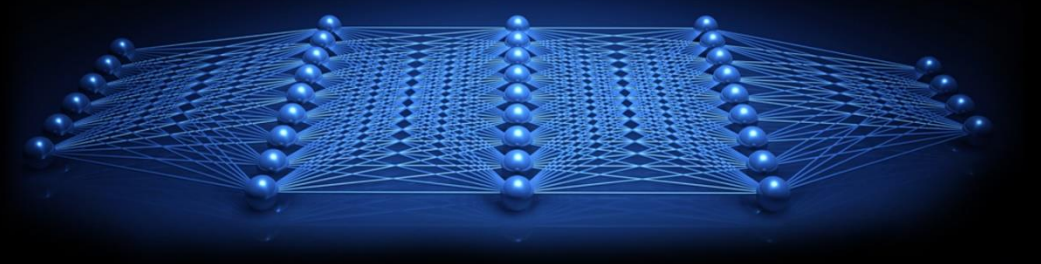
به نام خداوند بخشنده و مهربان

دانشگاه گیلان ۱۳۵۳ University of Guilan 1974

پروژه چهارم درس یادگیری عمیق

دکتر سیّد ابوالقاسم میرروشندل



طرّاح: رضا خان محمّدی

Multi-label Text Classification

داستان پروژه:

کندریک که قصد بر یادگیری زبان شیرین فارسی گرفته اما او در تشخیص *فاعل* و *شباهت جملات* در این زبان به مشکل برخورده است. او از شما کمک خواسته تا در این امر خطیر، او را با آموزش شبکه‌ای عصبی یاری دهید تا ۱) میزان شباهت دو جمله فارسی را طبقه‌بندی کرده و ۲) فاعل را تشخیص دهد.

هدف پروژه:

هدف از این پروژه، آشنایی دانشجو با برچسب‌زنی داده (Data Labelling) طبقه بندی چندبرچسبی (Multi-label Classification)، پردازش متن (Text Processing) و شبکه‌های عصبی بازگشتی (Recurrent Neural Networks) می باشد.

شرح پروژه:

در این پروژه از دیتاست PerSICK استفاده می‌کنیم:

score	sentence1	sentence2
4.5	گروهی از بچه ها در حیاط بازی می کنند و پیرمردی در پس زمینه ایستاده است	گروهی از پسران در حیاط بازی می کنند و مردی در پس زمینه ایستاده است
3.2	گروهی از کودکان در خانه مشغول بازی هستند و هیچ مردی در پس زمینه ایستاده نیست	گروهی از بچه ها در حیاط بازی می کنند و پیرمردی در پس زمینه ایستاده است
4.7	پسران جوان در فضای باز بازی می کنند و مرد در همان نزدیکی لبخند می زند	بچه ها در بیرون از خانه و در کنار یک مرد با لبخند بازی می کنند
3.4	بچه ها در بیرون از خانه و در کنار یک مرد با لبخند بازی می کنند	گروهی از بچه ها در حیاط بازی می کنند و پیرمردی در پس زمینه ایستاده است
3.7	پسران جوان در فضای باز بازی می کنند و مرد در همان نزدیکی لبخند می زند	گروهی از بچه ها در حیاط بازی می کنند و پیرمردی در پس زمینه ایستاده است

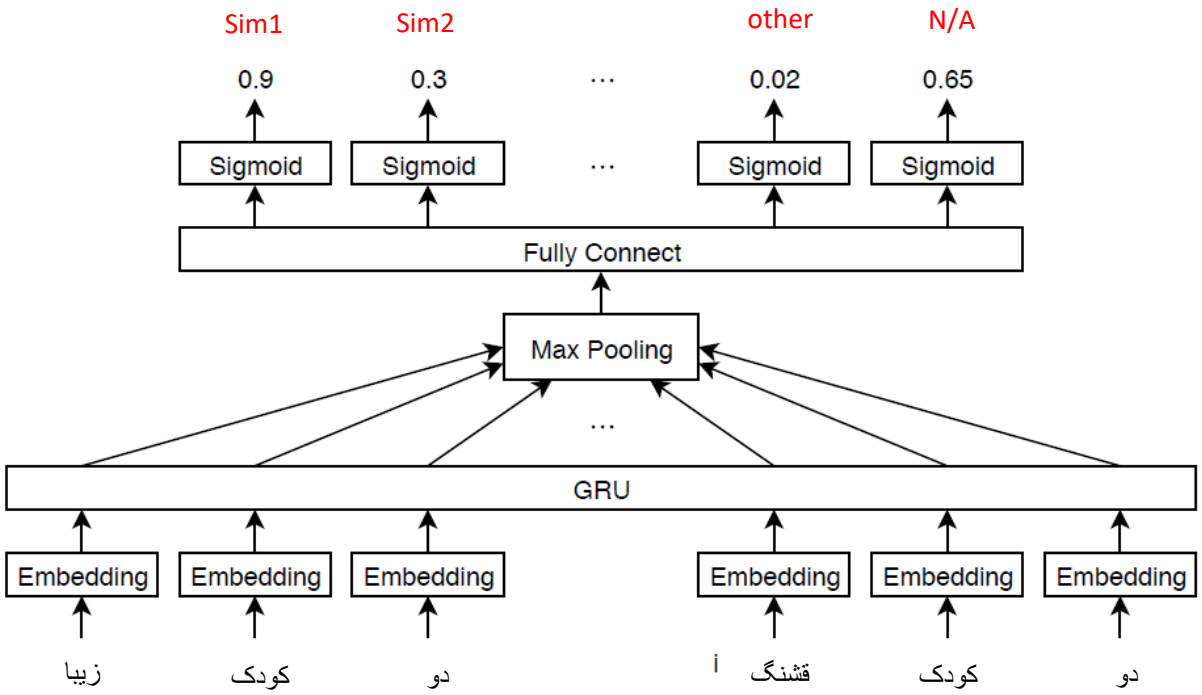
همانطور که مشاهده می‌کنید، این دیتاست سه‌هزار سطر و سه ستون کلی را شامل می‌شود. هر سطر شامل دو جمله sentence1 و sentence2 بوده و میزان شباهت این دو جمله با score مشخص شده‌است. این دیتاست مناسب تسک تشخیص شباهت متنی (Textual Similarity) بوده که در طی آن دو جمله به عنوان ورودی به شبکه عصبی داده‌شده و میزان شباهت آن دو تشخیص داده می‌شود. در حل این مسئله دو راه حل کلی وجود دارد: ۱) حل به روش رگرسیون که در طی آن مقدار دقیق عددی به کمک شبکه عصبی پیش‌بینی شده و ۲) حل به روش طبقه‌بندی که شباهت به دسته های ۱ تا ۵ تقسیم شده و به کمک شبکه عصبی طبقه آن پیش‌بینی می‌شود. شما بایستی در طی این پروژه از روش دوم استفاده کنید، بدین صورت که میزان شباهت بین جفت جملات (score) را طوری گرد کنید تا مسئله از حالت شباهت عددی به شباهت دسته‌ای تغییر کند.

ولی این انتهای ماجرا نیست! با نگاهی دقیق‌تر به دیتاست متوجه می‌شوید که جفت جملات اکثرا در خصوص یکی از عناوین روبرو هستند: کودک یا کودکانی که کار X را انجام می‌دهند، گربه یا سگی که کار Y را انجام می‌دهند، مردی یا زنی که ... . پس می‌توان نتیجه گرفت که به صورت کلی می‌توان این جفت جملات را می‌توان متناسب با موضوع نیز دسته‌بندی کرد. در طی این پروژه شما بایستی این جفت جملات را بر اساس *فاعل* دسته‌بندی کنید. دسته فاعل‌هایی که مدل شما بایستی توانایی طبقه‌بندی کردنشان را داشته باشد برابر است با :

- مرد / پسر: (مثال: مردی در حال پریدن به داخل یک استخر خالی است. پسر جوانی از دیواری که از سنگ ساخته شده بالا می رود)
- زن / دختر: (مثال: زنی کلاه مصری بر سر دارد. دختر جوان با پیراهن صورتی با آرامش روی چمن ها دراز کشیده است)
- کودک / کودکان : (مثال: یک کودک خردسال در حال بالا رفتن از یک دیوار سنگ نوردی است که در داخل خانه قرار دارد)
- حیوانات: (مثال: بچه گربه ها گرسنه هستند. سگی در حال گاز گرفتن قوطی است)
- دیگر: (مثال: مردم به برخی لباس های جمع شده در مجاورت جنگل نگاه می کنند)
- N/A: (برای مواردی که فاعل دو جمله یکی نبوده یا غیر قابل تشخیص است)

همانطور که متوجه شده‌اید دیتاست PerSICK ستون هایی مبنی بر فاعل جملات نداشته و این وظیفه شماست تا آن را استخراج دهید. بدین منظور می‌توانید از تکنیک Dependency parsing کتابخانه‌های معروف پردازش زبان فارسی همچون Parsivar و Hazm استفاده کنید.

تا به اینجای کار دو تسک معرفی کرده‌ایم: طبقه‌بندی میزان شباهت و فاعل دو جمله ورودی. هدف اصلی این پروژه اجرای **همزمان** این دو تسک بوده و شما بایستی بدین منظور از طبقه‌بندی چندبرچسبی استفاده کنید. شبکه عصبی شما با گرفتن دو ورودی مجزا به ازای هر جمله، ۱۱ برچسب را پیش‌بینی کند که ۵ برچسب مختص میزان شباهت بوده (۱ تا ۵) و ۶ برچسب باقی‌مانده همان دسته فاعل ذکر شده (حیوانات، کودکان، دیگر، و...) در بالاست. بدیهی است که شبکه عصبی شما بایستی یادگرفته تا یکی از برچسب‌های ۱ تا ۵ و یکی دیگر از ۶ برچسب فاعلی را انتخاب کند.



شکل بالا تنها **شمای** کلی پروژه را نشان می‌دهد (دقت کنید که ساختار شبکه دقیق نیست).

نکات تکمیلی:

- محدودیتی در انتخاب هایپرپارامترها و ساختار دقیق شبکه‌ها وجود ندارد. هرچند دانشجو موظف است تا با انتخاب مقادیر درست و آموزش بهتر مدل، دقت نهایی مدل پایانی را افزایش دهد.
- مقادیر Accuracy، Precision، Recall، F1 Score، و Confusion Matrix بایستی به ازای استفاده از لایه‌های بازگشتی Simple RNN، GRU، و LSTM گزارش شوند. به طوری که خروجی شما دارای سه سطر (به ازای استفاده از لایه‌های بازگشتی متفاوت در شبکه) و ستون‌هایی به تعداد پارامترهای ارزیابی ذکر شده داشته باشد.
- در آموزش شبکه‌های عصبی، نحوه استفاده **درست** از داده‌هایی که در اختیار دانشجو قرار داده شده دارای اهمیت بالایی بوده و صحت کار مدل‌ها را تعیین می‌کند.
- گزارش پروژه بایستی کامل بوده و دقیق به سوالات و بخش‌های مربوطه پاسخ داده‌باشد.
- پروژه دانشجو بایستی نمودارهای تغییر Loss و Accuracy هر دو مجموعه آموزش و تست را داشته باشد.
- پیاده سازی به صورت گروهی (**حداکثر ۲ نفره**) است و **هیچ محدودیتی** برای زبان برنامه نویسی و فریمووورک یادگیری عمیق مورد استفاده وجود ندارد.
- بحث و بررسی میان دانشجویان آزاد است اما هر دانشجو موظف است پروژه را به تنهایی انجام دهد و در هنگام تحویل حضوری، به تمام جزئیات کد کاملاً مسلط باشد. با موارد تقلب و کپی کردن، طبق تشخیص دوستان حل تمرین، برخورد جدی خواهد شد.
- توجه کنید که کدهای شما باید خوانا و دارای کامنت گذاری مناسب باشد.
- زمان بندی و چگونگی تحویل حضوری پروژه، متعاقباً اعلام خواهد شد.