دانشکده مهندسی کامپیوتر

جمع‌آوری و برچسب‌گذاری خودکار مجموعه داده برای آموزش مدل‌های شرح متراکم ویدئو

پروژه کارشناسی مهندسی کامپیوتر

دانشجویان:

امیرحسین احمدی

محمد صدرا خاموشی فر

استاد راهنما:

دکتر بهروز مینایی بیدگلی

دکتر عیسی زارع پور

شهریور ۱۴۰۲

تأییدیه هیأت داوران جلسه‌ی دفاع از پروژه

نام دانشکده: مهندسی کامپیوتر

نام دانشجویان: امیرحسین احمدی، محمد صدرا خاموشی فر

عنوان پایان‌نامه: جمع‌آوری و برچسب‌گذاری خودکار مجموعه داده برای آموزش مدل‌های شرح متراکم ویدئو

تاریخ دفاع: شهریور ۱۴۰۲

رشته: مهندسی کامپیوتر

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضاء
۱	استاد راهنما	دکتر بهروز مینایی بیدگلی	استاد	دانشگاه علم و صنعت	
۲	استاد راهنما	دکتر عیسی زارع پور	استادیار	دانشگاه علم و صنعت	

تأییدیه صحت و اصالت نتایج

باسمه تعالی

اینجانبان امیرحسین احمدی و محمد صدرا خاموشی فر به شماره دانشجویی‌های ۹۷۵۲۲۲۹۲ و ۹۷۵۲۱۲۶۱ دانشجویان رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نماییم که کلیه نتایج مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانبان تحت نظارت و راهنمایی عضو هیأت علمی دانشگاه علم و صنعت ایران بدون هر گونه دخل و تصرف انجام گرفته و به موارد نسخه‌برداری شده از آثار دیگران، مطابق مقررات و ضوابط، ارجاع داده شده و مشخصات کامل منابع را در فهرست منابع ذکر کرده‌ایم. این پایان‌نامه قبلاً برای احراز هیچ مدرکی ارائه نگردیده است.

در صورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مولفان و منصفان و قانون ترجمه، تکثیر و نشریات و آثار صوتی، ضوابط و مقررات آموزشی و پژوهشی، انضباطی و غیره) با اینجانبان رفتار خواهد شد و حق هرگونه اعتراض در خصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نماییم. در ضمن، مسئولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده اینجانبان خواهد بود و دانشگاه هیچ گونه مسئولیتی در این خصوص نخواهد داشت.

کلیه نتایج و حقوق حاصل از این پایاننامه متعلق به دانشگاه علم و صنعت ایران است. هرگونه استفاده از نتایج علمی و عملی و واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری ترجمه و اقتباس از این پایاننامه بدون موافقت کتبی دانشگاه علم و صنعت ایران ممنوع است. نقل مطالب با ذکر منبع بلامانع است.

نام و نام خانوادگی:
امیرحسین احمدی
محمد صدرا خاموشی فر

امضا و تاریخ:

مجوز بهره‌برداری از پایان نامه

بهره‌برداری از این پایان‌نامه در چارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.

☐ بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پایان‌نامه تا تاریخ ..... ممنوع است.

استاد راهنما:
دکتر بهروز مینایی بیدگلی
دکتر عیسی زارع پور

تاریخ:

امضا:

تقدیر و تشکر

از اساتید گرانقدر، جناب آقایان دکتر بهروز مینایی، دکتر عیسی زارع پور و به ویژه دکتر صالح اعتمادی و مهندس محمدجواد پیرهادی که در طول مدت تحقیق، ما را از رهنمودها و تجارب با ارزش خویش بهره‌مند ساختند، صمیمانه سپاسگزاریم. همچنین برخود لازم می‌دانیم تا از حمایت‌ها و محبت‌های بی‌دریغ خانواده و دوستان عزیزمان صمیمانه تشکر و قدردانی کنیم.

چکیده

تبدیل ویدئو به متن از مسائل پیچیده در پردازش تصویر است که در زمینه‌های مختلفی مانند تولید خودکار عنوان، تعامل انسان و کامپیوتر، کمک به افراد معلول و ساده‌سازی ویدئوهای آموزشی به صورت متنی کاربرد دارد. تسک شرح متراکم ویدئو[۱] به مسئله محلی‌سازی[۲] رویدادهای مهم در ویدئو و شرح هر کدام از آن ها در قالب یک جمله کوتاه می‌پردازد. هدف این پروژه جمع‌آوری مجموعه داده‌ای است که بتوان از آن برای پیش آموزش مدل‌های یادگیری عمیق، بینایی کامپیوتر و پردازش زبان‌های طبیعی در این حوزه استفاده نمود. این مدل‌ها به عنوان ورودی یک ویدیو کوتاه را دریافت می‌کنند، سپس به عنوان خروجی بخش های مهم آن را استخراج کرده و در یک جمله درباره آن توضیح می‌دهد.

تا به امروز مجموعه داده‌های مختلفی برای آموزش مدل های شرح متراکم ویدئو جمع‌آوری شده است. اکثر آن‌ها شامل کمتر از ده هزار ویدئو می‌شوند و به صورت دستی برچسب‌گذاری شده‌اند. به عنوان مثال مجموعه داده YouCook2[۱] که یکی از مشهور ترین مجموعه داده‌های این حوزه است، شامل حدود ۲۰۰۰ ویدئو است که هر کدام به صورت دستی و توسط یک گروه بزرگ برچسب گذاری شده‌اند.

بیشتر تمرکز این پژوهش بر روی برچسب‌گذاری خودکار ویدئوها است تا بتوان مجموعه داده‌های بزرگتری را در زمان کوتاه‌تر جمع‌آوری کرد. ویدئوهای این مجموعه‌ها، از مجموعه داده‌های بزرگتری مانند HowTo100M[۲] جمع‌آوری شده و تمام ویدئوهای استفاده شده در آن‌ها نیز از سایت YouTube جمع‌آوری شده‌اند. در این پژوهش دو روش برای برچسب‌گذاری داده‌ها ارائه خواهد شد. روش اول با استفاده از توضیحات متنی ارائه شده برای ویدیوها و دیگری با استفاده از داده‌های خام زیرنویس ویدیوها است. در انتها نیز مدلی با استفاده از داده‌های جمع‌آوری شده آموزش داده خواهد شد تا بتوان تاثیرگذاری داده‌ها را با سایر مجموعه داده‌ها در این حوزه مقایسه کرد.

واژه‌های کلیدی: شرح متراکم ویدئو، برچسب‌گذاری خودکار، YouTube

[۱] Dense Video Captioning

[۲] Localization

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱: مقدمه

۱-۱ شرح مسئله

برقراری ارتباط درباره دنیای تصویری از طریق زبان یکی از توانایی‌های مهم انسان به عنوان باهوش‌ترین موجودات است. حتی یک کودک ۵ ساله می‌تواند اشیا را لمس کند، حرکات خود را ببیند و با زبان خود آن‌ها را توصیف کند. بزرگسالان نیز با کسب توانایی‌های بیشتر می‌توانند فیلم ببینند، کتاب بخوانند و آن‌ها را یاد بگیرند. این ارتباط بین زبان و ویدئو حال توانسته به هوش مصنوعی تعمیم داده شود که بتواند محتوای تصویری را درک کند و در مورد آن با انسان‌ها ارتباط برقرار کند. هوش مصنوعی هنوز با چالش‌های مهمی در زمینه‌های بازیابی متن به ویدئو، محلی‌سازی رویدادهای ویدئو، شرح ویدئو و ... رو به رو است و پیشرفت در این زمینه‌ها برای بسیاری از کاربردها مانند جست‌وجوی آرشیوهای ویدئو و ارتباط انسان با کامپیوتر نیاز است[۲].

شرح متراکم ویدئو[۳] نیز یکی از تسک‌های بسیار چالش برانگیز در پردازش تصویر و متن است. توصیف دقیق و منسجم رویدادها در یک ویدئو نیازمند درک جامع و کامل محتوای ویدئویی و شناخت رویدادهای مهم آن است. برای رسیدن به این هدف فارق از معماری انتخابی برای آموزش مدل، همواره نیاز به یک مجموعه داده ویدئویی بزرگ وجود دارد. در واقع با توجه به تنوع بسیار زیاد محتواهای ویدئویی، برای یک یادگیری مناسب احتمالا نیاز به میلیون‌ها ویدئو و توضیحات متنی داریم. با این حال مجموعه داده‌های موجود در این زمینه در مقیاس هزاران محتوای ویدئویی هستند که به صورت دستی برچسب‌گذاری شده‌اند. جمع‌آوری چنین مجموعه‌داده‌هایی بسیار هزینه‌بر و مقیاس کردن آن بسیار کار سختی است. همچنین برچسب‌گذاری یک کار ذهنی است و کسی که مسولیت آن را بر عهده دارد همواره باید از لحاظ ذهنی با ثبات باشد[۳].

در این پژوهش ما روش متفاوتی را برای بدست آوردن ویدئوهای لازم و رویدادهای آن برای شرح متراکم ویدئو بررسی می‌کنیم. می‌دانیم که ویدئوهای آموزشی و روایی در مقادیر زیاد در دسترس هستند (مانند YouTube) و حجم زیادی از داده‌های بصری و متنی را ارائه می‌کنند. همچنین مجموعه داده‌های بزرگی با اهداف متفاوت از شرح متراکم ویدئو جمع‌آوری شده اند که می‌توانند کار ما را برای پیدا کردن محتوای ویدئویی مورد نیاز بسیار ساده‌تر کنند. مجموعه داده HowTo100M یکی از مجموعه داده‌های بزرگ در حوزه پردازش تصویر است. این مجموعه با بیش از ۱.۲ میلیون ویدئو آموزشی روایی که در آن انسان‌ها را در حال انجام بیش از ۲۳۰۰۰ کار مختلف به تصویر می‌کشد به همراه زیرنویس تمام ویدئوها کمک بزرگی به ما در انجام این پژوهش خواهد کرد[۲].

---

[۳] شرح متراکم ویدئو ارائه یک راه حل برای دو مسئله متفاوت است: ۱- محلی‌سازی رویدادهای مهم درون ویدئو ۲- توضیحی درباره هر رویداد در حد یک جمله

۱-۲ اهداف پژوهش

هدف از انجام این پژوهش یافتن راهی برای جمع‌آوری یک مجموعه داده برای شرح متراکم ویدئو است که بتوان آن را به صورت خودکار و بدون نیاز به نیروی انسانی زیاد و متخصص برچسب‌گذاری کرد و شامل تعداد بالایی ویدئو برای پیش آموزش جامع مدل‌های این حوزه باشد. همانطور که گفته شد، اهمیت این کار آنجاست که مدل‌های شرح متراکم ویدئو می‌توانند به صورت End-to-End[۴] با استفاده از مجموعه بزرگی از ویدئوها و رویدادهای مربوط به آن آموزش ببینند. همچنین با توجه به تنوع بالای ویدئوها مدل‌ها می‌توانند به صورت جامع آموزش دیده و برای کارهای مختلف استفاده شوند.

[۴] در حوزه هوش مصنوعی End-to-End Learning به تکنیکی گفته می‌شود که مدل تمام مراحل از ورودی اولیه تا خروجی نهایی را آموزش می‌بینند.

فصل ۲: کارهای مرتبط

تعداد قابل توجه‌ای از پژوهش‌های حوزه بینایی کامپیوتر مبتنی بر درک محتوای ویدئویی و متنی به صورت مشترک و در کنار یکدیگر است. این پژوهش‌ها شامل شرح خودکار تصویر و ویدئو، پاسخگویی به سوالات تصویری، بازیابی محتوای تصویری بر اساس پرسش‌های متنی، یافتن زمان رویدادهای موجود در ویدئو (شرح متراکم ویدئو) و یا خلاصه‌سازی ویدئو با استفاده از زبان طبیعی هستند. همانطور که پیش‌تر اشاره شد، در این پژوهش تمرکز ما بر روی شرح متراکم ویدئو است.

۲-۱ شرح متراکم ویدئو

شرح متراکم ویدئو یا Dense video captioning در سال ۲۰۱۷ توسط Ranjay Krishna[4] معرفی شد. این تسک در واقع نقطه تلاقی دو تسک محلی‌سازی و شرح رویداد است. اکثر روش‌های موجود برای شرح متراکم ویدئو شامل یک مرحله محلی‌سازی و یک مرحله شرح رویداد به دنبال آن است[۴-۸]. اما برای بهبود بخشیدن تعامل بین تسک‌ها، در کارهای اخیر ماژول‌های شرح رویداد و محلی سازی را به صورت مشترک و همزمان آموزش می‌دهند[۹-۱۲].

۲-۲ مجموعه داده‌ها

به دلیل نوظهور بودن شرح متراکم ویدئو، تعداد مجموعه داده‌هایی که به طور اختصاصی برای این حوزه جمع‌آوری شده‌اند بسیار محدود است. در این بخش نگاهی کلی به آمار و ارقام این مجموعه داده‌ها می‌پردازیم و نقاط ضعف و قدرت آن‌ها را مورد بررسی قرار می‌دهیم.

۲-۲-۱ مجموعه داده YouCook2

مجموعه YouCook2[۱] در سال ۲۰۱۷ جمع‌آوری شده است. این مجموعه داده شامل ۱۷۹۰ ویدئو برش نخورده از مراحل آشپزی است. طول ویدئوهای این مجموعه به طور میانگین چیزی حدود ۳۲۰ ثانیه می‌باشد و به صورت دستی با ۷.۷ جمله برای هر ویدئو برچسب‌گذاری شده است. همانطور که بالاتر هم به آن اشاره شد، این مجموعه داده فقط برای دسته بندی آشپزی ارائه شده و نسبت به مجموعه‌های دیگر تعداد کمتری ویدئو را شامل می‌شود.

شکل ۲-۱ نمونه‌ای از مجموعه داده YouCook2 در مورد ساخت ساندویچ [۱]

۲-۲-۲ مجموعه داده ViTT

این مجموعه داده[۱۳] در سال ۲۰۲۰ جمع‌آوری شده و به نسبت YouCook2 جدیدتر می‌باشد. این مجموعه از مجموعه داده بزرگ‌تری به نام YouTube-8M[۱۴] جمع‌آوری شده است. در این مجموعه برخلاف YouCook2 تمرکز ویدئوها فقط روی آشپزی نیست و محدوده وسیع‌تری از موضوعات را شامل می‌شود. البته همچنان اکثریت ویدئوها به دسته آشپزی تعلق دارند.

این مجموعه شامل ۷۶۷۲ ویدئو است که در حدود ۳۰۰۰ تای آن مربوط به دسته آشپزی می‌باشند که به آن ViTT-Cooking می‌گویند. به طور میانگین طول هر ویدئو ۲۵۰ ثانیه است و با ۷.۱ جمله برچسب‌گذاری شده است. حدود ۵۰۰۰ ویدئو یک بار و مابقی ویدئوها ۲ بار برچسب‌گذاری شده‌اند که برخی مدل‌ها[۵] برای اضافه کردن تعداد ویدئوها هر کدام از برچسب‌گذاری‌ها را یک مثال جدا در نظر می‌گیرند.

همانطور که مشاهده می‌شود در این مجموعه سعی شده است ایرادات مربوط به YooCook2 برطرف شود، ولی همچنان تعداد ویدئوها به نسبت مجموعه داده‌هایی مانند HowTo100M بسیار محدود است و ویدئوها گستردگی لازم را ندارند.

۲-۲-۳ مجموعه داده ActivityNet-Captions

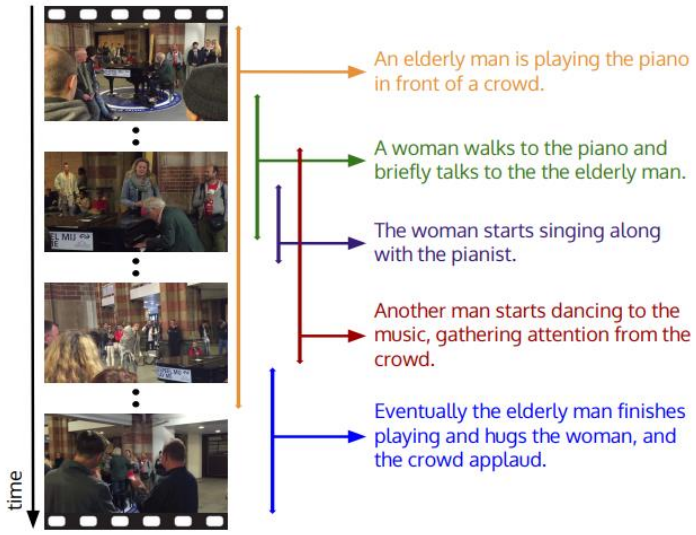
این مجموعه[۴] که در سال ۲۰۱۷ ارائه شده شامل ۱۴۹۳۴ ویدئو برش نخورده از فعالیت‌های انسانی است. برخلاف دو مجموعه قبلی که شامل محتوای گفتاری رونویسی شده[۶] بودند بیش از نیمی از ویدئوهای این مجموعه محتوای گفتاری رونویسی شده ندارند. به طور میانگین طول هر ویدئو ۱۲۰ ثانیه است و با ۳.۷ جمله

---

[۵] Vid2Seq[15] A. Yang *et al.*, "Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 10714-10726, 2023.

[۶] Transcribed speech content

برچسب‌گذاری شده است. برخی از ویدئوهای این مجموعه نیز همانند ViTT شامل ۲ یا چند برچسب‌گذاری متفاوت هستند.



شکل ۲-۲ نمونه‌ای مجموعه داده ActivityNet-Captions از پیانو زدن یک مرد در جمعیت [۴]

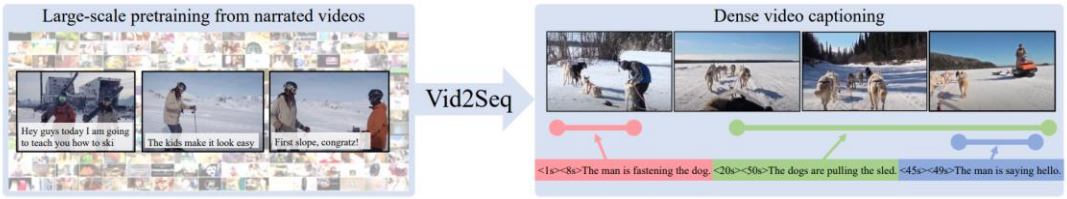
با اینکه تعداد ویدئوهای این مجموعه ۲ برابر ViTT است ولی طول ویدئوها و تعداد جملات آن بسیار کمتر می‌باشد که نقطه ضعف اصلی این مجموعه است. همچنین محتوای گفتاری رونویسی شده می‌تواند کمک خوبی برای پیش آموزش[۷] مدل‌های این حوزه باشد که در این مجموعه به تعداد محدودی وجود دارند.

۲-۲-۴ مجموعه داده HowTo100M

این مجموعه داده[۲] به طور اختصاصی برای شرح متراکم ویدئو جمع‌آوری نشده است ولی اخیرا در یک مقاله برای پیش آموزش یک مدل شرح متراکم ویدئو به نام Vid2Seq[۱۵] استفاده شده است. این مجموعه شامل ۱.۲۲۱ میلیون که معادل با ۱۵ سال ویدئو است می‌باشد. همان طور که از نام آن پیداست (How to) تمام ویدئوهای این مجموعه آموزشی روایی هستند که به می‌توان رویدادهای مختلف ویدئو و مراحل آموزش را از آن تشخیص داد. تمام ویدئوهای این مجموعه حاوی زیرنویس آن‌ها در YouTube و زمان هر زیرنویس می‌باشد. (ممکن است زیرنویس‌ها دستی یا به صورت تولید خودکار[۸] باشند)

[۷] Pre-train

[۸] Auto-generated



شکل ۲-۳ نمونه‌ای از عملکرد مدل Vid2Seq و پیش آموزش آن توسط HowTo100M [۱۵]

همانطور که گفته شد، این مجموعه هیچ برچسب‌گذاریی برای شرح متراکم ویدئو ندارد ولی در Vid2Seq از زیرنویس های هر ویدئو به عنوان رویدادهای آن استفاده شده است. درست است که ممکن است تعداد بالای رویدادها و دقت پایین آن‌ها دقت را پایین بیاورد، ولی از آنجایی که تعداد ویدئوهای این مجموعه به طور نمایی از دیگر مجموعه داده‌ها بیشتر است، می‌تواند گزینه بسیار خوبی برای پیش آموزش و آشنایی اولیه مدل با ویدئوها و رویدادها باشد. در Vid2Seq نیز به طرز هوشمندانه‌ای از آن برای پیش آموزش استفاده شده و سپس مدل بر روی سایر مجموعه داده‌ها با برچسب‌گذاری دقیق تر Finetune[۹] می‌شود.

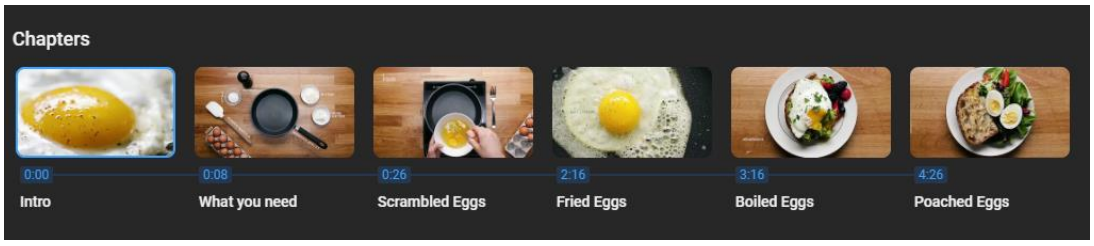
[۹] فرایندی است که در آن پارامترهای یک مدل باید به‌صورت خیلی دقیقی تنظیم شوند تا مدل با مشاهدات مشخصی تناسب پیدا کند.

فصل ۳: روش‌های پیشنهادی

۳-۱ برچسب گذاری با استفاده از Chapters

۳-۱-۱ ویژگی Chapters

حتما تا به حال برایتان پیش آمده که زمان تماشای یک ویدئو در YouTube نوار پیشرفت ویدئو را به صورت قسمت شده ببینید. هر کدام از این قسمت‌ها دارای یک عنوان به خصوص و توضیح مختص آن قسمت از ویدئو است. این یکی از ویژگی‌های مهم YouTube برای تولیدکنندگان محتوا به اسم Chapters است. Chapters ویژگی است که با آن می‌توان ویدئوهای طولانی را به بخش‌های کوتاه‌تر تقسیم کرد. یکی از استفاده‌های مهم این ویژگی در محتواهای آموزشی است. به طور مثال فرض کنید ویدئویی درباره آموزش پخت تخم مرغ دارید و آن را به سه بخش روشن کردن گاز، شکاندن تخم مرغ و اضافه کردن ادویه‌های مورد نیاز تقسیم می‌کنید. حالا بیننده این ویدئو اگر روشن کردن گاز را بلد باشد، می‌تواند بدون درنگ به مرحله شکاندن تخم مرغ برود و در وقتش صرفه جویی شود.



شکل ۳-۱ تصویر Chapters مربوط به یک ویدئو پخت تخم مرغ در YouTube - مرجع

۳-۱-۲ نحوه برچسب‌گذاری

در فصل‌های گذشته بررسی کردیم که محدودیت نیروی انسانی که تمرکز لازم را برای برچسب‌گذاری ویدئو داشته باشد یکی از مشکلات جمع‌آوری مجموعه داده‌ها است. اما اگر از خود تولیدکننده هر ویدئو بخواهیم تا برچسب گذاری آن را انجام دهد، به احتمال زیاد تمرکز و تخصص مورد نیاز برای این کار را دارد و می‌تواند کمک بزرگی در این راه باشد. ساختار Chapters بسیار نزدیک به ساختار رویدادهایی است که ما برای تسک شرح متراکم ویدئو نیاز داریم با این تفاوت که رویدادها می‌توانند جدا از هم باشند ولی Chapters را بازه‌های به هم چسبیده تشکیل می‌دهند.

۳-۱-۳ بررسی نقاط ضعف و قوت

همانطور که اشاره شد، محتواهای آموزشی روایی با احتمال بالاتری دارای Chapters هستند. به همین دلیل گزینه مناسبی برای جمع‌آوری داده‌ها به این روش محسوب می‌شوند. از این رو می‌توانیم از ویدئوهای

HowTo100M[۲] برای مجموعه داده خود استفاده کنیم. با این کار طیف وسیعی از ویدئوها با موضوعات مختلف را در نظر گرفته‌ایم. با این حال این مسئله وجود دارد که ممکن است تمام ویدئوهای موجود در HowTo100M دارای Chapters نباشند. با بررسی‌های انجام شده بر روی داده‌های حوزه آشپزی HowTo100M حدود ۰.۳ درصد ویدئوها دارای Chapters بودند. با تعمیم دادن این مقدار به ۱ میلیون ویدئو موجود احتمالا می‌توان چیزی حدود ۳۰۰۰ ویدئو از آن استخراج کرد که در مقایسه با YouCook2 (۲۰۰۰ ویدئو) آمار قابل قبولی به نظر می‌رسد. همچنین احتمالا می‌توان با استفاده از مجموعه داده‌های جدیدتر و بزرگ‌تر مانند YT-Temporal-1B[۱۶] به اعداد بالاتری نیز دست یافت.

۳-۲ برچسب گذاری با استفاده از زیرنویس

بسیاری از ویدئوهای YouTube دارای زیرنویس هستند. این زیرنویس‌ها ممکن است به صورت دستی و یا به صورت خودکار توسط خود YouTube قرار گرفته باشند. در مجموعه داده HowTo100M[۲] تمام ویدئوها به همراه زیرنویس‌ها موجود هستند. زیرنویس‌های هر ویدئو به صورت تعدادی جمله و زمان شروع نمایش و پایان نمایش آن ویدئوها در سایت موجود هستند. بنابراین ممکن است زمان زیرنویس‌های مجاور با یکدیگر تداخل داشته باشد.

۳-۲-۱ دلایل استفاده از زیرنویس

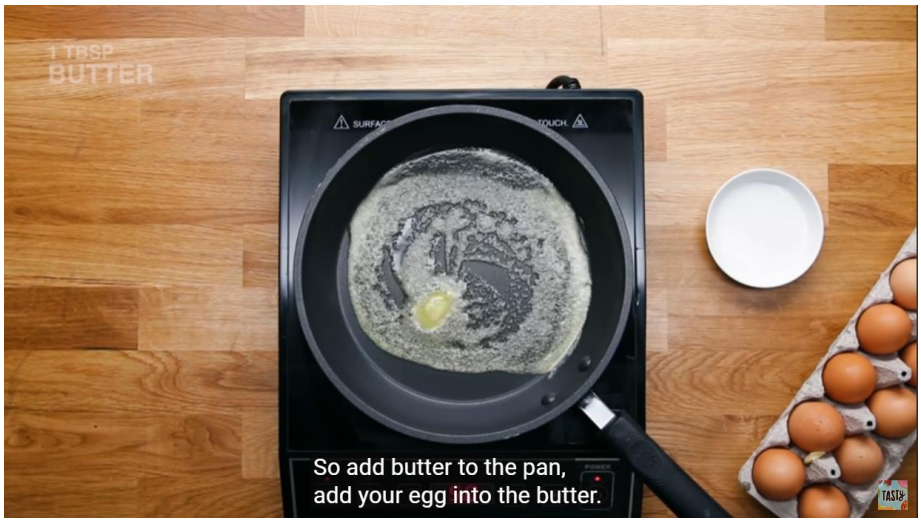
همانطور که پیش‌تر گفته شد، در مدل Vid2Seq[۱۵] از داده‌های HowTo100M برای پیش آموزش استفاده شده و هر کدام از زیرنویس‌ها به عنوان یک رویداد در نظر گفته شده است. با این که زیرنویس‌ها احتمالا تطابق چندانی با رویدادهای یک ویدئو ندارند ولی با توجه به تعداد بالای ویدئوهای این مجموعه داده می‌توان از نویز موجود در زیرنویس‌ها چشم‌پوشی کرد و به عنوان پیش آموزش از این داده‌ها استفاده کرد. حال در این پژوهش ما تلاش کردیم که بدون نیروی انسانی و با روشی خودکار، نویز موجود در زیرنویس‌ها را کم کرده و خطای آن‌ها را پایین بیاوریم.

۳-۲-۲ نحوه برچسب‌گذاری

در شرح متراکم ویدئو برای توصیف رویدادها نیاز به یک جمله داریم، اما زیرنویس‌ها به صورت جملات کامل نیستند و هر قسمت آن فقط بخشی از صحبت‌های درون ویدئو را شامل می‌شود. این بخش ممکن است قسمتی از یک جمله و یا قسمت‌هایی از جملات پشت سر هم باشند. برای بدست آوردن جملات کامل ما تصمیم گرفتیم که تمام زیرنویس‌های موجود برای یک ویدئو را به یکدیگر چسبانده و متن کامل ویدئو را بدست آوریم. سپس

جملات را با استفاده از مدل‌های زبان‌های طبیعی موجود از یکدیگر جدا کنیم. همچنین از آنجایی که زمان شروع و پایان هر زیرنویس را داریم، می‌توانیم زمان تقریبی هر کلمه را بدست آورده و در نتیجه پس از جدا کردن جملات زمان هر جمله را بدست آوریم.

حال که تمام جملات موجود در ویدئو بدست آمدند می‌توان از بین آن‌ها جملاتی که کلیدی‌تر به نظر می‌رسد را به عنوان رویداد در نظر بگیریم. برای این مورد فهرستی از فعل‌ها که نمایانگر انجام کاری هستند[۱۰] را بدست آورده و سپس جملات حاوی این فعل‌ها را به عنوان رویدادهای ویدئو در نظر می‌گیریم.



شکل ۳-۲ بخشی از زیرنویس یک ویدئو آشپزی که نمایانگر یکی از رویدادهای ویدئو است- مرجع

۳-۲-۳ بررسی نقاط ضعف و قوت

ایرادی که می‌توان به این روش برچسب‌گذاری گرفت این است که ممکن است صحبت‌های درون ویدئو لزوما راجع به رویدادهای آن نباشد یا حتی ممکن است یک ویدئو بی‌صدا باشد. این نکته درستی است ولی باید دو مورد را در نظر بگیرم. مورد اول این است که هدف از این نوع برچسب‌گذاری جمع‌آوری خودکار تعداد بالایی ویدئو است. درست است که ممکن است دقت برچسب‌ها به خوبی مجموعه داده‌هایی مانند YouCook2[۱] نباشد، ولی تعداد بالای آن‌ها می‌تواند کمک خوبی برای پیش آموزش مدل‌ها باشد. مورد دومی که باید در نظر گرفته شود این است که ویدئوهای موجود در مجموعه‌داده HowTo100M همه محتوای آموزشی روایی دارند. بنابراین کم پیش می‌آید که موضوع صحبت‌ها راجع به چیز متفاوتی باشد. برای مثال اگر ویدئوی آشپزی را در

[۱۰] Action verbs

نظر بگیریم، معمولا فرد آشپز در باره‌ی نحوه کار و مراحل انجامی صحبت می‌کند و می‌توان از صحبت‌هایش به عنوان رویدادهای احتمالی ویدئو استفاده کرد.

فصل ۴: پیاده‌سازی و آمار

۴-۱ استخراج Chapters از توضیحات

برای استفاده از Chapters تولیدکننده محتوا می‌بایست در توضیحات[۱۱] مربوط به ویدئو خود به ازای هر قسمت اضافه یک خط اضافه کند. در این خط می‌بایست ابتدا زمان شروع این قسمت و بعد از آن عنوان قسمت را وارد کند. فقط باید توجه کرد که قسمت‌ها باید از زمان ۰۰:۰۰ شروع شوند، ترتیب خط‌ها باید مانند ترتیب قسمت‌ها باشد، زمان‌ها باید به طور مرتب شده باشند، ویدئوها باید حداقل شامل ۳ قسمت باشند و طول هر قسمت باید حداقل ۱۰ ثانیه باشد.

۴-۱-۱ چالش‌ها

سایت YouTube ای‌پی‌آیی[۱۲] برای دریافت مستقیم Chapters ارائه نمی‌دهد. به همین علت ما نمی‌توانیم از Chapters خودکار تولید شده توسط خود سایت استفاده کنیم. در طول زمان پیاده سازی راه‌کارهایی مانند استفاده از ابزار متن باز برای دریافت محتوای ویدئوها امتحان شد. به طور مثال این ابزار[۱۷] با Crawl کردن صفحه ویدئو می‌توانست Chapters دستی و خودکار آن را بدست آورد. ولی نتوانست برای تعداد بالای ویدئوها جوابگو باشند. بنابراین مجبور شدیم که به Chapters دستی ویدئوها بسنده کنیم و آن را از توضیحات ویدئو با استفاده از ای‌پی‌آی ارائه شده توسط خود YouTube بدست آوریم. البته تعداد درخواست‌ها در این روش نیز محدود بود و روزانه فقط ۱۰۰۰۰ درخواست قابل انجام بود و برای برداشتن محدودیت نیاز به مکاتبه با YouTube داشت که در طول مدت محدود این پژوهش به سرانجام نرسید.

۴-۱-۲ جزئیات

برای استفاده از ای‌پی‌آی YouTube ابتدا نیاز است که از Google Developer Console یک API Key دریافت کرده و با استفاده از آن درخواست‌ها را ارسال کنید. ما پس از دریافت API Key توضیحات مربوط به ۱۰۰۰۰ ویدئوی آشپزی را دریافت کردیم. سپس زمان‌های به فرمت MM:SS و HH:MM:SS موجود در توضیحات را به همراه عنوان آن‌ها استخراج کرده و ذخیره کردیم. Chapters توضیحاتی که با ۰۰:۰۰ شروع نمی‌شوند را به در نظر نمی‌گیرد ولی ما به دلیل کمبود محتوا این توضیحات را نیز در نظر گرفتیم. همچنین توضیحاتی که شامل کمتر از ۳ زمان را مشخص کرده بودند حذف کردیم تا فقط داده‌ها با مفهوم‌تر شوند.

---

[۱۱] Description

[۱۲] API

۴-۱-۳ نتایج

از مجموع ۱۰۰۰۰ ویدئو بررسی شده فقط ۲۹ ویدئو دارای Chapters بودند. در نتیجه با توجه به محدودیت‌های بررسی شده و تعداد کم ویدئو تمرکز بیشتری رو این بخش گذاشته نشد. ولی در آینده اگر گوگل ای‌پی‌آی مربوط به Chapters و راهی برای بیشتر کردن محدودیت درخواست‌ها به وجود بی‌آید با این روش می‌توان مجموعه داده‌ای بزرگ بدون هزینه زیاد برای شرح متراکم ویدئو جمع‌آوری کرد.

۴-۲ استخراج رویدادها از زیرنویس

۴-۲-۱ محاسبه زمان حدودی هر کلمه

در مجموعه داده HowTo100M به ازای هر ویدئو تعدادی جمله (زیرنویس ها) و زمان شروع و پایان نمایش آن‌ها در ویدئو موجود است. در مرحله اول نیاز داریم که تمام زیرنویس‌ها را به هم بچسبانیم تا بتوانیم متن اصلی را بدست آوریم. چالش اصلی این بخش زمان زیرنویس‌ها بود. همانطور که احتمالا در سایت YouTube مشاهده کرده‌اید، زمان نمایش داده شدن یک بخش از زیرنویس بخش بعدی به خط بالایی می‌رود و تا مدتی رو صفحه می‌ماند تا بیننده بتواند آن را بخواند. به همین دلیل زمان نمایش زیرنویس‌ها با یک دیگر تداخل دارد.

در این روش ما سعی داشتیم تا زمان حدودی هر کلمه را با تقسیم زمان آن زیرنویس به تعداد کلماتش بدست آوریم. سپس با جدا کردن جملات از یکدیگر بتوانیم زمان حدودی جمله را از زمان کلمه اول و آخر آن متوجه شویم. اما اگر زمان زیرنویس‌ها با یکدیگر تداخل داشته باشند، ممکن است زمان کلمه آخر حتی زمان عقب تری نسبت به کلمه اول جمله داشته باشد. به همین علت زیرنویس‌هایی که با یکدیگر تداخل دارند را به صورت یک دسته در نظر گرفتیم و سپس زمان حدودی کلمات هر دسته را از تقسیم کل زمان آن دسته به تعداد کلماتش بدست آوردیم.

۴-۲-۲ نقطه گذاری و جداسازی جملات

زیرنویس‌های موجود در YouTube ممکن است به صورت دستی و توسط تولیدکننده محتوا یا به صورت خودکار تولید شده باشند. در هر صورت تضمینی وجود ندارد که این متون دارای نقطه گذاری مناسب باشند و ممکن است انتها هر جمله بعدی به جمله بعد متصل باشد. به همین دلیل برای این مرحله نیاز به نقطه‌گذاری[۱۳] جملات داریم تا بتوان جملات را از یکدیگر تشخیص داد که برای آن می‌توان از مدل‌های زبان‌های طبیعی

[۱۳] Punctuation

استفاده کرد. ابزار NeMo[۱۸] یک ابزار هوش مصنوعی محاوره‌ای است که برای پژوهش‌های گوناگون حوزه زبان‌های طبیعی کاربرد دارد. ما با استفاده از مدل "punctuation_en_bert" که این ابزار برای نقطه گذاری ارائه می‌دهد استفاده کردیم و متن بدست آمده را نقطه گذاری کردیم. سپس تا انتهای هر نقطه را به عنوان یک جمله و زمان کلمات اول و آخر هر جمله را به عنوان زمان شروع و پایان آن جمله در نظر گرفتیم.

با توجه به اینکه استفاده از ابزار ارائه شده برای نقطه گذاری زمان زیادی می‌برد و نیاز به پردازنده‌های گرافیکی قدرتمند برای پردازش کل یک میلیون ویدئو HowTo100M ما توانستیم ۱۰۰۰۰ ویدئو مربوط به بخش آشپزی را در این مرحله پردازش کنیم. هدف از اینکار مقایسه این مجموعه داده جدید در برابر YouCook2 است که بتوان تاثیر داده آموزشی بیشتر را در برابر دقت پایین‌تر برچسب‌گذاری‌ها را مشاهده کرد. اما در آینده با پردازش تمام ویدئوها می‌توان چندین برابری این روش را مشاهده و استفاده نمود.

۴-۲-۳ انتخاب رویدادها از میان جملات

با توجه به اینکه ویدئوهای این مجموعه آموزشی روایی هستند می‌توان نتیجه گرفت که رویدادهای این ویدئوها بیانگر انجام یک عملی می‌باشند. حال اگر فهرستی از فعل‌هایی که بیان‌گر انجام کاری هستند داشته باشیم، می‌توانیم جملات مهم که کاندیدای رویدادها می‌باشند را جدا کنیم. مجموعه داده HowTo100M فهرستی از عناوین تسک‌های انجام شده در ویدئوهای خود ارائه می‌دهد. ما نیز با جستوجوی این فعل‌ها در جملات استخراج شده، رویدادهای هر ویدئو را بدست می‌آوریم.

۴-۲-۴ آمار و نتایج بدست آمده

پس از بررسی برچسب ویدئوهایی که حاوی حداقل یک رویداد بودند، متوجه نکات ارزشمندی شدیم که در ادامه به آنها پرداخته خواهد شد.

- در اکثر ویدئوها رویدادهای انتخابی مطابق با کار انجام شده در تصویر است. اما در برخی به دلیل عدم همگامی بین ویدئو و زیرنویس، رویدادها از بین رفته و یا زمان اشتباهی را نشان می‌دهند.
- چالش دیگر در شناسایی رویدادها، وجود چندین معنی برای بسیاری از افعال بود. این مشکل از افعال چندمعنی در زبان انگلیسی ناشی می‌شود. ما برای یافتن رویدادها از فهرستی از افعال که بیان‌گر انجام کاری می‌باشند استفاده کردیم. با این حال ممکن است این افعال چند معنی داشته باشند و معنی آن‌ها در جمله، انجام کاری را نشان ندهد.
- چالش آخری که در شناسایی رویدادها با آن مواجه شدیم، استفاده از افعال آینده در زیرنویس بود. در برخی موارد، زیرنویس ویدئوها حدود ۳ تا ۴ ثانیه قبل از وقوع یک رویداد نمایش داده می‌شود.

از بین حدود ۱۰۰ داده‌ای که بررسی کردیم، بیش از ۵۰ درصد رویدادها به درستی استخراج شده بودند، حدود ۲۰ درصد رویداد اشتباهی را تشخیص داده، و تقریبا ۳۰ درصد از رویدادها شناسایی نشده بودند.

به عنوان مثال، در یک ویدئو که در مورد چندین برنامه آشپزی با مدت زمانی بین ۲ تا ۴ دقیقه بود، تمامی رویدادها به درستی استخراج شدند. جملاتی که از این ویدئو به عنوان رویداد تشخیص داده شده است به ترتیب به این صورت می‌باشند :

جدول ۴- ۱ رویدادهای مربوط به یک برنامه آشپزی

جملات انگلیسی	ترجمه فارسی
Dip cookie cutters into flour and cut the dough into desired shapes	کاترهای شیرینی را به آرد آغشته کرده و خمیر را به شکل دلخواه برش دهید.
Next, sprinkle the shapes with granulator, and colored sugar, then bake at ۳۷۵ degrees for seven to eight minutes until the edges are light brown	بعد شکل ها را با گرانول و شکر رنگی بپاشید سپس با دمای ۳۷۵ درجه به مدت هفت تا هشت دقیقه بپزید تا لبه ها قهوه ای روشن شوند.
Remove the cookies from the baking sheets and cool on wire racks.	کوکی ها را از ورقه های پخت خارج کرده و روی توری های سیمی سرد کنید.

این نشان می‌دهد که در صورتی که مدت زمان ویدئوها کوتاه باشد، به دلیل تراکم زیرنویس‌ها و رویدادها، دقت در شناسایی رویدادها افزایش می‌یابد و بیشتر آنها به درستی تشخیص داده می‌شوند. همچنین دستوری بودن جملات به هنگام تشخیص رویداد بسیار کمک کننده بوده است و باعث می‌شود که رویدادها درست‌تر مشخص شوند. در مورد ویدئوهای طولانی‌تر که بین ۵ تا ۱۰ دقیقه به طول می‌انجامند، بسیاری از رویدادها به درستی استخراج نمی‌شوند زیرا نسبت تعداد کلمات و جملاتی که به خود رویداد اشاره کنند کاهش یافته و این امر باعث می‌شود دقت تشخیص رخداد ها کاش یابد.

فصل ۵: ارزیابی بر روی مدل

۵-۱ مدل GVL

[۱۹]GVL یک فریم‌ورک آموزش ویدئویی و زبانی[۱۴] به طور مشترک برای ویدئوهای برش نخورده است. این مدل توانسته است به بهترین نتایج در شرح متراکم ویدئو بر روی مجموعه داده‌هایی مانند ActivityNet Captions، YouCook2 و ... برسد و نتایج قابل رقابتی در سایر تسک‌های تولید و فهم زبان بدست آورد.

۵-۲ نحوه ارزیابی

همانطور که گفته شد، هدف از این پژوهش جمع‌آوری داده هایی برای پیش آموزش مدل‌های شرح متراکم ویدئو بود. برای ارزیابی این مجموعه داده‌ها تصمیم گرفتیم تا ابتدا یک مدل GVL را یک بار بدون پیش آموزش بر روی یکی از مجموعه داده‌های استفاده شده در آن آموزش دهیم. سپس مدل GVL دیگری را بار دیگر با پیش آموزش بر روی داده‌های جمع‌آوری شده پژوهش، بر روی همان مجموعه داده آموزش دهیم. از مقایسه نتایج این دو مدل می‌توان تاثیر پیش آموزش بر روی داده‌ها را مشاهده نمود.

برای این کار ما مجموعه داده YouCook2 را در نظر گرفتیم. این مجموعه داده حاوی ۲۰۰۰ ویدئو آشپزی است. مجموعه داده جمع‌آوری شده ما نیز شامل ۱۰۰۰۰ ویدئو آشپزی است که به صورت خودکار برچسب گذاری شده‌اند. با توجه به تعداد ۵ برابری ویدئوها و موضوعات نزدیک به همه ویدئوهای این دو مجموعه، مجموعه داد ما می‌تواند داده پیش آموزش مناسبی برای YouCook2 باشد.

۵-۳ جزئیات و نتایج آموزش

ابتدا از ویدئوهای هر دو مجموعه فیچرهای TSP[20] استخراج کردیم که به زمان حساس هستند و برای تسک‌های محلی سازی ویدئو کاربرد دارند. سپس مدل اول را با تنظیمات مربوط به بهترین مدل GVL برای YouCook2 در ۲۰ ایپاک آموزش دادیم. مدل دوم را نیز با همان تنظیمات ابتدا در ۲۵ ایپاک توسط مجموعه داده‌مان پیش آموزش داده و سپس آن را مانند مدل اول با YouCook2 آموزش دادیم.

برای شرح متراکم ویدئو، GVL چهار استاندارد مختلف ارائه می‌دهد:

۱. METEOR یک معیار ارزیابی است که میزان شباهت بین خروجی مدل و مرجع را با در نظر گرفتن کلمات هم معنا اندازه گیری می‌کند.
۲. Recall تعداد رویدادهای واقعی در ویدئو که توسط مدل تشخیص داده شده‌اند را نسبت به کل رویدادها نشان می‌دهد.

[۱۴] Joint video-language learning

۳. Precision تعداد رویدادهای تشخیص داده شده توسط مدل را که با رویدادهای مرجع هم‌خوانی دارند را نسبت به کل رویدادهای تشخیص داده شده نشان می‌دهد.
۴. soda_c بر اساس فاصله زمانی بین رویدادهای تشخیص داده شده و رویدادهای مرجع، کیفیت خروجی را بررسی می‌کند.

جدول ۱-۵ مقایسه معیارهای ارزیابی مدل GVL با پیش آموزش و بدون پیش آموزش

معیارها	بدون پیش آموزش	با پیش آموزش
METEOR	0.020	**0.022**
Recall	0.27	0.24
Precision	0.022	**0.025**
soda_c	0.035	0.033

همانطور که در جدول قابل مقایسه است، با استفاده از پیش آموزش توانستیم معیار METEOR را افزایش دهیم. بنابراین این پیش آموزش توانسته به مدل در شناخت بیشتر کلمات و تطبیق تصویر با کلمات کمک کند.

از طرفی دیگر معیار Precision افزایش و معیار Recall کاهش یافته است. از این می‌توان نتیجه گرفت که احتمالا مدل با پیش آموزش در ایپاک‌های بالاتر حساسیت بیشتری در تشخیص رویدادها پیدا کرده است. از این رو هم دقت رویدادهای انتخابی بالاتر رفته (Precision) و هم بسیاری از رویدادها را از دست داده‌ایم (Recall). همچنین به دلیل آموزش مدل با تنظیمات پیش فرض GVL می‌توانستیم حدس بزنیم که مدل می‌تواند Overfit شود.

معیار soda_c نیز کاهش داشته که می‌تواند ناشی از نحوه تخمین زمان رویدادها باشد. که این مورد می‌تواند با استفاده از الگوریتم‌های جدید موجود در زمینه تخمین زمان بهبود یابد.

در مجموع می‌توان نتیجه گرفت که با بهبود بخش انتخاب جملات از کاهش معیار soda_c جلوگیری کرد و با تنظیم بهتر مدل نتایج بهتری در Recall گرفت. همچنین با افزایش داده‌ها می‌توان درک بهتری از کلمات به مدل داد و نتایج بهتری در معیارهای METEOR و Precision گرفت.

فصل ۶: جمع‌بندی و کارهای آینده

۶-۱ جمع‌بندی

در این پژوهش تلاش کردیم تا بتوانیم با استفاده از روش‌های خودکار مجموعه داده‌هایی بزرگ مناسب برای پیش آموزش مدل‌های شرح متراکم ویدئو جمع‌آوری کنیم. ابتدا بررسی کردیم که مشکلات اصلی مجموعه داده‌های کنونی چیست و چطور می‌توان آن‌ها را بهتر کرد. مشکل اصلی این مجموعه‌ها برچسب‌گذاری هزینه‌بر آن‌ها بود. به همین علت تلاش کردیم مجموعه داده‌هایی را به صورت خودکار برچسب‌گذاری کنیم. می‌دانستیم که دقت برچسب‌گذاری این مجموعه‌ها هرگز نمی‌تواند با برچسب‌گذاری به صورت دستی رقابت کند. اما به لطف برچسب‌گذاری خودکار می‌توان مجموعه داده‌های بزرگی را برای پیش آموزش مدل‌های شرح متراکم بدست آورد. نتایج بدست آمده نیز نشان می‌دهد که استفاده از این مجموعه‌ها برای پیش آموزش می‌تواند کمک بزرگی باشد و آموزش را برای داده‌های با دقت بالاتر آسان‌تر کند.

۶-۲ کارهای آینده

با توجه به نتایج بدست آمده و همچنین محدودیت‌های موجود، می‌توان گفت جای پیشرفت زیادی وجود دارد. کارهای زیادی وجود دارند که می‌تواند در آینده باعث بهبود این نتایج شوند که در ادامه به تعدادی از آن‌ها می‌پردازیم

- همانطور که اشاره کردیم، در این پژوهش دو روش برای برچسب‌گذاری ارائه شد که روش اول به مرحله بررسی نتایج نرسید. در آینده اگر بتوان محدودیت‌های موجود در تعداد درخواست‌ها را با مکاتبه برداشت می‌توان این روش را برای تمام داده‌ها امتحان نمود و آن را مانند روش دوم بررسی نمود.
- در روش اول ما فقط به Chapters تولید شده توسط تولیدکنندگان محتوا بسنده کردیم. اما در آینده اگر بتوان راهی برای استفاده از Chapters خودکار پیدا کرد می‌توان تعداد داده‌ها را بسیار افزایش داد.
- استفاده از مجموعه داده‌های بزرگتر و جدیدتر به جای HowTo100M می‌تواند بهبود بیشتری در نتایج هر دو روش ببخشد.
- در روش دوم فقط داده‌های بخش آشپزی مورد بررسی قرار گرفت، در آینده با در نظر گرفتن تمام دسته‌بندی‌ها تعداد ویدئوی بیشتری در دسترس خواهیم داشت و همچنین می‌توان مدل‌های جامع‌تر و برای اهداف مختلف آموزش داد.
- در مرحله تشخیص زمان هر کلمه الگوریتم‌های جدیدی معرفی شده‌اند. البته که این الگوریتم‌ها مانند whisper-timestamped[۲۱] نیازمند پردازش ویدئوها برای تشخیص زمان کلمات می‌باشند، ولی با استفاده از آن‌ها می‌توان زمان دقیق تر رویدادها را بدست آورد و استفاده کرد.

- برای این پژوهش در مرحله انتخاب رویدادها به شکل ساده ای جملاتی که حاوی یکسری فعل‌های خاص بودند در نظر گرفته شدند. اما در آینده می‌توان تمرکز بیشتری در این زمینه گذاشت و الگوریتم‌های بهتری برای انتخاب رویدادها به کار برد.
- تمرکز این پژوهش بیشتر بر روی نحوه‌ی برچسب‌گذاری داده‌ها بود و برای انتخاب ویدئو اقدامی صورت نگرفت. اما در آینده می‌توان تلاش کرد تا ویدئوهایی که زیرنویس بهتری برای روش دوم و یا برای روش اول داده مناسب‌تری دارند انتخاب شوند. همچنین با توجه به خودکار بودن برچسب گذاری فقط محدودیت‌های سخت افزاری و YouTube می‌توانند مانع از هرچه بزرگتر شدن این مجموعه داده‌ها گردند.
- ولاگ‌ها از محتواهایی هستد که با توجه به ذات روایی بودنشان می‌توانند منبع خوبی برای جمع‌آوری ویدئو برای روش دوم باشند و در آینده می‌توان بر روی جمع‌آوری آن‌ها تمرکز کرد.

فصل ۷: مراجع

[١] Zhou, L. a. Xu, C. a. Corso, and J. J, "Towards Automatic Learning of Procedures From Web Instructional Videos," in *AAAI Conference on Artificial Intelligence*: AAAI, 2018, pp. 7590--7598.

[٢] A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev ,and J. Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 27 Oct.-2 Nov. 2019 2019, pp. 2630-2640, doi: 10.1109/ICCV.2019.00272 .

[٣] J .Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 5288-5296, doi: 10.1109/CVPR.2016.571 .

[٤] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-Captioning Events in Videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 22-29 Oct. 2017 2017, pp. 706-715, doi: 10.1109/ICCV.2017.83 .

[٥] V. E. Iashin and E. Rahtu, "A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer," *ArXiv,* vol. abs/2005.08271, 2020.

[٦] V. E. Iashin and E. Rahtu, "Multi-modal Dense Video Captioning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 4117-4126, 2020.

[٧] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 7190-7.٢٠١٨ ,١٩٨

[٨] T. Wang, H. Zheng, M. Yu, Q. Tian, and H. Hu, "Event-Centric Hierarchical Representation for Dense Video Captioning," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 31, pp. 1890-1900, 2020.

[٩] A. Chadha, G. Arora, and N. Kaloty, "iPerceive: Applying Common-Sense Reasoning to Multi-Modal Dense Video Captioning and Video Question Answering," *ArXiv,* vol. abs/2011.07735, 2020.

[١٠] S. Chen and Y.-G. Jiang, "Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 8421-8431, 2021.

[١١] C. Deng, S. Chen, D. Chen, Y. He, and Q. Wu, "Sketch, Ground, and Refine: Top-Down Dense Video Captioning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 234-243, 2021.

[١٢] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-End Dense Video Captioning with Parallel Decoding," *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 6827-6837, 2021.

[١٣] G. Huang, B. Pang, Z. Zhu, C. Rivera, and R. Soricut, "Multimodal Pretraining for Dense Video Captioning," Suzhou, China, December 2020: Association for Computational Linguistics, in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 470-490. [Online]. Available: https://aclanthology.org/2020.aacl-main.48. [Online]. Available: https://aclanthology.org/2020.aacl-main.48

[١٤] S. Abu-El-Haija *et al.*, "YouTube-8M: A Large-Scale Video Classification Benchmark," *ArXiv,* vol. abs/1609.08675, 2016.

[١٥] A. Yang *et al.*, "Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 10714-10726, 2023.

[١٦] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision," *ArXiv,* vol. abs/2108.10904, 2021.

[١٧] *YouTube-operational-API: YouTube operational API works when YouTube Data API v3 fails*. Accessed: 2023/9/15. [Online]. Available: https://github.com/Benjamin-Loison/YouTube-operational-API

[١٨] *NeMo: NeMo :a toolkit for conversational AI*. Accessed: 2023/9/15. [Online]. Available: https://github.com/NVIDIA/NeMo

[١٩] T. Wang, J. Zhang, F. Zheng, W. Jiang, R. Cheng, and P. Luo, "Learning Grounded Vision-Language Representation for Versatile Understanding in Untrimmed Videos," *ArXiv,* vol. abs/2303.06378, 2023.

[٢٠] H. Alwassel, S. Giancola, and B. Ghanem, "TSP: Temporally-Sensitive Pretraining of Video Encoders for Localization Tasks," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* ,pp. 3166-3176, 2020.

[٢١] J. Louradour, "whisper-timestamped," *GitHub repository,* 2023 2023. [Online]. Available: https://github.com/linto-ai/whisper-timestamped.

Abstract

Video captioning is one of the complex tasks in image processing, which is used in various fileds such as automatic title generation, human-computer interaction, helping disabled people, and simplifying educational videos in the form of text. The dense video captioning aims to localize important events in the video and describe them with short sentences. The goal of this research is to collect a dataset that can be used to train deep learning, computer vision, and natural language processing models in this field. These models receive a short video as input, then extract its important parts as output and explain it in one sentence.

To date, various datasets have been collected for training dense video captioning models. Most of them contain fewer than 10k videos, also they are manually labeled. For example, the YouCook dataset, which is one of the most popular datasets in this field, contains about 2000 videos, each manually labeled by a large group of experts.

Most of this research has been focused on automatic labeling so that we can collect larger datasets in a shorter time. Our videos are collected from larger datasets such as HowTo100M[2] and their videos are also collected from YouTube. In this research, two methods for data labeling will be presented. The first method is using the raw description provided by YouTube for videos. The second one is using raw subtitles of the videos. Finally, a VDC model will be trained using the collected data so that the impact of the data can be compared with other datasets.

Keywords: dense video captioning, automatic labeling, YouTube

Iran University of Science and Technology
School of Computer Engineering

Data collection and automatic labeling for dense video captioning models

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Bachelor of Science in Computer Engineering

By:
Amirhossein Ahmadi
Mohamad Sadra Khamooshifar

Supervisor:
Dr. Behrouz Minaei-Bidgoli
Dr. Issa Zarepour

September 2023