به نام خدا

رباتیک اجتماعی و شناختی

مدرس : علیرضا طاهری

نیمسال دوم ۰۲-۱۴۰۲

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی مکانیک

تمرین سری سوم | " شبکه‌های بازگشتی و ترنسفورمر" | مهلت ارسال: ۳۰ اردیبهشت ۱۴۰۳

- هم فکری و هم کاری شما در انجام تمارین مانعی ندارد اما پاسخ ارسالی هر فرد در نهایت حتما باید توسط خود او حل و نوشته شده باشد.
- در صورت هم فکری و یا استفاده از منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید. کمک گرفتن از LLM ها در حل تمارین مجاز است!!
- نتایج و پاسخ های خود را در یک فایل فشرده (ترجیحاً به نام HW1-Name-StudentNumber) در سامانه قرار دهید. پیشنهاد می شود برای بخش نرم افزاری از زبان برنامه نویسی پایتون در یکی از محیط های Jupyter notebook و یا Google Colab برای کدنویسی و تست کدهای خود استفاده نموده و فایل کدها را به فرم IPYNB ارسال کنید.
- در صورت داشتن هرگونه سوال و ابهام با دستیاران آموزشی این بخش در ارتباط باشید:
  - سیده نیلوفر حسینی (s.n.hoseini1379@gmail.com)
  - علی مجیبی (alimojibi99@gmail.com)

سوالات تشریحی (۲۰۰ نمره)

۱- با دانش خود و یا بهره گیری از جستجوهای اینترنتی یا سایر منابع، به پرسش‌های زیر پاسخ دهید:

الف) مشکل Vanishing Gradient را توضیح دهید و بیان کنید چگونه در شبکه های LSTM از این مشکل جلوگیری می‌شود؟

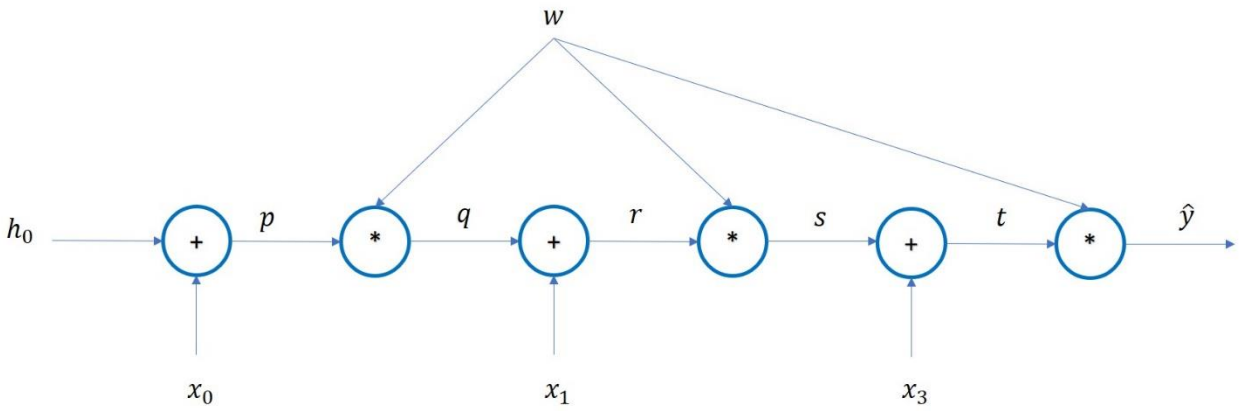
ب) چگونه گیت ورودی، گیت فراموشی و گیت خروجی در یک شبکه LSTM به توانایی آن در یادگیری وابستگی های طولانی مدت کمک می کند؟

پ) شبکه‌های RNN و LSTM چگونه داده‌های گمشده یا ناقص را در توالی مدیریت می‌کنند؟

ت) درمورد اصطلاح Teacher Forcing در آموزش شبکه‌های عصبی جست‌وجو کنید و مزایا و معایب آن را بیان کنید.

ث) معماری Vision Transformer را به صورت مختصر شرح دهید. به نظر شما این معماری چگونه قابلیت تعمیم برای استفاده ورودی از جنس فیلم را دارد؟

۲- شبکه بازگشتی یک بعدی با گراف محاسباتی زیر را درنظر بگیرید. در این شبکه تنها پارامتر اسکالر W را داریم.



شکل ۱: گراف محاسباتی برای یک شبکه بازگشتی ساده

الف) یک رابطه بازگشتی برای شبکه مذکور بنویسید.

ب) با نوشتن روابط انتشار به جلو مقدار $p, q, r, s, t, \hat{y}$ را برحسب ورودی‌ها محاسبه کنید.

پ) تابع خطا را به صورت $(y - \hat{y})^2$ در نظر بگیرید و سپس با استفاده از الگوریتم انتشار خطا به عقب و گراف محاسباتی شکل ۱، مشتقات جزئی روی هرکدام از یال‌های خروجی از W را بدست آورید.

ت) با استفاده از نتیجه قسمت پ، مشتق خطا نسبت W را محاسبه کنید.

ث) با توجه به روابطی که در قسمت‌های قبل بدست آوردید، درمورد دو مشکلی که می‌تواند آموزش شبکه‌های عصبی بازگشتی را دشوار کند توضیح دهید.

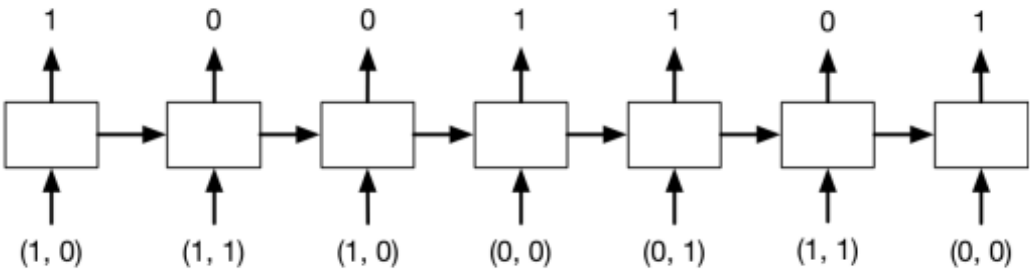
۳- در این مسأله می‌خواهیم از یک شبکه بازگشتی برای پیاده سازی جمع اعداد باینری استفاده کنیم. به‌عنوان مثال جمع زیر را درنظر بگیرید.

$$[100111 + 110010 = 1011001]$$

برای سادگی عملیات جمع را از کم اهمیت ترین بیت (بیت سمت راست) شروع می‌کنیم. به عبارتی ورودی‌های شبکه را به‌صورت زیر درنظر می گیریم. توجه کنید که دنباله‌ها را با ۱ بیت اضافه pad می‌کنیم.

$$Input1: 1, 1, 1, 0, 0, 1, 0$$
$$Input2: 0, 1, 0, 0, 1, 1, 0$$
$$output: 1, 0, 0, 1, 1, 0, 1$$

به‌نوعی شبکه را می‌توان به‌صورت زیر نمایش داد

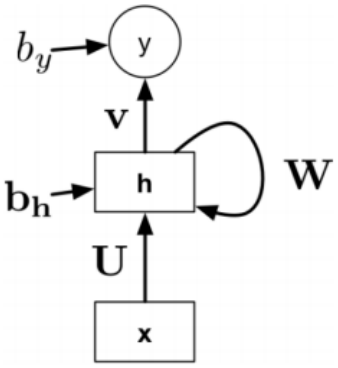


شکل ۲: یک شبکه بازگشتی برای جمع اعداد باینری

هدف طراحی وزن‌ها و بایاس‌ها برای این شبکه بازگشتی است. برای سادگی فرض کنید که هر واحد RNN، ۳ نورون مخفی دارد. همچنین از تابع فعال‌ساز پله با رابطه زیر استفاده کنید:

$$f(x) = \begin{cases} 1 & x > 0 \\ 0 & Otherwise \end{cases}$$

مطلوب است محاسبه ماتریس‌های $\mathbf{U}, \mathbf{V}, \mathbf{W}$ و همچنین مقادیر بایاس $\mathbf{b}_\mathrm{h}, b_y$ برای معماری زیر:



شکل ۳: نمایش فشرده برای شبکه بازگشتی

۴- در مقاله معرفی مدل ترنسفومر (Attention is all you need)، از کدگذاری موقعیت به صورت زیر استفاده شده است.

$$\mathbf{p}_t = \begin{bmatrix} \sin \omega_1 t \\ \cos \omega_1 t \\ \sin \omega_2 t \\ \cos \omega_3 t \\ \vdots \\ \sin \omega_{\frac{d}{2}} t \\ \cos \omega_{\frac{d}{2}} t \end{bmatrix}_{d \times 1}$$

که در آن

$$\omega_k = \frac{1}{10000^{2k/d}}$$

الف) توضیح دهید که پارامترهای d و t بیانگر چه هستند؟

ب) در بخشی از مقاله مذکور، نویسندگان به علت انتخاب این شیوه از کدگذاری موقعیت اشاره می‌کنند:

"We chose this function because we hypothesized it would allow the model to easily learn to attend by relative positions, since for any fixed offset k, $PEpos + k$ can be represented as a linear function of $PEpos$."

در این قسمت می‌خواهیم اثباتی برای این عبارت ارائه دهیم. ابتدا برای یک جفت sin و cos با فرکانس ω_k نشان دهید ماتریس **M** ی وجود دارد که:

$$\mathbf{M}. \begin{bmatrix} \sin \omega_k t \\ \cos \omega_k t \end{bmatrix} = \begin{bmatrix} \sin\big(\omega_k (t + \varphi)\big) \\ \cos\big(\omega_k (t + \varphi)\big) \end{bmatrix}$$

پ) با استفاده از نتیجه قسمت ب ثابت کنید تبدیل خطی T وجود دارد که به کمک آن می‌توان نوشت:

$$\mathbf{T}^{(k)}. \mathbf{p}_t = \mathbf{p}_{t+k}$$

سوالات عملی (۸۰۰ نمره)

۱- با استفاده از معماری LSTM، مدلی برای پیش‌بینی سری زمانی مربوط به دادگان تغییرات آب و هوایی دهلی ارائه دهید و دقت مدل خود را ارزیابی کنید. نوت بوک مربوطه به سوال پیوست است.

۲- هدف از این سوال طبقه بندی تصاویر با استفاده از مدل‌های از پیش آموزش داده شده Vision Transformer است. برای این کار به دلخواه یکی از مدل های از پیش آموزش داده شده Vision Transformer (ViT) را انتخاب نموده و این مدل را برای تسک طبقه بندی تصاویر روی دادگان CIFAR10، Fine Tune کنید. نوت بوک این سوال پیوست است. توجه کنید که حداقل دقت قابل قبول برای این سوال ۸۰٪ است.

۳- لب خوانی خودکار توسط سامانه های رباتیک اجتماعی به صورت بالقوه در تشخیص بهتر علائم زبان اشاره ایرانی کاربرد دارد (جهت اطلاع شما، متاسفانه به علت عدم آشنایی افراد عادی در جامعه و حتی معلمین افراد ناشنوا/با مشکلات شنوایی در کشور با زبان اشاره ایرانی، افراد ناشنوا از همان ابتدای دوران تحصیل به لب خوانی عادت کرده و (به جای تعامل مستمر دوطرفه با دیگران از طریق حرکات دست و دهان به صورت همزمان،) برای رفع نیازهای خود، ناچار به خبره شدن در زمینه لب خوانی هستند!! بگذریم ...). در این سوال مجموعه دادگانی در قالب داده های آموزش و آزمون به شکل فیلم های کوتاه (بدون صدا) با پسوند mp4 در اختیار شما قرار می گیرد که شما می بایست با طراحی شبکه ای مناسب، یک مسئله دسته بندی انجام دهید. تعداد کلمات استفاده شده در این سوال ۵ کلمه (شامل ۱: ایران، ۲: خوشحال، ۳: معلم، ۴: سلام، ۵: خداحافظ) است که توسط ۱۰ فرد مختلف، هر یک به تعداد ۴ مرتبه، بیان شده اند. تعداد افراد آموزش ۸ (شماره های ۱ تا ۸) و تعداد افراد تست ۲ (شماره های ۹ و ۱۰) می باشد. نام گذاری فایل‌ها بدین صورت است که عدد بعد از s شماره فرد، عدد بعد از w شماره کلمه و عدد بعد از r مرتبه تکرار را نشان می‌دهد (برای مثال s1-w2-r3 فایل مربوط به تکرار سوم کلمه ۲: خوشحال توسط فرد شماره ۱ است).

الف) یک شبکه تلفیقی پیچشی-بازگشتی ارائه دهید تا دسته بندی این ۵ کلمه را انجام دهد. ساختار مورد نظر خود را برای انجام این تکلیف طبقه بندی توضیح دهید. ماتریس درهم ریختگی را بر روی دادگان آزمون ارائه کنید.

ب) سوال بخش الف را به کمک طراحی یک شبکه تلفیقی پیچشی-ترنسفورمر مجدداً انجام دهید و نتایج را با بخش الف مقایسه نمایید.

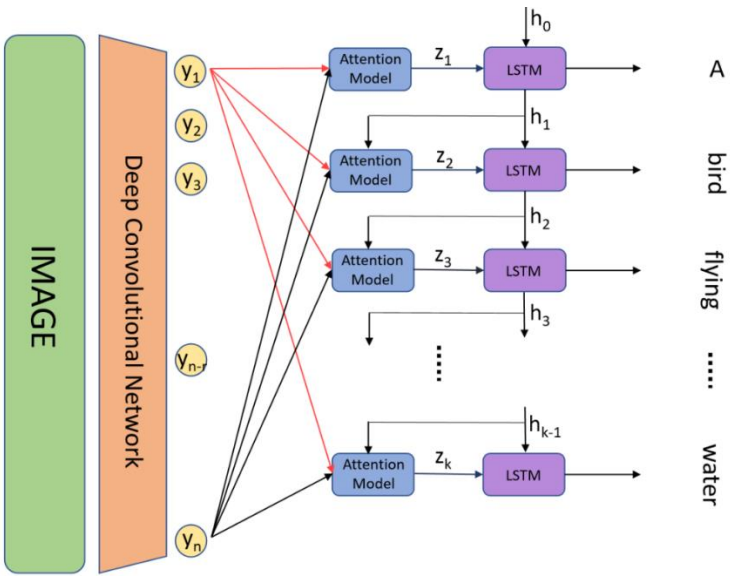
پ) سوال بخش الف را با یک شبکه پیچشی سه بعدی حل کرده و نتایج را با دو بخش قبل مقایسه نمایید.

حداقل دقت قابل قبول برای این سوال، ۷۰ درصد می باشد.

لینک دانلود دادگان آموزش و آزمون:

https://drive.google.com/drive/folders/19po4DgGLAobWcn2Je-hp9lawUWXLYWv2?usp=sharing

۴- در این سوال هدف پیاده سازی یک مدل برای توصیف عکس[۱] بر روی دادگان flickr8k است. شکل ۴ شماتیکی از مدلی که قصد توسعه آن را داریم نشان می‌دهد. برای این سوال بایستی نوت بوک مربوطه را گام به کام کامل فرمایید. جزئیات لازم برای پیاده سازی در نوت‌بوک آورده شده است. ***توجه کنید که در این سوال برای مکانیزم توجه استفاده ازتوابع از پیش آماده شده مجاز نیست.***



شکل ۴: پیاده سازی مدل توصیف عکس با استفاده از ۱- شبکه های پیچشی ۲- مکانیزم توجه ۳- شبکه LSTM

نکات مورد توجه در زمینه ارائه فایل های سوالات عملی:

- ترجیحاً صفحات html از صفحه گوگل کولب یا جوپیتر که هم دارای کدها و هم پاسخ ها در زیر سلول های مربوطه می باشند را برای ما ذخیره و ارسال نمایید.
- پیشنهاد می شود بهترین مدل خود در هر سوال عملی را با فرمت h5. ذخیره کنید تا در صورت نیاز ما در آینده به آن ها، دسترسی ها مقدور باشد.
- در بخش عملی الزامی به ارائه گزارش نبوده و تنها کافی است توضیحات و فرضیات مورد نظر و همچنین بحث روی نتایج حاصله، به صورت مختصر و قابل فهم در کد پایتون نوشته شود.

[۱] Neural image captioning