
 <p>دانشگاه صنعتی امیرکبیر</p>	<p>دوره آموزشی پردازش زبان طبیعی (NLP)</p> <p>تمرین پنجم</p> <p>مهلت تحویل:</p> <p>۲۰ شهریورماه ۱۴۰۰</p>	 <p>آکادمی همراه</p>
---	--	---

در این تمرین قصد داریم با استفاده از مجموعه داده Microsoft Research Paraphrase Corpus یک شبکه سیامس^۱ مبتنی بر شبکه RNN برای Paraphrase Identification آموزش دهیم. برای دسترسی به مجموعه داده آموزش^۲، ارزیابی^۳ و آزمون^۴ به ترتیب از دستورات زیر استفاده کنید. ستون‌های این مجموعه داده به ترتیب زیر هستند:

label، id1، id2، sentence1 و sentence2

این ستون‌ها به ترتیب برچسب مجموعه داده، شناسه جمله اول، شناسه جمله دوم، جمله اول و جمله دوم هستند.

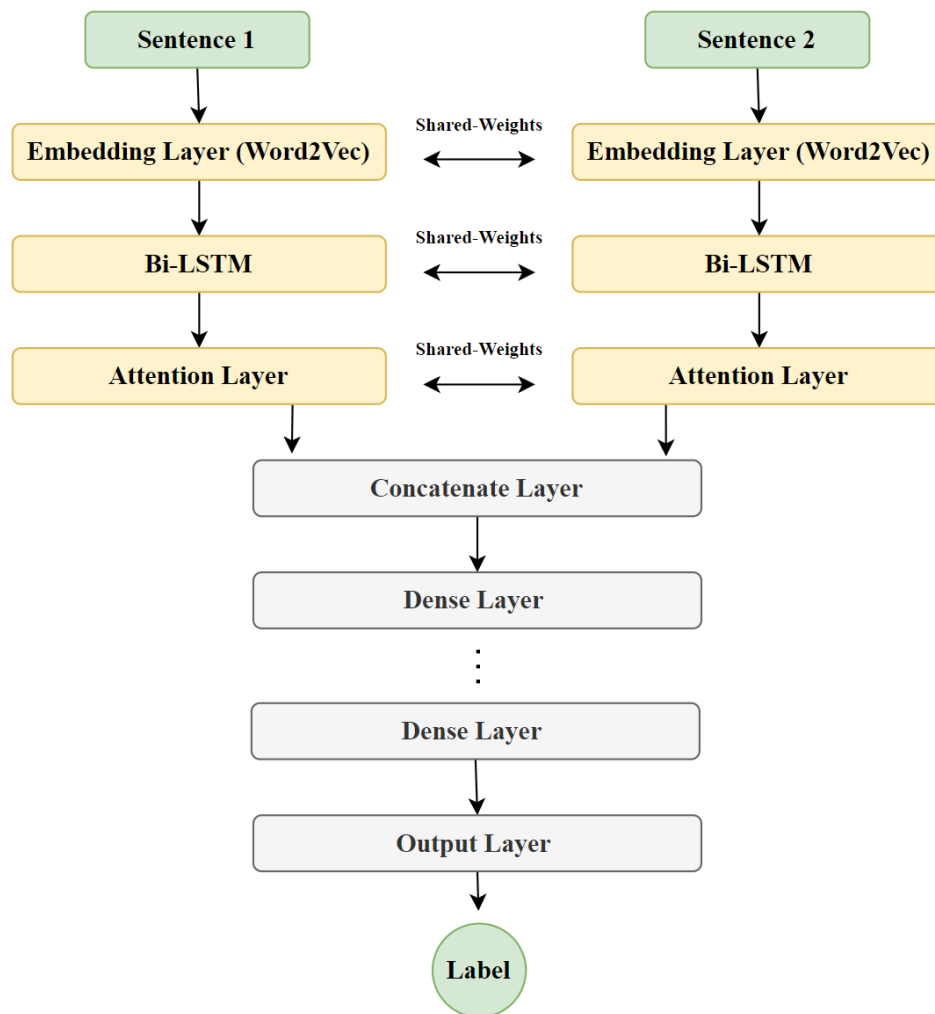
```
!gdown --id 17eLq5Ng5yfbX9tLq2tdiOdscO0ea8xK5
!gdown --id 1Y-68UagW14hwJXNyaR2HnML5jEwvjQ1j
!gdown --id 1P_HLLvGq15gsgDl5P6QYrHkTlkcytMdr
```

در واقع هدف وظیفه Paraphrase Identification در پردازش زبان طبیعی این است که بررسی کنیم آیا دو جمله موجود در مجموعه داده یک مفهوم را منتقل می‌کنند و یا به عبارتی آیا دو جمله بازنویسی شده یکدیگر هستند یا خیر. در صورتی که دو جمله بازنویسی شده یکدیگر باشند برچسب داده برابر با یک و در غیر این صورت صفر خواهد بود.

نکته ۱: برای پیاده‌سازی این تمرین می‌توانید از کتابخانه‌های Keras، TensorFlow و یا PyTorch استفاده کنید.

نکته ۲: در این تمرین قرار است از معماری شبکه سیامس استفاده شود. در معماری شبکه‌های مبتنی بر شبکه سیامس، دو شبکه موازی با معماری یکسان استفاده و وزن‌ها بین دو شبکه موازی به اشتراک گذاشته می‌شود.

¹ Siamese
² Train
³ Validation
⁴ Test



سوال ۱) در معماری شبکه این تمرین قصد داریم از بازنمایی Word2Vec مدل skip-gram استفاده کنیم. در این قسمت مدل skip-gram را بر روی مجموعه داده آموزشی که در اختیارتان قرار گرفته آموزش دهید. بازنمایی کلمات موجود در پیکره را ذخیره کنید تا در قسمت بعد به عنوان ماتریس وزن‌ها برای آموزش شبکه مورد استفاده قرار دهید.

نکته ۳: لینک دسترسی به فایل بازنمایی کلمات را پس از ذخیره در درایو خود در فایل کولب تمرین قرار دهید تا در صورت نیاز هنگام تصحیح تمرین مورد استفاده قرار گیرد.

سوال ۲) در این قسمت شبکه‌ای با معماری فوق را با استفاده از مجموعه داده آموزش آموزش دهید و معیارهای accuracy، f1-measure و recall را برای مجموعه داده آزمون گزارش دهید. توجه کنید که معیارهای precision، recall و f1-measure را دوبار محاسبه کنید یک بار با این فرض که لیبل هدف است و یک بار با این فرض که لیبل صفر لیبل هدف است. سپس میانگین مقادیر precision، recall و f1-measure را هم به صورت micro و هم به صورت macro گزارش کنید. مقدار loss function و accuracy تمام اپیک‌ها در زمان آموزش شبکه را برای داده آموزش و ارزیابی در نموداری رسم کنید.

نکته ۴: پیدا کردن مقادیر بهینه پارامترها و تعداد لایه‌های Dense بعد از لایه Concatenate برای آموزش شبکه به عهده شما می‌باشد.

نکته ۵: بهترین مدل را با استفاده از accuracy مدل بر روی مجموعه داده ارزیابی انتخاب کنید و در درایو خود ذخیره کنید. لینک دسترسی به آن را در فایل کولب تمرین قرار دهید تا در صورت نیاز هنگام تصحیح تمرین مورد استفاده قرار گیرد.

سوال ۳) در معماری شبکه برای سوال قبل از شبکه مکرر Bi-LSTM استفاده شد. در این قسمت لایه Bi-LSTM را یک بار با لایه LSTM و یکبار با لایه GRU جایگزین کنید. معیارهای ارزیابی خواسته شده در سوال ۲ را برای مجموعه داده آزمون برای هر دو معماری جدید گزارش کنید. مانند سوال ۲ مقدار loss function و accuracy تمام ایپاک‌ها در زمان آموزش شبکه را برای داده آموزش و ارزیابی در نموداری رسم کنید.

نکته ۶: بهترین مدل را برای هر دو معماری جدید با استفاده از accuracy مدل بر روی مجموعه داده ارزیابی انتخاب کنید و در درایو خود ذخیره کنید. لینک دسترسی به آن را در فایل کولب تمرین قرار دهید تا در صورت نیاز هنگام تصحیح تمرین مورد استفاده قرار گیرد.

سوال ۴) نتایج سوال ۲ و سوال ۳ را با هم مقایسه کنید و تحلیلی از آن بنویسید.

سوال ۵) (سوال امتیازی اضافه) وزن‌های لایه Attention را به بهترین نحو در نموداری نمایش دهید و تحلیل خود را از آن بنویسید.