



یادگیری عمیق

پاییز ۱۴۰۲

استاد: دکتر فاطمی زاده

گردآورندگان: ارشاک رضوانی، محمدجواد محمدی و پرهام محمدی

مهلت ارسال: چهارشنبه ۲۰ دی

VAE & RNN

تمرین چهارم

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۶ روز و در مجموع ۲۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- همکاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ‌های ارسال‌شده باید توسط خود او نوشته شده باشد. (دقت کنید در صورت تشخیص مشابهت غیرعادی برخورد جدی صورت خواهد گرفت.)
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- نتایج و پاسخ‌های خود را در یک فایل با فرمت zip به نام HW۴-Name-StudentNumber در سایت [Quera](#) قرار دهید. برای بخش عملی تمرین نیز لینک گیت‌هاب که تمرین و نتایج را در آن آپلود کرده‌اید قرار دهید. دقت کنید هر سه فایل نوت‌بوک تکمیل شده بخش عملی را در گیت‌هاب قرار دهید.
- لطفا تمامی سوالات خود را از طریق کوثرای درس مطرح بکنید (برای اینکه تمامی دانشجویان به پاسخ‌های مطرح شده به سوالات دسترسی داشته باشند و جلوی سوالات تکراری گرفته شود، به سوالات در بسترهای دیگر پاسخ داده نخواهد شد).
- دقت کنید کدهای شما باید قابلیت اجرای دوباره داشته باشند، در صورت دادن خطا هنگام اجرای کدتان، حتی اگر خطا بدلیل اشتباه تایپی باشد، نمره صفر به آن بخش تعلق خواهد گرفت.

سوالات عملی (۱۰۰۰ نمره)

۱. (۲۸۰ نمره) در این مسئله قصد داریم تا به کمک مدل‌های Sequential یک مسئله‌ی Binary Classification را حل کنیم. این مسئله مربوط به داده‌های زیستی است و یک نوع مسئله‌ی Molecular Property Prediction محسوب می‌گردد. پیشینی کردن رفتار احتمالی ترکیبات شیمیایی یکی از مسائل اساسی در حوزه‌ی زیستی است. در ادبیات این حوزه به صورت گسترده از روش‌های مبتنی بر یادگیری عمیق استفاده می‌شود. یک دیتاست در اختیار شما قرار گرفته است که نام آن BBBP است. نام این دیتاست مخفف شده‌ی blood-brain barrier penetration است. این دیتاست شامل ۲۰۵۳ نمونه است. هر نمونه یک مولکول شیمیایی است که فرمول آن توسط فرمت SMILES داده شده است. برای هر نمونه مشخص شده است که جزو کلاس نفوذپذیر یا نفوذناپذیر است. کلاس هر نمونه در ستون p_np با مقادیر ۰ یا ۱ مشخص شده است. در این مسئله قدم به قدم تلاش می‌کنید تا یک Representation از مولکول‌ها به کمک مدل‌های RNN بسازید. با این کار هر مولکول را با یک بردار توصیف خواهد شد که اطلاعات مولکول را در خود کد کرده است. سپس از آن برای پیشینی کردن کلاس مولکول‌ها استفاده کنید.

(آ) در قدم اول سعی کنید ویژگی‌ها و جنس داده‌ای که در اختیار دارید را به دقت بررسی کنید تا بتوانید حدس اولیه‌ای از چالش‌های آن داشته باشید. در بخش اول باید با فرمت SMILES آشنا شوید و در

گزارش توضیح دهید که طول رشته‌ی مولکول‌هایی که در این دیتاست وجود دارد چه طیفی دارد و توزیع آن را تحلیل کنید. تنوع token های این رشته‌ها و فرکانس رخداد آنها را نیز بدست آورده و در گزارش بیاورید.

(ب) با استفاده از one-hot-encoder کاراکترهای رشته‌ی SMILES را کد کنید. بنابراین هر نمونه به یک بردار عددی تبدیل میشود. سپس یک شبکه‌ی عصبی FC روی آن آموزش دهید و دقت آن را بر روی داده‌ی تست ارزیابی کنید. از این دقت به عنوان baseline برای مقایسه با دقت بدست آمده در بخش‌های بعدی استفاده کنید.

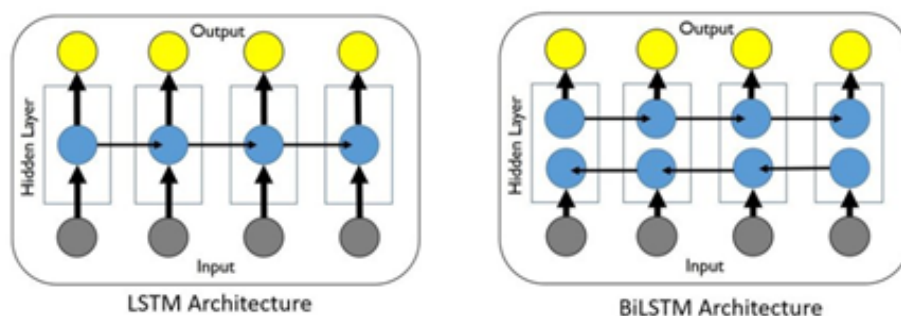
(ج) یک شبکه‌ی LSTM پیاده‌سازی کنید که رشته‌ی مولکول‌ها به آن داده میشود. خروجی آن را flatten کنید و با چند لایه‌ی fully connected کلاس نهایی را پیشبینی کنید. سعی کنید تا حدی پارامترهای مختلف را بهینه کنید تا عملکرد قابل قبولی بدست آورید.

(د) در این مرحله در مدل قبلی خود، شبکه‌ی LSTM را با BiLSTM جایگزین کنید و بدون تغییر پارامترها آموزش و ارزیابی را مجدداً انجام دهید.

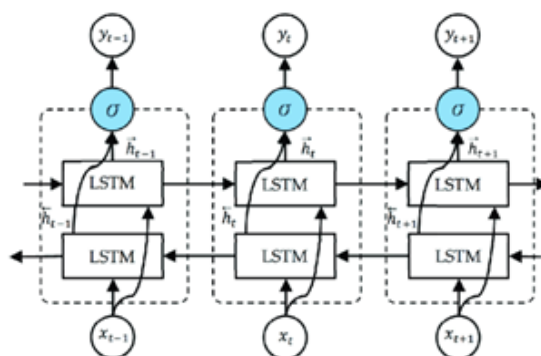
(ه) برای اینکه دقت را با اطمینان قابل قبولی بدست آورید، برای هریک از قسمت‌های ۲ و ۳ از Cross Validation استفاده کنید و دقت بدست آمده از آن را به عنوان دقت نهایی هر بخش گزارش دهید.

(و) با استفاده از ساختار شبکه‌های RNN و Bidirectional RNN و جنس داده علت تفاوت دقت بدست آمده بین این دو شبکه را تحلیل کنید و شرح دهید.

(ز) داده‌های تست را بر اساس طول رشته‌ی SMILES متناظر به ۱۰ bin افزایش دهید. نتایج شبکه‌های Sequential را با Baseline که در قسمت ۲ بدست آوردید روی هریک از این bin ها مقایسه کنید و علت شهودی روند بهبود یا تضعیف عملکرد مدل‌های Sequential روی این bin ها را توضیح دهید.



شکل ۱: معماری LSTM و BiLSTM

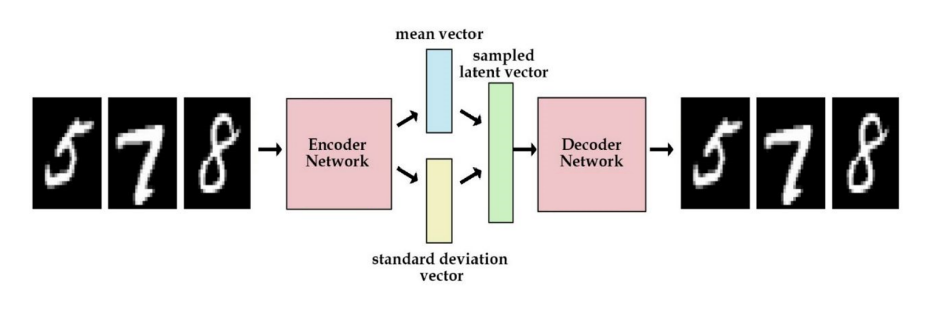


شکل ۲: معماری BiLSTM با جزئیات بیشتر

۲. (۲۸۰ نمره) لطفاً به فایل نوتبوک با نام `VAE_CVAE.ipynb` مراجعه نمایید. در این تمرین عملی، هدف آشنایی با مدل‌های VAE و CVAE مبتنی بر MLP می‌باشد. پس از طراحی معماری و آموزش شبکه ساده خودتان در هر یک از این حالات، شباهت‌ها و تفاوت‌های آنها را ملاحظه خواهید کرد. همچنین با ترسیم توزیع احتمال فضای نهان به ازای تصاویر از کلاس‌های مختلف، دید بهتری به مدل‌های VAE و CVAE خواهید داشت.

برای این منظور، شما مدل VAE را یکبار با استفاده از لایه‌های MLP پیاده‌سازی می‌نمائید تا بتوانید تصاویری از نوع دادگان ارقام MNIST را تولید نمائید. سپس همین کار را برای مدل CVAE انجام می‌دهید و با استفاده از لایه‌های MLP بر اساس همان دادگان ارقام MNIST پیاده‌سازی نمائید.

در نهایت، برای مدل‌های VAE و CVAE از نوع MLP، توزیع احتمال لایه نهان‌شان را به ازای تصاویری از کلاس‌های مختلف (۰ تا ۹) رسم کنید.



شکل ۳: پیاده‌سازی مدل‌های VAE و CVAE برای دیتاست MNIST

نکات مهم در انجام این تمرین:

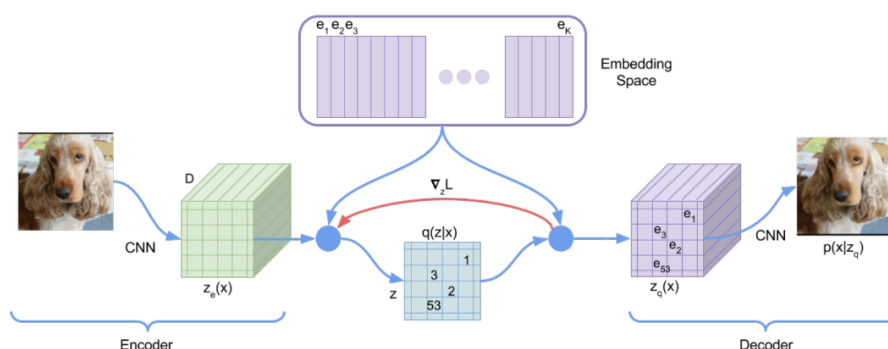
- (آ) برای مدل‌های VAE و CVAE از نوع MLP، لایه نهان را دو بعدی در نظر بگیرید تا نمایش توزیع احتمال آن بر روی صفحه مقذور باشد.
- (ب) تعداد لایه‌ها در هر یک از قسمت‌های encoder و decoder برای هر کدام از مدل‌های VAE و CVAE حداقل ۲ لایه و حداکثر ۴ لایه می‌تواند باشد.

۳. (۱۶۰ نمره) برای این سوال کلیه اشعار حکیم ابولقاسم فردوسی در اختیار شما قرار گرفته‌است. هدف این سوال پیاده‌سازی یک مدل sequence to sequence با استفاده از fine tuning یک مدل pre-trained برای تولید شعر در سبک فردوسی است! برای انجام این تسک می‌توانید از مدل GPT2 که روی داده‌های زبان فارسی آموزش دیده است، مانند [HooshvareLab/gpt2-fa](https://huggingface.co/HooshvareLab/gpt2-fa) از کتابخانه Transformers هاگینگ فیس استفاده کنید.

- (آ) ابتدا مجموعه دادگان را پردازش کنید و دادگان مناسب برای آموزش شبکه مورد نظر را به دست آورید. برای آموزش بهتر مدل، دادگان داده‌شده را به بیت‌های مجزا تقسیم کنید و یک بیت را برای ورودی در نظر بگیرید و بیت پس از آن را خروجی قرار دهید. ست‌های ترین و تست را نیز جدا کنید.
- (ب) داده‌های مجموعه شعر فارسی خود را با استفاده از tokenizer ای که قادر به پردازش زبان فارسی است، tokenize کنید، مدل pre-trained ذکر شده دارای tokenizer مناسب نیز است. دیتا لودرهای ترین و تست را نیز بسازید.
- (ج) مدل pre-trained را لود کنید. حلقه آموزش و تابع زیان مانند (crossentropy) را تنظیم کنید. مدل را روی مجموعه داده‌های شعر خود آموزش دهید.
- (د) از مدل آموزش‌دیده برای تولید جمله جفت‌شده با توجه به یک جمله ورودی استفاده کنید و تعدادی از اشعار خروجی را گزارش کنید. همچنین عملکرد مدل را روی ست تست نیز بسنجید. می‌توانید هر یک از معیارهای Perplexity یا BLEU را استفاده کنید.

۴. (۲۸۰ نمره) در این تمرین از شما می‌خواهیم که یک نوع از خودرمزگذارها را به نام Vector Quantized Variational Autoencoder یا به اختصار VQ-VAE را پیاده‌سازی کنید. در خودرمزگذارهای عادی صرفاً سعی می‌کردیم یک بردار به داده ورودی نسبت دهیم و هیچ مفهوم احتمالاتی در فضای نهان آموزش‌دیده‌شده وجود نداشت و خط واصل بین دو نقطه در فضای نهان کاملاً بی‌معنا بود. سپس در درس دیدیم که Variational Autoencoder ها این مشکل را حل می‌کنند و یک توزیع احتمال پیوسته روی فضای نهان یاد می‌گیرند که به ما اجازه حرکت در داخل این فضای نهان را می‌دهند و تمامی نقاط داخل فضا بعد از بازسازی، معنادار هستند. حال در بسیاری از مسائل همچنان دنبال یک فضای نهان گسسته و در عین حال معنادار هستیم و برای این منظور VQ-VAE ساخته‌شده است. مثلاً در تسک تولید تصاویر سگ و گربه از نژادهای مختلف تمامی نقاط یک فضای نهان پیوسته معنا درستی ندارند و نیاز است که به صورت گسسته در این فضا حرکت کنیم (موجودی که بین سگ و گربه باشد وجود ندارد!) و در عین حال بتوانیم داده جدید تولید کنیم. شما در این تمرین باید مراحل زیر را قدم به قدم انجام داده و خروجی‌های خواسته شده را تولید کنید.

(آ) برای یادگیری جزئیات این مدل و پیاده‌سازی آن باید مقاله اصلی این روش را از این [لینک](#) مطالعه کرده و توضیح دهید که VQ-VAE چه برتری نسبت به VAE عادی دارد. شکل زیر نمای کلی این مدل را نشان می‌دهد:



شکل ۴: نمای کلی مدل VQ-VAE

(ب) شما باید مدل خود را روی دیتاست MNIST پیاده کنید و موارد خواسته شده را که در ادامه گفته شده است، انجام دهید. پس یک قطعه کد برای لود کردن این دیتاست و استفاده از آن را بنویسید، می‌توانید از کد سوال قبلی نیز استفاده کنید.

(ج) اجزای اصلی مدل را پیاده‌سازی کنید:

- انکودر (Encoder)
- دیکودر (Decoder)
- کدبوک (Code Book)

برای پیاده‌سازی انکودر و دیکودر در مدل VQ-VAE روی داده تصویر باید از لایه‌های کانولوشنال استفاده کنید و تصاویر ورودی را flatten نکنید! توجه کنید که در مقاله گفته شده است که در فرآیند یادگیری، گرادیان‌های بعد از کدبوک مستقیماً به قبل از کدبوک کپی می‌شوند و نیازی به محاسبه مشتق فرآیند (تابع) خود کدبوک نیست. می‌توان این فرآیند کپی کردن گرادیان‌ها را با رابطه زیر نوشت:

$$B = A + (f(A) - A).detach() \quad (1)$$

متد $detach()$ باعث می‌شود که متغیر موردنظر به هنگام گرفتن مشتق، عدد ثابت حساب شود که یعنی در فرآیند back-propagation تاثیری نداشته‌باشد. در رابطه بالا A ورودی قبل از کدبوک، B خروجی کدبوک و تابع $f(.)$ خود فرآیند (تابع) کدبوک یا همان Vector Quantization است. **توجه شود**

که تمامی جزئیات پیاده‌سازی مورد نیاز شما برای ساخت کدبوک، در مقاله توضیح داده شده است.

(د) پس از ساختن بلوک‌های مورد نیاز مدل و کامل کردن متد forward، باید تابعی برای ترین کردن مدل با استفاده از تابع زیان زیر بنویسید (تابع زیان زیر دقیقاً همان لاس توضیح داده شده در مقاله است).

$$\mathcal{L} = \log(p(x | z_q(x))) + \|z_e(x) \cdot detach() - e\|_2^2 + \beta \|z_e(x) - e \cdot detach()\|_2^2 \quad (2)$$

(ه) جمله اول خطای بازسازی است (از VAE عادی به یاد داریم که می‌توان آن را به MSE جایگزین کرد)، z_e خروجی انکودر، e خروجی کدبوک است.

(و) پس از انجام مراحل بالا شما باید مدل خود را روی داده گفته شده آموزش دهید، برای ساخت کدبوک، **سه بردار دو بعدی** در نظر بگیرید تا بتوان آن‌ها را نمایش داد. نمودار هزینه بر حسب زمان را نیز نشان دهید.

(ز) ورودی و خروجی مدل، خروجی‌های کدبوک و سه بردار دو بعدی کدبوک را برای تعدادی از ورودی‌ها نشان دهید.

(ح) در یک فایل گیف (ساخت انیمیشن) سه بردار کدبوک را در صفحه دو بعدی در طول زمان نشان دهید.

(ط) حال می‌خواهیم تاثیر زیاد کردن تعداد و ابعاد بردارهای کدبوک را بررسی کنیم:

- ابتدا هر تصویر تک کاناله ورودی را در سه عدد رندوم بین صفر تا یک ضرب کنید و هر کدام را به عنوان یک کانال رنگ (rgb) استفاده کنید و دیتاستی رنگی بسازید.
- حال تعداد بردارها را از سه تا عددی بزرگ مثل پنجاه زیاد کنید (در چند مرحله) و اثر آن را در خروجی مدل گزارش کنید.
- سپس تعداد بردارها را عددی کوچک و ثابت در نظر بگیرید و ابعاد آن را زیاد کنید (مثلاً سه بردار پنجاه بعدی) و اثر آن را در خروجی مدل گزارش کنید.