پروژه

یادگیری عمیق - دانشکده مهندسی برق - زمستان ۱۴۰۱

پروژه درس یادگیری عمیق طراحی سیستم مولتی‌مودال برای تحلیل احساسات است. در این پروژه ابتدا با مجموعه دادگان این حوزه آشنا خواهید شد و سپس شروع به آموزش مدل‌هایی بر پایه شبکه عصبی برای تحلیل احساسات داده مولتی‌مودال (شامل متن, تصویر و صوت) خواهید کرد. مجموعه دادگان استفاده در این پروژه همگی به زبان انگلیسی هستند.

قواعد پروژه :

- راه ارتباطی با تیم پروژه تنها از طریق گروه درس در تلگرام و یا بخش پرسش و پاسخ کوئرا بوده و اعضای تیم پروژه به سوالات مستقیم پاسخ نخواهند داد.
- پروژه با احتساب فاز صفر مجموعا چهار فاز خواهد داشت و مجموعا ۶ روز تاخیر مجاز. پس از این مدت به ازای هر روز ۲ درصد از نمره بخش مربوطه از دست خواهد رفت. توجه فرمایید در فاز چهارم امکان تاخیر وجود نداشته و پس از ددلاین این فاز، تحویل پروژه خواهید داشت.
- پس از ارسال کد هر فاز امکان ایجاد تغییر در کد خود برای فازهای بعدی پروژه را خواهید داشت اما ملاک ارزیابی هر فاز کد آپلود شده برای آن فاز می‌باشد نه کد ارائه شده در انتهای پروژه.
- آپلود پروژه از طریق کوئرا انجام می‌شود. برای راحتی دوستان در پروژه استفاده از GitHub اجباری نمی‌باشد اما توصیه اکید می‌شود به منظور مدیریت بهتر کار گروهی از ابزارهای مربوطه استفاده بفرمایید.

فاز سه

در فاز اول با نحوه پردازش دادگان چهره آشنا شده و مدلی برای تشخیص احساسات بر مبنای آن پیاده نمودید. در فاز دوم نیز همین کار را براساس دادگان متنی انجام دادید. در این فاز، سعی در ترکیب representationهای بدست آمده از دادگان تصویر و متن جهت بهبود عملکرد مدل نهایی در تشخیص احساسات را خواهید داشت. همچنین با نحوه استفاده از مدالیته های مختلف در کاربردهای Weak-supervision و Unsupervised آشنا خواهید شد.

نکته: تاکنون روش های متعددی برای مسائل مورد بررسی در این فاز ارائه شده است. همچنین پیاده سازی و آموزش صحیح و کامل این مدل ها معمولا چالش برانگیز و دشوار می باشد. لذا در این فاز از پروژه تمرکز اصلی بر روی خلاقیت و تبحر شما در بکارگیری اطلاعات مدالیته های مختلف و طرز صحیح پیاده سازی آن می باشد. **البته که دقت مدل شما نیز ارزش و نمره خود را خواهد داشت.**

• بخش اول - ترکیب مدالیته های تصویر و متن

در بخش اول فاز سوم به ترکیب کارهای انجام شده در دو فاز قبل خواهید پرداخت.

- زیر بخش اول - Concatenation

به عنوان اولین روش ترکیب اطلاعات متنی و تصویری می بایست که از بهترین روش بدست آمده خود از فاز اول و دوم پروژه بهره جسته و دو بردار representation (بردارهای قبل از لایه classification) تصویر(فاز اول) و متن(فاز دوم) را با یکدیگر Concatenate نمایید و سپس با آموزش شبکه ای مانند MLP از بردار حاصل شده برای تشخیص احساسات استفاده کنید. لازم است که نتایج خود را با عملکرد مدل های بدست آمده از فاز اول و دوم مقایسه کرده و در صورت بهبود عملکرد مدل در تشخیص احساسات، میزان آن را گزارش نمایید. انتظار می رود که خروجی مدل دارای عملکردی بهتر از (یا حداقل برابر با) دو فاز قبلی پروژه داشته باشد، چرا که میزان اطلاعات داده شده به عنوان ورودی مدل نهایی ما بزرگتر یا مساوی دو فاز قبل می باشد. برای گزارش و مقایسه عملکرد مدل های مذکور می توانید از Accuracy، F1 score و Confusion Matrix استفاده نمایید.

- زیر بخش دوم- Pre-trained Transformer Backbones

در این بخش می بایست که از مدل های Transformer آموزش داده شده به عنوان Backbones استفاده کرده و Finetuning یک مدل تشخیص احساسات را با استفاده از مدالیته های متن و تصویر انجام دهید. برای بدست آوردن یک نگاه کلی درمورد انواع روش های ترکیب این مدالیته های مختلف با استفاده از معماری Transformerها می توانید به این مقاله[1] مراجعه نمایید. همچنین به عنوان دو مثال از

[1] https://arxiv.org/pdf/2206.06488.pdf

Transformerهای چند-مدالیته با وزن های آموزش دیده در دسترس، میتوانید به این دو مقاله[32] رجوع کنید.

نکته: دو مقاله نامبرده صرفا به عنوان مثال ذکر شده اند و هیچ الزامی به اخذ آن ها به عنوان Backbone خود ندارید. در انتخاب معماری کاملا آزاد بوده و پیشنهاد میشود که از مدل های بروز برای انجام این تسک استفاده نمایید. در این زیربخش از پروژه انتظار می رود که با الهام گرفتن از روش های موجود برای ترکیب کردن بردارهای ویژگی حاصل از مدالیته های مختلف، از یک معماری آموزش دیده Transformer به عنوان Backbone استفاده نموده و بردارهای ویژگی حاصل از متن و تصویر را به گونه منطقی با یکدیگر ترکیب کرده و سپس شبکه Transformer انتخابی خود را بر روی دیتاست Finetune نمایید. از آنجا که از یک مدل Transformer آموزش دیده (و ترجیحا SOTA) به عنوان backbone استفاده می شود، انتظار داریم که معیارهای Accuracy و F1 score در مقایسه با نتایج حاصل از زیربخش اول قابل مقایسه (و ترجیحا بهتر) باشند. البته لازم به ذکر است که به دلیل پیچیدگی های موجود در پیاده سازی و نتیجه مطلوب گرفتن از معماری Transformerها و همچنین محدودیت های Computational resource دانشجویان، تمرکز اصلی در هنگام نمره دهی این زیربخش برروی ترکیب کردن درست و منطقی بردارهای ویژگی، و بکارگیری موفق یک مدل Transformerای SOTA برای دادگان چند مدالیته و Finetunning موفقیت آمیز آن خواهد بود(Crash نکردن نتایج گرفته شده شما معیار هست).

- ## بخش دوم- دادگان چندمدالیته برای یادگیری Weakly Supervised

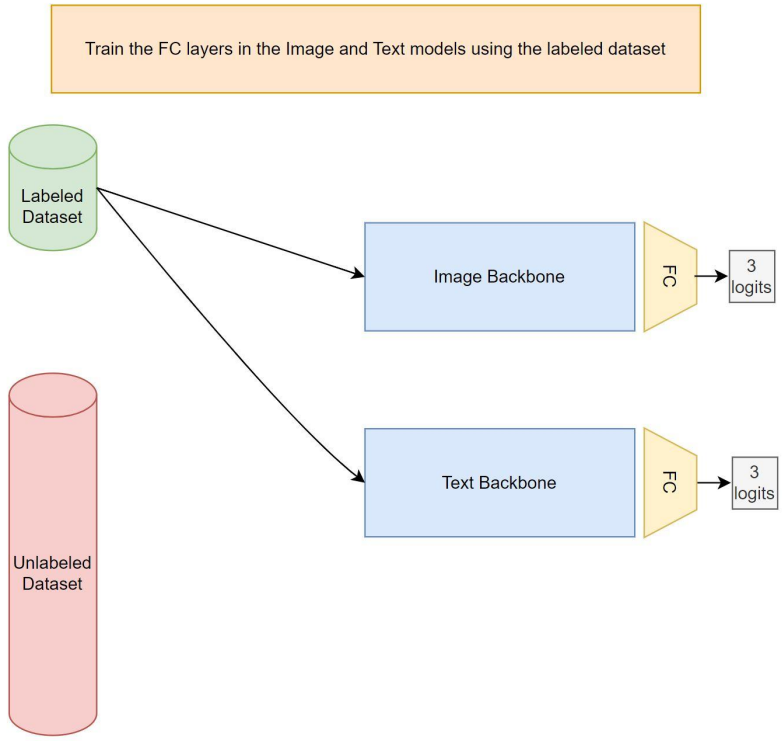
تاکنون، مجموعه دادگان ما بطور کلی شامل تصویر، متن و لیبل مربوطه بوده است. حال فرض کنید که علاوه بر دیتاست لیبل دار MSCTD، ما یک مجموعه دادگان بسیار بزرگ ولی بدون لیبل از جفت تصویر-متن داریم (به عنوان مثال، این دادگان بدون لیبل را می توانید مجموعه تصاویر و زیرنویس سریال دلخواه خود فرض کنید). با انجام این فرضیات، در این بخش از پروژه میخواهیم که با مفاهیم Weakly Supervised Learning و Domain Adaptation آشنا شده و هدف نهایی ما رسیدن به یک مدل تشخیص احساسات مبتنی بر تصاویر برای مجموعه دادگان بدون لیبل خواهد بود. مجموعه دادگان "لیبل دار" و "بدون لیبل" ما الزاما از یک حوزه (Domain) نبوده و عملا می خواهیم با بکارگیری خلاقانه دادگان چند مدالیته لیبل دار خود، مدلی برای تشخیص احساسات مبتنی بر تصویر برای دادگان بدون لیبل بدست آوریم، که خود این دادگان بدون لیبل نیز تضمینی برای نشات گرفتن از حوزه(Domain) دادگان لیبل دار ندارند.

برای رسیدن به هدف ذکر شده لازم است که گام های زیر را انجام دهید:

<u>گام اول</u>: ابتدا یک مدل مبتنی بر Image و یک مدل مبتنی بر Text درنظر گرفته و با بکارگیری Backbone مناسب بر هرکدام از آن ها، وزن لایه های FC آنها را در جهت تشخیص صحیح احساسات با استفاده از مجموعه دادگان لیبل دار بروزرسانی می نماییم(یعنی همان کاری که در دو فاز پیشین پروژه انجام داده اید). این فرآیند در شکل زیر قابل مشاهده می باشد:

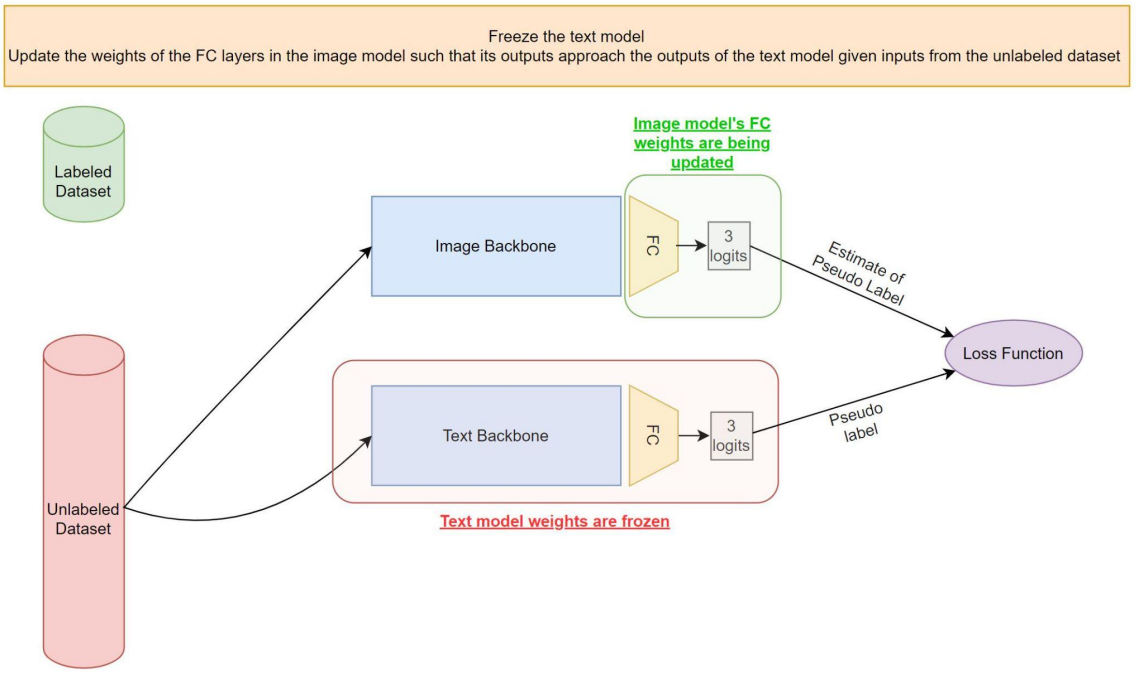
---

[2] https://github.com/airsplay/lxmert
[3] https://github.com/jayleicn/ClipBERT



پس از اتمام فرآیند آموزش مربوط به گام 1، عملا دو مدل خواهیم داشت که یکی از آن ها مبتنی بر تصویر و دیگری مبتنی بر متن بوده و این دو مدل قادر به تشخیص احساسات در حوزه(Domain) مجموعه دادگان لیبل دار ما می باشند.

<u>گام دوم</u>: همانطور که در ابتدا گفته شد، هدف نهایی ما بدست آوردن مدلی مناسب جهت تشخیص احساسات مبتنی بر تصویر برای مجموعه دادگان بدون لیبل بوده است. در این راستا، می توانیم از خروجی مدل مبتنی بر متن آموزش دیده در مرحله قبل به عنوان یک Pseudo label استفاده نماییم. عملکرد این فرآیند بدین صورت است که در ابتدا یک نمونه از مجموعه دادگان بدون لیبل را نمونه برداری کرده و سپس متن و تصویر آن را به عنوان ورودی به مدل های آموزش دیده مرحله قبل می دهیم. سپس خروجی حاصل شده از مدل مبتنی بر متن را به عنوان Pseudo label درنظر گرفته و وزن های لایه های FC مدل مبتنی بر تصویر را جهت میل کردن خروجی آن به Pseudo label بروزرسانی می نماییم. این فرآیند در شکل زیر نشان داده شده است (دقت کنید که در این روش، بهتر است بردار شامل هر سه لاجیت را به عنوان pseudo-label در نظر بگیرید و نه argmax آن‌ها) :



با انجام این فرآیند، خروجی های حاصل شده مدل مبتنی بر تصویر به مرور به خروجی حاصل شده از مدل مبتنی بر متن میل می کند. می توان استدلال نمود که این رویکرد بطور کلی می تواند موجب بهبود عملکرد مدل مبتنی بر تصویر ما بشود، چرا که: 1. مدل های زبانی بطور کلی برای تشخیص احساسات مناسب تر می باشند و زبان عموما در تشخیص احساسات بهتر از چهره عملکرد دارد. 2. مدل های زبانی موجود معمولا روی دادگان بسیار وسیعی آموزش می بینند و عملا logitهای ساخته شده توسط آن ها حاوی اطلاعات بیشتری میتواند باشد. به زبان ساده تر، مدل های زبانی عموما بدلیل آموزش دیدن روی دادگان وسیع میتوانند representationهای بهتری از متن ورودی ایجاد کنند و دخالت دادن این representationهای خوب در فرآیند آموزش تشخیص احساسات مبتنی بر عکس می تواند مفید بوده و به بهبود عملکرد مدل نهایی کمک کند.

توجه نمایید که در مرحله دوم، وزن های مدل مبتنی بر متن Freeze بوده و بروزرسانی براساس نتایج دادگان بدون لیبل برروی آن ها انجام نشده است. بلکه خروجی خود این language model به عنوان لیبل جهت آموزش دادن مدل مبتنی بر تصویر بکاررفته است. به زبان دیگر، مدل مبتنی بر تصویر ما با استفاده از لیبل های نویزی و غیردقیق حاصل از خروجی مدل مبتنی بر متن، وزن های خود را جهت بهبود عملکردش برروی مجموعه دادگان بدون لیبل بروزرسانی می کند. این فرآیند استفاده از روش های خلاقانه برای تولید لیبل هایی که لزوما دقیق و تماما صحیح نمی باشند، با هدف آموزش و بهبود عملکرد یک مدل برروی مجموعه بزرگی از دادگان بدون لیبل را Weakly Supervised Learning می گویند.

همچنین، همانطور که در ابتدا گفته شد، مجموعه دادگان "لیبل دار" و "بدون لیبل" ما الزاما از یک حوزه (Domain) نبودند و ما با بکارگیری رویکرد مذکور توانستیم از "مجموعه دادگان چند مدالیته لیبل دار" خود جهت بهبود عملکرد مدل مبتنی بر تصویر برروی "مجموعه دادگان بدون لیبل" بهره ببریم. به وضوح می توان دید که مدل مبتنی بر تصویر ما پس از اتمام گام دوم می بایست که دارای عملکرد بهتری برروی دادگان بدون

لیبل داشته و اصطلاحا می توان گفت که مدل ما به این حوزه (Domain) از دادگان Adaptation داشته است.

در این بخش بطور گام به گام نشان دادیم که چگونه می توان از مدالیته های مختلف برای انجام تسک Domain Adaptation via Weakly Supervised Learning استفاده نماییم.

شما می بایست که در ابتدا یک مجموعه دادگان بدون لیبل به دلخواه خود انتخاب نمایید(مثلا سریال دلخواه خود به همراه زیر نویس آن، و یا دیتاست های آماده حاوی جفت Image-Caption). مراحل مذکور را با استفاده مدل مبتنی بر متن و مدل مبتنی بر تصویر خود(یعنی مدل های بدست آمده از فاز اول و دوم پروژه) طی کرده، و نتایج مدل مبتنی بر تصویر حاصل شده از انجام فرآیند Domain Adaptation via Weakly Supervised Learning را برروی دادگان بدون لیبل انتخابی خود گزارش نمایید. همچنین می بایست که نتایج بدست آورده شده را با نتایج حاصل از حالتی که مدل مبتنی بر تصویرتان صرفا با استفاده از خود تصویر و لیبل متناظر با آن(یعنی بدون دخالت مدالیته متن) آموزش دیده باشد(عملا مدل اولیه مبتنی بر تصویر که در فاز 1 بدست آوردید)، مقایسه نمایید. گزارش دادن معیارهای Accuracy، F1-score و رسم Confusion Matrix در همه مراحل کفایت می کند.

نکته: پیاده سازی بی عیب و نقص هرآنچه که گفته شد، و همچنین گرفتن نتیجه کاملا مناسب از آن، به عنوان هدف اصلی ما از مطرح کردن این بخش از فاز سوم پروژه **نبوده است.** هدف ما در این بخش از فاز سوم، آشنایی شما با مفاهیم ارائه شده و طی کردن گام به گام این مراحل جهت کسب درک عمیق تری از موضوعات مطرح شده می باشد(به همین دلیل بر لزوم استفاده از مدل هایی که خودتان در فازهای پیشین پروژه بدست آوردید و پرهیز نمودن از بکارگیری مدل های بزرگ و پیچیده تاکید شده است)

تاکید اصلی در فرآیند نمره دهی بر فهم درست مسائل عنوان شده، و طی نمودن صحیح مراحل تحقق آن خواهد بود. البته که عملکرد مدل نهایی شما و نتایج گزارش شده نیز ارزش و نمره خود را خواهد داشت.