به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

مدرس: دکتر احمد کلهر

تمرین شماره ۴

خرداد ماه ۱۴۰۵

فهرست مطالب

فهرست مطالب

شبکه‌های عصبی بازگشتی عمیق ۲

بخش اول: سوالات تئوری ۳

Backpropagation Through Time (BPTT) (۲۰ نمره) ۳

سوال دوم: معماری دروازه‌ای (Gating Architecture) (۴۰ نمره) ۴

بخش دوم: سوالات پیاده‌سازی ۶

زیربخش اول: پیش‌پردازش داده‌ها و واژه‌نامه (۵ نمره) ۶

زیربخش دوم: معماری مدل و بارگذاری داده (۱۰ نمره) ۷

زیربخش سوم: منحنی‌های یادگیری و مقایسه عملکرد (۱۵ نمره) ۸

زیربخش چهارم: تشخیص رفتار گرادیان نسبت به گام‌های زمانی (Gradient Diagnostics & BPTT)

(۲۰ نمره) ۸

شبکه‌های عصبی بازگشتی عمیق

این تمرین به بررسی جامع مبانی نظری و کاربردهای عملی شبکه‌های عصبی بازگشتی عمیق در پردازش دنباله‌ها می‌پردازد. در بخش تئوری، مفاهیم پس‌انتشار در زمان (BPTT)، تحلیل طیفی پدیده‌های ناپدید شدن و انفجار گرادیان، کنترل گرادیان با تکنیک برش (clipping)، و معماری‌های دروازه‌دار (LSTM و GRU) مورد بحث و اثبات ریاضی قرار می‌گیرند. در بخش پیاده‌سازی نیز یک سیستم دسته‌بندی متن طراحی شده و پایداری بهینه‌سازی مدل‌ها از طریق تحلیل تجربی جریان گرادیان ارزیابی می‌شود.

بخش اول: سوالات تئوری

Backpropagation Through Time (BPTT) (۲۰ نمره)

یک شبکه عصبی بازگشتی ساده (vanilla RNN) را در نظر بگیرید که روی یک دنباله ورودی $(x_1, x_2, \ldots, x_T)$ عمل می‌کند. (hidden state) $h_t \in \mathbb{R}^d$ در گام زمانی t به صورت زیر تعریف می‌شود:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

فرض کنید L_T یک تابع هزینه اسکالر (scalar loss) است که صرفاً در گام زمانی نهایی T بر اساس خروجی حاصل از h_T محاسبه می‌شود.

- **بخش الف: زنجیره گرادیان (The Gradient Chain)** (۵ نمره)

$\frac{\partial L_T}{\partial W_{hh}}$ را به صورت تحلیلی و صریح به دست آورید. پاسخ شما باید به طور صریح حاصل‌ضرب ژاکوبین‌های زمانی (temporal Jacobians) $\frac{\partial h_j}{\partial h_{j-1}}$ را نشان دهد تا مشخص شود چگونه اطلاعات گذشته (past information) بر حالت فعلی تأثیر می‌گذارد.

- **بخش ب: تحلیل طیفی ماتریس ژاکوبین (Spectral Analysis of the Jacobian)** (۱۰ نمره)

ماتریس ژاکوبین زمانی $J_{j-1} = \frac{\partial h_j}{\partial h_{j-1}}$ را مجزا کنید. اثبات کنید که نرم عملگری ماتریس (matrix operator norm) این ژاکوبین توسط نرم طیفی (spectral norm) یا بزرگ‌ترین مقدار تکین (largest singular value) ماتریس وزن W_{hh} محدود می‌شود. با استفاده از این نتیجه، شرایط دقیق ریاضی را به دست آورید که تحت آن‌ها با میل کردن طول دنباله T به سمت بی‌نهایت، گرادیان به صورت نمایی منفجر (explode) یا ناپدید (vanish) می‌شود.

- **بخش ج: مقابله با انفجار گرادیان** (۵ نمره)

برش نرم گرادیان (gradient norm clipping) را توضیح دهید. چگونه فرآیند برش (clipping)، بردار به‌روزرسانی وزن را هنگامی که نرم L_2 گرادیان از یک آستانه از پیش تعریف‌شده τ فراتر می‌رود، تغییر می‌دهد. توضیح دهید چرا این چارچوب با موفقیت gradient exploding را مهار می‌کند اما در حل مسئله ناپدید شدن گرادیان (vanishing gradient problem) ناتوان است.

سوال دوم: معماری دروازه‌ای (Gating Architecture) (۴۰ نمره)

شبکه‌های حافظه طولانی کوتاه‌مدت (LSTM) و واحدهای بازگشتی دروازه‌دار (GRU) مکانیزم‌های دروازه‌ای داخلی را معرفی می‌کنند تا بر محدودیت‌های ساختاری شبکه‌های عصبی بازگشتی ساده (vanilla RNN) غلبه کنند.

- **بخش الف: حفظ گرادیان در LSTM** (۱۰ نمره)

به صورت ریاضی نشان دهید چرا به‌روزرسانی جمعی (additive update) $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$ در مقایسه با به‌روزرسانی ضربی (multiplicative update) در شبکه‌های بازگشتی ساده (vanilla RNNs)، گرادیان‌ها را در بازه‌های زمانی طولانی حفظ می‌کند.

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

که در آن f_t دروازه فراموشی (forget gate)، i_t دروازه ورودی (input gate)، $\tilde{c}_t$ حالت سلول کاندید (can-didate cell state) و $\odot$ نشان‌دهنده ضرب هادامارد (Hadamard product) یا همان ضرب مؤلفه‌به‌مؤلفه است.

راهنمایی: مشتق جزئی $\frac{\partial c_t}{\partial c_{t-1}}$ را محاسبه کرده و رفتار آن را زمانی که $f_t \approx 1$ است تحلیل کنید.

- **بخش ب: محدودیت‌های ساختاری GRU در مقایسه با LSTM** (۱۰ نمره)

واحد GRU مسیر مجزای (cell state) را به طور کامل حذف می‌کند و تنها یک دنباله حالت پنهان را از طریق فرمول درون‌یابی خطی (linear interpolation) زیر حفظ می‌کند:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

که در آن z_t دروازه به‌روزرسانی (update gate) و $\tilde{h}_t$ حالت پنهان کاندید (candidate hidden state) است.

به صورت ریاضی اثبات کنید که یک GRU را می‌توان به عنوان یک نسخه با محدودیت ساختاری از یک LSTM در نظر گرفت؛ برای این کار، فرضیات دقیق یا کاهش پارامترهای لازم برای نگاشت مستقیم به‌روزرسانی‌های حالت یک سلول LSTM به چارچوب GRU را مشخص کنید. موازنه‌ها (trade-offs) حاصل از این فشرده‌سازی ساختاری را از نظر تعداد پارامترها و ظرفیت حافظه مورد بحث قرار دهید.

- **بخش ج: طراحی دستی LSTM برای محاسبه زوجیت** (۲۰ نمره)

یک شبکه LSTM تک‌بعدی ($d = ۱$) طراحی کنید که زوجیت (parity) یک دنباله بیت ورودی $(x_۱, x_۲, \ldots, x_t)$ را با $x_t \in \{۰, ۱\}$ ردیابی کند. تابع زوجیت p_t در گام زمانی t به صورت زیر تعریف می‌شود که نشان‌دهنده فرد بودن تعداد یک‌ها تا گام زمانی t است:

$$p_t = \left(\sum_{i=۱}^{t} x_i\right) \pmod{۲} \in \{۰, ۱\}$$

مقادیر دقیق وزن‌ها (شامل ماتریس‌های وزن $W_{xf}, W_{xi}, W_{xo}, W_{xc}, W_{hf}, W_{hi}, W_{ho}, W_{hc}$)، بایاس‌ها (b_f, b_i, b_o, b_c) و حالت اولیه $h_۰$ را به صورت دستی به گونه‌ای تعیین کنید که زوجیت دنباله (یعنی p_t) همواره در حالت سلول (یعنی c_t) ذخیره شود (به طوری که $c_t = p_t$). همچنین اثبات ریاضی کوتاهی برای صحت طرح ارائه دهید.

برای تسهیل در محاسبات عددی، فرض کنید توابع فعال‌ساز دروازه‌ها و حالت به صورت ایده‌آل و پله‌ای/بخش‌بند زیر تعریف شده‌اند:

$$\sigma(x) = \begin{cases} ۱ & \text{if} \quad x > ۰ \\ ۰ & \text{otherwise} \end{cases}$$

$$\tanh(x) = \begin{cases} ۱ & \text{if} \quad x > ۰ \\ ۰ & \text{if} \quad x = ۰ \\ -۱ & \text{otherwise} \end{cases}$$

راهنمایی: از هم‌ارزی منطقی عملگر تفاضل متقارن یا همان XOR استفاده کنید:

$$\text{XOR}(a, b) = (a \wedge \neg b) \vee (\neg a \wedge b)$$

چگونه می‌توانید این هم‌ارزی منطقی را با معادله به‌روزرسانی حالت سلول (cell state update) در LSTM یعنی $c_t = f_t \odot c_{t-۱} + i_t \odot \tilde{c}_t$ مرتبط کنید؟

بخش دوم: سوالات پیاده‌سازی

در این بخش، روند پیاده‌سازی پروژه دسته‌بندی متن (Sentiment Analysis) و تحلیل گرادیان‌های مدل‌های بازگشتی قرار داده شده است. شما باید کدهای مربوطه را در فیلترهای مشخص‌شده پیاده‌سازی کرده و گزارش، نمودارها و تحلیل‌های خواسته شده را در گزارش کار خود به طور دقیق ارائه دهید. در هر زیربخش، وظایف پیاده‌سازی گام‌به‌گام به همراه پرسش‌های تحلیلی و الزامات گزارش‌دهی مربوطه به صورت یکجا آورده شده است.

زیربخش اول: پیش‌پردازش داده‌ها و واژه‌نامه (۵ نمره)

در این زیربخش، شما باید فرآیند آماده‌سازی متون خام و تبدیل آن‌ها به داده‌های عددی قابل فهم برای شبکه را طراحی و پیاده‌سازی کنید.

۱. **پیاده‌سازی توکنایزر قانون‌محور فاصله‌گذار:** یک توکنایزر ساده مبتنی بر تفکیک با فاصله (whitespace rules-based tokenizer) بنویسید که ابتدا کل متن را حروف کوچک کرده، تمامی کاراکترهای غیرآلفانومریک و علائم نگارشی را حذف و سپس عبارات را با فاصله تفکیک کند.

◄ **تحلیل و گزارش:** مزایا و محدودیت‌های جدی توکنایزرهای ساده مبتنی بر قوانین فاصله‌گذار را در مقایسه با الگوریتم‌های پیشرفته زیرکلمه‌ای (Subword Tokenizers) مانند BPE یا WordPiece تحلیل کنید (به مواردی نظیر نحوه مواجهه با کلمات خارج از واژه‌نامه، چندریختی کلمات و اندازه بهینه واژه‌نامه اشاره کنید).

۲. **ساخت واژه‌نامه فرکانسی ($V = ۵۰۰۰$):** کلاسی برای واژه‌نامه بسازید که فرکانس کلمات مجموعه آموزش را محاسبه کرده و حداکثر ۵۰۰۰ کلمه پرتکرار را نگه دارد. شناسهٔ 0 را به طور انحصاری به پدینگ (<pad>) و شناسهٔ 1 را به توکن ناشناخته (<unk>) اختصاص دهید و کلمات کم‌تکرار خارج از واژه‌نامه را به شناسهٔ 1 نگاشت کنید.

◄ **تحلیل و گزارش:** گام‌های منطقی ساخت واژه‌نامه و نحوه نگاشت عبارات کم‌تکرار به شناسهٔ ناشناخته و تخصیص شناسه پدینگ را توضیح دهید.

۳. **بارگیری فایل‌ها و تقسیم‌بندی داده‌ها:** کدهای بارگیری فایل‌های متنی rt-polarity.pos و rt-polarity.neg را مستقیماً از پوشهٔ محلی data با رمزگذاری latin-1 پیاده‌سازی کنید. با اعمال بذر تصادفی ثابت (seed=42)، داده‌ها را به صورت کاملاً متوازن (تعداد برابر از نمونه‌های مثبت و منفی در هر بخش) به ۸۵۳۰ نمونه آموزشی،

۱۰۶۶ نمونه اعتبارسنجی و ۱۰۶۶ نمونه تست تقسیم کنید.

◄ **تحلیل و گزارش:** میانگین و انحراف معیار طول جملات (تعداد توکن‌ها) را در هر سه مجموعهٔ آموزش، اعتبارسنجی و تست محاسبه و گزارش کنید تا از همگونی و توازن توزیع طول در تقسیم‌بندی داده‌ها اطمینان حاصل شود.

زیربخش دوم: معماری مدل و بارگذاری داده (۱۰ نمره)

در این زیربخش، نحوه مدیریت بهینه دسته‌های متغیر و طراحی معماری مدل‌های بازگشتی را در کدهای خود اجرا می‌کنید.

۱. **پیاده‌سازی تابع پدینگ پویا و آماده‌سازی دسته‌ها:** یک تابع پدینگ پویا سفارشی (custom collate_fn) برای پد کردن پویا دنباله‌های هر دسته (batch) بنویسید و از فشرده‌سازی دنباله‌ها (sequence packing) برای پردازش بهینه استفاده کنید.

◄ **تحلیل و گزارش:** توضیح دهید چرا پدینگ پویا در سطح دسته (dynamic batch-level padding) نسبت به پدینگ ثابت تا حداکثر طول کل داده‌ها (static global padding) از نظر فضا و زمان محاسباتی کارآمدتر است. همچنین نحوه عملکرد درونی pack_padded_sequence و اهمیت آن در جلوگیری از محاسبه گرادیان و پردازش غیرضروری روی صفرهای پدینگ را تشریح کنید.

► **پاسخ:** در پدینگ ثابت، همه جملات تا طول بلندترین جمله کل دیتاست پد می‌شوند (به‌طور میانگین اتلاف ٪۶۰ ≈ از محاسبات برای دیتاست ما با میانگین ۵.۱۸ توکن). در پدینگ پویا، طول پد تنها محدود به بلندترین جمله هر دسته می‌شود که بسیار کوتاه‌تر است. تابع pack_padded_sequence دنباله‌های پدشده را به یک ساختار فشرده تبدیل می‌کند (PackedSequence) که سلول بازگشتی فقط توکن‌های واقعی را پردازش می‌کند و گرادیان روی صفرهای پدینگ محاسبه نمی‌شود؛ این هم کارایی را بالا می‌برد و هم از آلودگی گرادیان جلوگیری می‌کند.

۲. **طراحی معماری مدل‌های بازگشتی:** مدل‌های بازگشتی (Vanilla RNN، LSTM و GRU) را برای دسته‌بندی متن پیاده‌سازی کنید. در پیاده‌سازی خود، بعد فضای embedding را برابر $D = ۶۴$ (با اعمال padding_idx=0)، بعد لایه پنهان را برابر $H = ۱۲۸$ تنظیم کرده و برای طبقه‌بندی نهایی، حالت پنهان متناظر با آخرین کلمه واقعی دنباله (با نادیده گرفتن مقادیر پد شده) را استخراج و به لایهٔ خروجی ارسال کنید.

◄ **تحلیل و گزارش:** نحوه استخراج حالت پنهان آخرین کلمه واقعی دنباله را در هر سه معماری (به‌ویژه با توجه به تفاوت در خروجی مدل LSTM که دارای زوج حالت پنهان و حالت سلول c_n است) بنویسید و

علت خطای ناشی از برداشت ساده حالت پنهان نهایی بدون در نظر گرفتن طول واقعی دنباله را شرح دهید.

زیربخش سوم: منحنی‌های یادگیری و مقایسه عملکرد (۱۵ نمره)

در این زیربخش، به آموزش مدل‌ها پرداخته و نتایج تجربی حاصله را برای تحلیل رفتار تعمیم‌پذیری ارائه می‌دهید.

۱. **آموزش مدل‌ها و استخراج آمارها:** هر سه مدل Vanilla RNN، LSTM و GRU را با استفاده از هایپرپارامترهای مناسب آموزش دهید. زمان کل آموزش هر مدل (به ثانیه)، تعداد پارامترهای قابل آموزب و دقت نهایی آن‌ها را روی مجموعه‌های اعتبارسنجی و تست اندازه‌گیری و ثبت کنید.

◄ **تحلیل و گزارش:** جدولی ارائه دهید که برای هر سه مدل شامل تعداد پارامترها، زمان آموزش (ثانیه)، دقت اعتبارسنجی و دقت تست باشد. با تکیه بر آمارها، سرعت آموزش و پویایی همگرایی سه مدل را تحلیل کنید. چرا سرعت محاسباتی Vanilla RNN بیشتر است اما دقت نهایی کمتری دارد؟

۲. **رسم منحنی‌های آموزش و اعتبارسنجی:** یک شکل دوپنلی (1x2 subplots) رسم کنید. پنل سمت چپ مربوط به منحنی خطای آموزش و اعتبارسنجی (Loss) و پنل سمت راست مربوط به منحنی دقت آموزش و اعتبارسنجی (Accuracy) برای هر سه مدل باشد.

◄ **تحلیل و گزارش:** نمودارهای دوتایی هم‌جوار یادگیری را با رعایت کامل راهنماها و برچسب‌های محورها در گزارش خود قرار دهید. وضعیت بیش‌برازش ملایم یا شدید را در هر سه مدل بررسی کنید. کدام مدل زودتر یا بیشتر بیش‌برازش می‌شود و چرا؟ این رفتار را با ظرفیت محاسباتی و تعداد پارامترهای مدل‌ها توجیه کنید.

زیربخش چهارم: تشخیص رفتار گرادیان نسبت به گام‌های زمانی (Gradient Diagnostics & BPTT) (۲۰ نمره)

در این زیربخش، به تحلیل تجربی جریان گرادیان در گام‌های زمانی در BPTT پرداخته و زوال یا ثبات گرادیان را به چالش می‌کشید.

۱. **استخراج گرادیان فضای تعبیه‌ساز در طول زمان:** دسته‌ای شامل حداقل ۳۲ نمونه از داده‌های اعتبارسنجی که طول آن‌ها حداقل ۳۰ کلمه است جدا کرده و طول همگی را دقیقاً به $T = ۳۰$ توکن اول محدود کنید. نمونه‌ها را به لایه تعبیه‌ساز داده و خروجی را با دستور embedded.clone().detach().requires_grad_(True)

برای ثبت گرادیان آماده کنید. بردار تعبیه‌ساز ورودی را به مدل فرستاده، خطا را محاسبه کنید و با فراخوانی پس‌انتشار از طریق دستور loss.backward()، گرادیان خطا نسبت به بردارهای تعبیه‌ساز ورودی یعنی $g_t = \frac{\partial \mathcal{L}}{\partial e_t}$ را برای هر گام زمانی $t \in [۱, ۳۰]$ استخراج کنید.

◄ **تحلیل و گزارش:** توضیح دهید که چگونه محاسبه گرادیان نسبت به فضای تعبیه‌ساز ورودی ($e_t \in \mathbb{R}^{۶۴}$) در طول گام‌های زمانی، جریان واقعی گرادیان و پدیده زوال یا ثبات آن را در فرآیند پس‌انتشار در زمان (BPTT) نشان می‌دهد و چرا این تحلیل برای مقایسه مستقیم رفتار دو مدل Vanilla RNN و LSTM کارآمد است.

۲. **محاسبه نرم گرادیان و رسم نمودار جریان:** میانگین نرم $L_۲$ این گرادیان‌ها را در طول گام‌های زمانی بر روی کل دسته برای دو مدل Vanilla RNN و LSTM محاسبه کرده و آن‌ها را در قالب یک نمودار خطی واحد با عنوان gradient_diagnostics.png رسم نمایید.

◄ **تحلیل و گزارش:** نمودار خطی حاصل (gradient_diagnostics.png) را در گزارش کار خود بگنجانید. با ترکیب تحلیل‌های ریاضی بخش تئوری و رفتار تجربی در شکل، فروپاشی نمایی گرادیان در Vanilla RNN به سمت گام‌های اولیه را اثبات کنید. همچنین با فرمول‌بندی جریان خطا در LSTM، توضیح دهید چگونه گیت فراموشی کاملاً باز و مکانیسم Constant Error Carousel (CEC) مانع از افت شدید گرادیان تا گام نخست ($t = ۱$) می‌شوند. نسبت تجربی گرادیان ($g_۱/g_{۳۰}$) را برای هر دو مدل محاسبه و گزارش نموده و تفسیر آن را در یادگیری وابستگی‌های بلندمدت بنویسید.

نکات تحویل

- مهلت ارسال این تمرین تا پایان روز "جمعه ۱۵ خرداد ماه" خواهد بود.
- این زمان قابل تمدید نیست و درصورت نیاز می‌توانید از گریس استفاده کنید.
- در نظر داشته باشید که حداکثر مهلت آپلود تمرین در سامانه تا ۷ روز پس مهلت تحویل است و پس از آن سامانه بسته خواهد شد.
- پیاده سازی با زبان برنامه نویسی پایتون باید باشد و کدهای شما باید قابل اجرا بوده و به همراه گزارش آپلود شوند.
- انجام این تمرین به صورت یک نفره می‌باشد.
- در صورت مشاهده هر گونه تشابه در گزارش کار یا کدهای پیاده‌سازی، این امر به منزله تقلب برای طرفین در نظر گرفته خواهد شد.
- استفاده از کدهای آماده بدون ذکر منبع و بدون تغییر به منزله تقلب خواهد بود و نمره تمرین شما **صفر** در نظر گرفته می‌شود.
- در صورت رعایت نکردن فرمت گزارش کار نمره گزارش به شما تعلق نخواهد گرفت.
- تحویل تمرین به صورت **دستنویس** قابل پذیرش نیست.
- تمامی تصاویر و جداول مورد استفاده در گزارش کار باید دارای توضیح (caption) و شماره باشند.
- بخش زیادی از نمره شما مربوط به گزارش کار و روند حل مسئله است.
- لطفا گزارش، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه بارگذاری نمائید.
  - HW4_[Lastname]_[StudentNumber].zip
- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با موضوع NNDL_HW4 با دستیاران آموزشی در ارتباط باشید:
  - دستیار تمرین

amirh.bonakdar@ut.ac.ir

با آرزوی سلامتی و موفقیت روزافزون