دانشکده مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۰-۹۹

مدرس: حمید بیگی

تمرین سری سوم | **یادگیری عمیق** | زمان تحویل: ۱۸ اردیبهشت

نکات زیر را رعایت کنید:

فایل گزارش را به همراه تمامی کدها در یک فایل فشرده و با عنوان #HW3_STD در سایت Quera.ir بارگذاری نمایید.

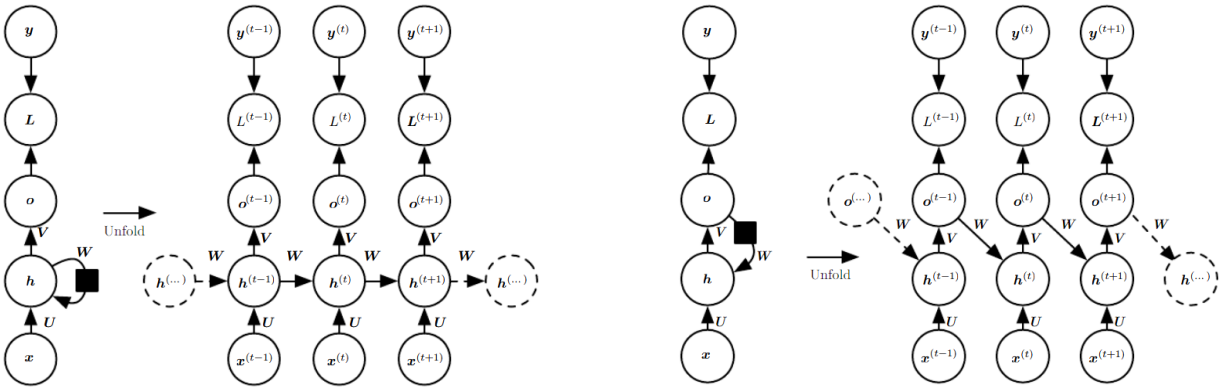
بخش‌های پیاده سازی مربوط به هر سوال را در فایل مربوطه با شماره‌ی آن سوال و در پوشه‌ای برای آن سوال قرار دهید.

سوالات خود را از طریق Piazza مطرح کنید.

سوالات تئوری

مسئله۱. (۲۰ نمره)

در یک شبکه بازگشتی می‌توانیم به جای اتصالات از $h(t-1)$ به $h(t)$ ، از اتصال $o(t-1)$ به $h(t)$ استفاده کنیم. این دو پیکربندی بازگشتی در شکل ۱ نشان داده شده است. در این شکل در هر گام شبکه بازگشتی، $x(t)$ ورودی، $h(t)$ بردار نهان، $o(t)$ خروجی شبکه و $y(t)$ برچسب های مورد انتظار هستند. تابع $L(t)$ نیز تابع هزینه‌ی شبکه بازگشتی در گام t است که با هدف نزدیک کردن مقادیر $o(t)$ به $y(t)$ بهینه می‌شود.



(ب) اتصالات از بردار نهان به بردار نهان

(آ) اتصالات از خروجی به بردار نهان

شکل۱: دو نوع پیکربندی در شبکه‌های بازگشتی

الف. استفاده کردن از پیکربندی شکل۱آ چه مزیتی نسبت به پیکربندی شکل۱ب می‌تواند داشته باشد؟ این دو پیکربندی را از نظر قدرت مدلسازی نیز با هم مقایسه کنید.

ب. برای یادگرفتن توزیع یک دنباله، مدل‌هایی مشابه شکل۱ را می‌توان با استفاده از تکنیک **teacher forcing** آموزش داد. یکی از مشکلات شناخته شده این روش **exposure bias** است. تکنیک **teacher forcing** و مشکل **exposure bias** را توضیح دهید.(برای اطلاعات بیشتر در مورد تکنیک معرفی شده می‌توانید به آدرس **[۱]** در قسمت منابع مراجعه کنید.)

ج. یکی از راه‌حل هایی که برای رفع مشکل **exposure bias** ارائه شده، روش **schedule sampling** است. این روش را مختصرا توضیح دهید.

د. مسئله‌ی **gradient vanishing/explosion** را توضیح دهید. این مسئله برای آموزش شبکه بازگشتی چه مشکلی ایجاد می‌کند؟

ه. راه‌حل ساده و در عین حال موثر برای حل مشکل **gradient explosion** روش **norm clipping** و برای حل مشکل **gradient vanishing** روش **soft constraint** است. این دو روش را توضیح دهید.

مسئله ۲. (۲۰ نمره)

مکانیزم توجه[۱] برای از بین بردن گلوگاه اطلاعات[۲] بین دیکودر و انکودر معرفی شده است. به این صورت که به جای آخرین بردار نهان انکودر، دیکودر به تمامی بردارهای نهان انکودر دسترسی دارد. این مکانیزم به صورت زیر فرموله می‌شود و در هر گام شبکه‌ی تکرار شونده‌ی دیکودر مورد استفاده قرار می‌گیرد:

$$a_t(s) = \frac{\exp \text{score}(h_d^{(t)}, h_e^{(s)})}{\sum_{s'} \exp \text{score}(h_d^{(t)}, h_e^{(s')})} \quad (۱)$$

$$c_t = \sum_{s'} a_t(s') h_e^{(s')} \quad (۲)$$

$$\hat{h} = \tanh W_c[c_t; h_d^{(t)}] \quad (۳)$$

$$y_t = \text{softmax } \hat{W_s} \hat{h} \quad (۴)$$

که در آن $\boldsymbol{h_d^{(i)}}$ بردار نهان دیکودر، $\boldsymbol{h_e^{(j)}}$ بردار نهان انکودر و $\boldsymbol{y_t}$ خروجی گام **t** ام دیکودر می‌باشد. تابع $\textbf{score}(\boldsymbol{h_d^{(t)}}, \boldsymbol{h_e^{(s)}})$ را می‌توان به سه روش زیر تعریف کرد:

$$\text{score}(h_d^{(t)}, h_e^{(s)}) = \begin{cases} \mathbf{h_d^{(t)\top} h_e^{(s)}} & dot \\ \mathbf{h_d^{(t)\top} W_a h_e^{(s)}} & general \\ v_a^\top \tanh \mathbf{W_a}[\mathbf{h_d^{(t)}}; \mathbf{h_e^{(s)}}] & tanh\ layer \end{cases}$$

[۱] Attention Mechanism
[۲] Information Bottleneck

الف. این سه تابع را از نظر توان مدل کردن، هزینه‌ی محاسباتی و عبور گرادیان در مرحله بازانتشار خطا مقایسه کنید. شما کدام یک را برای یک شبکه Seq2Seq انتخاب می‌کنید؟

ب. در ادبیات یادگیری عمیق، دو نوع مکانیزم توجه رایج هستند که توسط دو فرد مختلف معرفی شده اند: ۱- Bahdanau et. al. [۲] و ۲- Luong et. al. [۳] این دو ساختار را با هم مقایسه کنید و تفاوت های آن را ذکر کنید. کدام یک توانایی مدل کردن بیشتری دارد؟

ج. یکی از مشکلات رایج مکانیزم توجه، مخصوصا هنگامی که متن ورودی در طرف انکودر طولانی باشد، عدم توانایی این مکانیزم در پرداختن به تکه های مختلف متن ورودی است. به طور مثال ممکن است در تمامی گام های دیکودر، مکانیزم توجه فقط به یک یا دو کلمه ی خاص امتیاز بسیار بالایی بدهد و فقط آن ها را در نظر بگیرد. در این صورت مدل قادر نخواهد بود که از تمامی متن ورودی استفاده کند. برای حل این مشکل چه راهکاری پیشنهاد میدهید؟ توضیح دهید.

(راهنمایی: شما می‌توانید یک جمله ی جدید به تابع خطا/هزینه ی مدل اضافه کنید)

د. توجه سخت[۳] و توجه نرم[۴] را با هم مقایسه کنید و بگویید کدام یک را می‌توان با استفاده از روش پس انتشار خطا آموزش داد؟ چرا؟

مسئله ۳. (۱۰ نمره)

با در نظر گرفتن روش skip-gram برای word2vec به سوالات زیر جواب کوتاه دهید.

در این روش کلمه ای وارد شبکه شده و انتظار می‌رود که کلمات به سوالات زیر جواب کوتاه دهید.

الف. در این روش کلمه‌ای وارد شبکه شده و انتظار می‌رود که کلمات زمینه[۵] آن کلمه پیش‌بینی شود. اگر آموزش را به صورتی انجام دهیم که به جای کلمات زمینه، خود کلمه ی ورودی پیش‌بینی شود (مشابه Autoencoder)، چه مشکلی به وجود می آید؟

ب. در این روش دو ماتریس وجود دارد که در نهایت تعبیه[۶] کلمات به کمک این دو ماتریس به دست می آیند. ماتریس اول، ماتریسی است که کلمه ی ورودی را به فضای ویژگی می‌برد و دلیل اینکه مقادیر نهایی تعبیه برای کلمات باشد، مشخص است. توضیح دهید که چرا ماتریس دیگر نیز یک تعبیه برای کلمات است؟

ج. در سوال قبل توضیح داده شد که براساس دو ماتریس مذکور، دو تعبیه برای کلمات به دست می آید. چرا در این ساختار از دو ماتریس مختلف استفاده شده و اگر این دو ماتریس مقادیر اشتراکی[۷] داشته باشند، چه مشکلی به وجود می آید؟

سوالات عملی

[۳] Hard attention
[۴] Soft attention
[۵] Context
[۶] Embedding
[۷] Shared

مسئله۴. تعبیه کلمات (۱۰ نمره)

Word2vec را با رویکرد CBOW[۸]پیاده سازی کنید و استفاده از دادگان نظرات کاربران در وب سایت IMDB که مجموعه داده آن در torchtext.datasets موجود است آموزش دهید. ابعاد بردارها را ۱۰۰ و اندازه پنجره را ۵ در نظر بگیرید. برای آموزش از هر دو مجموعه داده آموزش و تست استفاده کنید.

مسئله۵. تحلیل احساسات (۲۰ نمره)

تحلیل احساسات[۹] فرآیند تشخیص مثبت یا منفی بودن حس نویسنده در دادهی متنی است. این تحلیل معمولا توسط کسب و کارها برای تشخیص احساسات کاربران در شبکه‌های اجتماعی، شناخت مشتری‌ها و ارزیابی شهرت برندها استفاده می‌شود.

در این مسئله از شما خواسته شده است که چند مدل شبکه عصبی بازگشتی با استفاده از دادگان نظرات کاربران در وب سایت IMDB که مجموعه داده آن در torchtext.datasets موجود است، آموزش دهید. فرآیند پیاده سازی و آموزش را با استفاده شبکهی LSTM انجام دهید. ابعاد تعبیه کلمات را ۱۰۰ و تعداد لایه‌ها را ۲ در نظر بگیرید. برای آموزش داده‌ها را در دسته‌های ۶۴ تایی دسته بندی کنید و فرآیند آموزش را ۱۰ دوره[۱۰] تکرار کنید. آموزش را یکبار با استفاده از لایه تعبیه آموزش داده شده در مسئله ۴ و یکبار با استفاده از یک تعبیه از پیش آموزش داده شده با استفاده از روش GloVe [۱۱]که روی ۶ میلیارد توکن آموزش داده شده است و ابعاد بردارهای تعبیه برابر ۱۰۰ است انجام دهید. در ۵ دور اول آموزش مدل، وزن‌های لایه تعبیه آموزش داده نشود(اصطلاحا freeze شود) و در ۵ دوره بعد وزن های لایه تعبیه مورد آموزش قرار گیرد. در نهایت مدل آموزش داده شده با استفاده از داده‌های تست ارزیابی گشته، میزان خطا و دقت در گزارش مربوطه ذکر گردد.

مسئله۶. سری‌های زمانی (۲۰ نمره)

شبکه‌های عصبی یکی از مدل‌های پرطرفدار جهت پیش بینی سری‌های زمانی[۱۲] هستند. در این مسئله از شما خواسته شده است با استفاده از مجموعه دادگانی که در اختیار شما قرار گرفته است مقدار آینده سری زمانی را پیش بینی کنید. مجموعه دادگان این مسئله شامل داده‌های مصرف انرژی است در مناطق مختلف است که به صورت ساعتی جمع آوری شده است. ابتدا پیش پردازش های لازم را بر روی داده‌ها انجام داده، سپس ۹۰ درصد داده‌ها را برای فرآیند آموزش و ۱۰ درصد باقی را برای تست نگه دارید. برای در فرآیند آموزش و تست، برای پیش بینی مقدار بعدی سری زمانی، داده‌های ۹۰ قدم قبلی را در نظر بگیرید. دو مدل LSTM و GRU را برای این مسئله پیاده سازی کنید. داده‌ها را در دسته‌های ۱۰۲۴ تایی دسته بندی کنید و مدل‌ها را برای ۵ دوره بر روی داده‌ها آموزش دهید. مدل‌های آموزش داده شده را با استفاده از معیار sMAPE[۱۳] ارزیابی کنید.

$$sMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|F_t| + |A_t|)/2}$$

موفق باشید.

---

[۸] Continuous Bag of Words
[۹] Sentiment analysis
[۱۰] Epoch
[۱۱] Global Vectors for Word Representation
[۱۲] Time Series
[۱۳] symmetric Mean Absolute Percentage Error

منابع

[1] https://machinelearningmastery.com/teacher-forcing-for-recurrent-neural-networks

[2] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

[3] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (201۵).