده شده. کرنل معادل 3×3 یعنی $\tilde{W} \in \mathbb{R}^{C_{out} \times C_{in} \times 3 \times 3}$ را تعریف کنید. دقیقاً مشخص کنید کدام درایه های $\tilde{W}$ ناصفر هستند و مقادیرشان چیست.

(ج) قضیه ی ادغام شاخه های موازی. ثابت کنید که اگر $W^{(1)}$ و $W^{(2)}$ دو کانولوشن 3×3 به صورت موازی اعمال شوند (ورودی یکسان، خروجی ها جمع می شوند):

$$W^{(fused)} * X = W^{(1)} * X + W^{(2)} * X$$

فرم دقیق $W^{(fused)}$ چیست؟

بخش دوم: نگاشت شناسه به عنوان کانولوشن و ادغام بلوک ResNet

در بخش اول دیدیم که هر شاخه‌ی موازی (از جمله شاخه‌ی شناسه) را می‌توان در یک کانولوشن 3×3 ادغام کرد. اما آیا نگاشت شناسه $\text{id}(X) = X$ همیشه به صورت یک کانولوشن نمایش پذیر است؟ جواب منوط به ابعاد ورودی و خروجی است. برای $C_{\text{in}} = C_{\text{out}} = C$ و $\text{stride} = 1$ ، نگاشت شناسه می‌تواند به صورت یک کانولوشن 3×3 نمایش داده شود (کرنل شناسه). اما اگر این شرط‌ها برقرار نباشند، وضعیت متفاوت می‌شود.

(آ) کرنل شناسه $W_{\text{id}} \in \mathbb{R}^{C \times C \times 3 \times 3}$ را به طور صریح بنویسید. کدام درایه‌ها ناصفر هستند و مقادیرشان چیست؟ نشان دهید که اعمال این کرنل روی ورودی X دقیقاً X را برمی‌گرداند.

(ب) حال دو حالت زیر را جداگانه بررسی کنید:

i. اگر $C_{\text{in}} \neq C_{\text{out}}$: نشان دهید که هیچ کرنل $W \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times 3 \times 3}$ وجود ندارد که $W * X = X$ برای همه‌ی X . راهنمایی: ابعاد ورودی و خروجی کانولوشن را بررسی کنید.

ii. اگر $\text{stride} = s > 1$: نشان دهید که خروجی کانولوشن با $\text{stride} = s$ ابعاد مکانی $\frac{H}{s} \times \frac{W}{s}$ دارد و بنابراین نمی‌تواند برابر $X \in \mathbb{R}^{C \times H \times W}$ باشد.

ResNet در این حالت‌ها از **پروجکشن میان‌بر** (Projection Shortcut) استفاده می‌کند: یک کانولوشن 1×1 با stride مناسب که ابعاد را تطبیق می‌دهد. توضیح دهید که چرا این کانولوشن 1×1 هم‌ارز نگاشت شناسه نیست و چه اطلاعاتی احتمالاً از دست می‌روند.

(ج) برای یک بلوک ResNet با شاخه اصلی W_{main} (شامل دو کانولوشن 3×3 متوالی هر کدام با BN) و اتصال میان‌بر شناسه، فرمول کرنل ادغام‌شده W_{fused} را پس از بازپارامتری‌سازی بنویسید. راهنمایی: ابتدا هر کانولوشن+BN را ادغام کنید، سپس شاخه‌های موازی را با هم جمع کنید.

بخش سوم: تحلیل نظری

(آ) **قضیه‌ی هم‌ارزی**. ثابت کنید که قدرت بازنمایی مدل ادغام‌شده دقیقاً هم‌ارز مدل چندشاخه‌ای است. یعنی برای هر ورودی X :

$$\mathcal{F}_{\text{multi}}(X) = \mathcal{F}_{\text{fused}}(X)$$

چه فرض‌هایی برای این هم‌ارزی لازم است؟

(ب) **تحلیل FLOPs**. برای یک بلوک ResNet با:

• شاخه اصلی: $C \rightarrow C$ با کرنل 3×3

• اتصال میان‌بر: $C \rightarrow C$ با کرنل 1×1 (پروجکشن)

FLOPs کل را قبل و بعد از ادغام برای ابعاد مکانی $H = W = 56$ و $C = 256$ محاسبه کنید. نشان دهید FLOPs تغییر نمی‌کند.

(ج) **تحلیل دسترسی حافظه**. نشان دهید که هزینه دسترسی حافظه برای مدل چندشاخه‌ای تقریباً برابر است با:

$$\text{MAC}_{\text{multi}} \approx 2 \cdot C^2 \cdot H \cdot W + 2 \cdot C \cdot H \cdot W$$

در حالی که برای مدل ادغام‌شده:

$$\text{MAC}_{\text{fused}} \approx C^2 \cdot H \cdot W + 2 \cdot C \cdot H \cdot W$$

چرا این تفاوت برای کارایی سخت‌افزاری مهم است؟ مفهوم **شدت حسابی** (Arithmetic Intensity) را توضیح دهید و نشان دهید چرا مدل ادغام‌شده از نظر سخت‌افزاری کارآمدتر است.

پرسش ۵. الگوریتم پسانتشار و چالش‌های آن برای RNN (۱۵ نمره)

بخش اول

یک نسخه ساده و تک‌بعدی RNN را به شکل زیر فرض کنید

$$h_t = \tanh(w \cdot h_{t-1}), \quad t \in \mathbb{N}$$

که در آن $h_t \in \mathbb{R}$ نشان‌گر حالت پنهانی شبکه و $w \in \mathbb{R}$ پارامتر قابل یادگیری است. همچنین فرض نمایید این شبکه تنها در گام پایانی T -ام خروجی می‌دهد به طوری که

$$\hat{y} = h_T.$$

برای تابع اتلاف نیز تابع مربعات خطا به شکل زیر برای برچسب y^* استفاده می‌شود:

$$L = \frac{1}{2}(h_T - y^*)^2.$$

برای آموزش شبکه‌های RNN از الگوریتم پسانتشار طی زمان (BPTT) استفاده می‌شود؛ در این راستا نشان دهید رابطه گرادیان خطا نسبت به پارامتر مدل، به شکل زیر است:

$$\frac{\partial L}{\partial w} = (h_T - y^*) \sum_{t=1}^T w^{T-t} \cdot (1 - h_t^2) h_{t-1} \cdot \prod_{j=t+1}^T (1 - h_j^2)$$

یادآوری: توجه کنید که رابطه $\frac{d}{dx} \tanh(x) = 1 - \tanh^2(x)$ برقرار است.

بخش دوم

(آ) فرض کنید $\gamma_i := w \cdot (1 - h_i^2)$. بازه مقادیر ممکن برای $|\gamma_i|$ را مشخص نمایید.

(ب) گام مشخص t را در نظر بگیرید و طول دنباله را $T \rightarrow \infty$ فرض کنید. مشخص نمایید تحت چه شرایطی چالش محو شدن گرادیان (Gradient Vanishing) گرادیان رخ می‌دهد؟

(ج) گام مشخص t را در نظر بگیرید و طول دنباله را $T \rightarrow \infty$ فرض کنید. مشخص نمایید تحت چه شرایطی چالش انفجار (exploding) گرادیان رخ می‌دهد؟

پرسش ۶. مقایسه معماری‌های LSTM و GRU (۱۵ نمره)

در درس با دو معماری LSTM و GRU به تفصیل آشنا شده‌اید. در این پرسش قصد داریم این دو معماری را با یکدیگر مقایسه نماییم و نشان دهیم تحت فرضیاتی این دو مدل هم‌ارز خواهند بود. در ادامه برای یادآوری روابط این دو معماری آورده شده است.

مدل LSTM:

$$\begin{aligned}i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (\text{گیت ورودی}) \\f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (\text{گیت فراموشی}) \\o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (\text{گیت خروجی}) \\\tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\h_t &= o_t \odot \tanh(c_t)\end{aligned}$$

مدل GRU:

$$\begin{aligned}z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (\text{گیت بروزرسانی}) \\r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (\text{گیت بازنشانی}) \\\tilde{h}_t &= \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \\h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t\end{aligned}$$

بخش اول

فرض کنید در مدل LSTM، گیت خروجی همواره باز است یعنی $o_t = 1$ و مقدار h_t بدون تابع فعالسازی $\tanh$ محاسبه می‌شود. همچنین در نظر بگیرید گیت‌های ورودی و فراموشی مکمل یکدیگر رفتار می‌کنند یعنی $f_t = 1 - i_t$ است.

(آ) با توجه به فرضیات فوق h_t را به عنوان تابعی از i_t ، x_t و h_{t-1} بازنویسی کنید.

(ب) رابطه بازنویسی شده برای h_t را با مدل GRU مقایسه کنید. شباهت و تفاوت آن دو را توضیح دهید.

بخش دوم

(آ) نقش گیت بازنشانی r_t در مدل GRU را توصیف کنید و برای مدل LSTM تغییری در رابطه $\tilde{c}_t$ بدهید به‌طوری که بتواند توانایی مشابه با گیت بازنشانی مدل GRU پیدا کند.

(ب) با استفاده از نقش گیت‌ها در مدل GRU و مدل LSTM با تغییرات فوق، توضیح دهید هر کدام از فرضیات به چه معنایی هستند و شباهت این دو مدل را بررسی کنید.

پرسش ۷. جریان گرادیان و معماری‌های کمینه RNN (۱۰ نمره)

قبلاً بررسی کردید که چگونه شبکه‌های بازگشتی استاندارد از مشکل محو شدن گرادیان رنج می‌برند و چگونه معماری‌هایی مانند LSTM و GRU با استفاده از مکانیزم‌های گیت‌گذاری این مشکل را کاهش می‌دهند. در این پرسش، قصد داریم جریان گرادیان حالت سلول را در یک LSTM از نظر ریاضی تحلیل کنیم و سپس به بررسی یک نسخه سفارشی و بسیار ساده‌شده از GRU بپردازیم.

بخش اول: چرخ و فلک خطای ثابت در LSTM (Constant Error Carousel)

معادله به‌روزرسانی اصلی برای حالت سلول c_t در LSTM به صورت زیر است:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

که در آن f_t گیت فراموشی، i_t گیت ورودی، و $\tilde{c}_t$ حالت سلول کاندید است.

(الف) در فرمول‌بندی اصلی LSTM، برای جلوگیری از انفجار یا محو شدن گرادیان از طریق مسیرهای پیچیده، خطا بریده (Truncate) می‌شود. فرض کنید گرادیان‌های گیت‌ها (f_t, i_t) و حالت کاندید $(\tilde{c}_t)$ نسبت به c_{t-1} برابر صفر در نظر گرفته شوند (یعنی ما تنها مسیر خطی مستقیم حالت سلول را در نظر می‌گیریم). تحت این فرض، گرادیان محلی $\frac{\partial c_t}{\partial c_{t-1}}$ را به دست آورید.

(ب) با استفاده از نتیجه بخش (الف)، گرادیان کلی $\frac{\partial c_T}{\partial c_t}$ را روی دنباله‌ای به طول T محاسبه کنید. از نظر ریاضی توضیح دهید که چگونه تنظیم بایاس گیت فراموشی b_f روی یک عدد مثبت بزرگ در زمان مقداردهی اولیه، به کاهش مشکل محو شدن گرادیان کمک می‌کند. این رفتار را با یک شبکه استاندارد RNN مقایسه کنید.

بخش دوم: واحد گیت‌دار کمینه

برای کاهش بار محاسباتی معماری GRU (که دارای دو گیت به‌روزرسانی z_t و بازنشانی r_t است)، محققان معماری واحد گیت‌دار کمینه (MGU) را پیشنهاد دادند که گیت‌های بازنشانی و به‌روزرسانی را در یک گیت فراموشی f_t ترکیب می‌کند. MGU به صورت زیر تعریف می‌شود:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ \tilde{h}_t &= \tanh(W_h x_t + U_h (f_t \odot h_{t-1}) + b_h) \\ h_t &= (1 - f_t) \odot h_{t-1} + f_t \odot \tilde{h}_t \end{aligned}$$

(الف) معادلات MGU را با GRU استاندارد مقایسه کنید. به‌طور مشخص، گیت‌های GRU استاندارد (r_t و z_t) را بر حسب گیت f_t در MGU بیان کنید تا نشان دهید چگونه MGU حالت خاصی از معماری GRU است.

(ب) رفتار دقیق MGU را در مقادیر حدی تابع فعال‌ساز سیگموید تعیین کنید:

- **حالت ۱:** اگر برای تمام ابعاد $1 \rightarrow f_t$ میل کند چه اتفاقی می‌افتد؟ در این حالت، MGU دقیقاً رفتار کدام معماری کلاسیک را تقلید می‌کند؟
- **حالت ۲:** اگر برای تمام ابعاد $0 \rightarrow f_t$ میل کند چه اتفاقی می‌افتد؟ کاربرد عملی این حالت برای مدل‌سازی توالی‌ها چیست؟