توصیف‌گر ترکیبی ژرف

سامان جهانگیری

چکیده

با وجود این‌که شبکه های عصبی که اخیراً در توصیف تصویر به کار گرفته شده اند، به نتایج قابل قبولی دست یافته اند، آن‌ها عمدتاً به منابع جفت عکس و جمله[۱] نیازمندند. در این‌جا، توصیف‌گر ترکیبی ژرف را معرفی خواهیم کرد تا روشی برای توصیف اشیاء جدیدی که در داده های جفت عکس ـ جمله وجود ندارند را ارائه کنیم. این کار، با استفاده از مجموعه داده های بزرگ مربوط به تشخیص اشیاء و منابع متنی غیر مرتبط، و با انتقال اطلاعات بین مفاهیم مشابه انجام می‌شود. سایر مدلهای توصیف‌گر ژرف فعلی، با وجود این‌که با مجموعه داده های بزرگ تشخیص چهره، مانند ImageNet، پیش آموزش[۲] داده می‌شوند، تنها می‌توانند اشیایی را توصیف کنند که در منابع جفت عکس ـ جمله حضور دارند. برعکس، این مدل می‌تواند اشیاء جدید و روابط آنها با سایر اشیاء را توصیف کند. توانمندی مدلمان برای توصیف اشیاء جدید را با برآورد کردن عملکرد آن روی MSCOCO و نمایش کمی نتایج برای عکسهای ImageNet که هیچ داده جفتی از اشیاء داخل آنها موجود نیست نشان می‌دهیم. نتایج نشان می‌دهد که DCC مزایای ویژه ای نسبت به روشهای موجود توصیف اشیاء جدید داخل عکس دارد.

۱ مقدمه

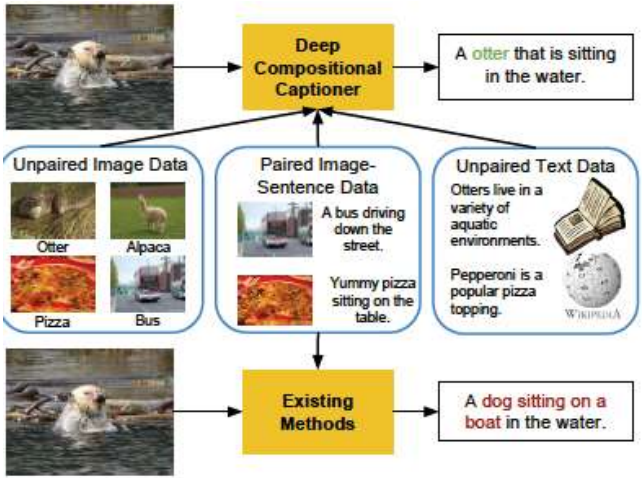

Figure ۱: مدل با ترکیب اطلاعات مربوط به داده های تصویری و منابع متنی مستقل، توصیفی برای عکسهای دیده نشده در منابع عکس ـ جمله ارائه می‌دهد.

در سالهای اخیر، شبکه های عصبی بازگشتی[۳] متعددی برای توصیف عکس معرفی شده اند که نتایج قابل قبولی به دست آورده اند [۲۹][۲۷][۳۲][۳۰][۳۶]. منابع گسترده جفت عکس و توصیف مانند MSCOCO و Flickr30k ، از عوامل مهم موفقیت این روشها به شمار می‌روند. با این وجود، این مجموعه داده ها شامل تعداد کمتری از اشیاء به نسبت مجموعه داده های مربوط به

[۱] Paired image-sentence data
[۲] Pre-train
[۳] Recurrent Neural Networks

تشخیص عکس، مانند ImageNet هستند.

بنابراین، با وجود این‌که سیستم‌های جدید تشخیص عکس قادر به تشخیص هزاران گونه از اشیاء هستند، بهترین روشهای توصیف عکس، فاقد توانایی ترکیب مفاهیم مختلف برای توصیف اشیاء جدید با استفاده از مفاهیم شناخته شده هستند، بدون اینکه مثالهایی از جفت عکس و جمله‌ی مربوط به آن شیء را دیده باشند. به همین خاطر توصیف‌گر ترکیبی ژرف (DCC) را معرفی خواهیم کرد که میتواند عناصر بصری قسمتهای لغت‌نامه‌ای را ترکیب کند تا توصیفاتی از اشیایی به دست آورد که در داده های جفت ناموجود، اما در منابع تشخیص عکس موجود هستند. عملکرد DCC مشابه مدلهای توصیف‌گر اخیر [۲۹][۲۷][۳۲][۳۰][۳۶] است که شبکه های بازگشتی[۴] و پیچشی[۵] را با هم ترکیب می‌کنند، با این تفاوت که میتواند اشیاء جدید را با ترکیب آنها با عبارات داخل زبان که از داده های جفت یاد گرفته است، توصیف کند. برای مثال تصویر سمور در عکس ۱ را در نظر بگیرید. برای ارائه یک توصیف دقیق، یک مدل توصیف‌گر نیاز دارد تا اجزای بصری تشکیل دهنده عکس مانند "otter"، "water" و "sitting" را تشخیص دهد و آنها را ترکیب کند تا یک جمله قابل قبول به دست آید. با اینکه مدلهای قبلی توصیف‌گرهای ژرف می‌توانند عناصر بصری را ترکیب کرده و جمله بسازند، DCC میتواند توصیفی برای یک شیء جدید مانند "otter" ارائه کند. به این ترتیب که میفهمد که کلمه "otter" مشابه "animal" است و بنابراین به همان روش می‌تواند در جملات مورد استفاده قرار گیرد. این مدل از چند عنصر کلیدی تشکیل شده است. یک دسته‌بندی‌گر لغت‌نامه‌ای[۶] و یک واحد زبانی[۷] که هر یک می‌توانند به طور جداگانه روی عکس و متن جفت نشده آموزش داده شوند، و همچنین توصیف‌گر ژرف که از ترکیب این دو به دست می آید و با داده های جفت عکس_توصیف آموزش داده می‌شود. مهمتر از همه، لایه چندگانه[۸] است که در آن اطلاعات مربوط به اشیاء شناخته شده در داده های جفت عکس_توصیف به اشیاء جدید که تنها در داده های غیر جفت موجودند انتقال داده می‌شود. همچنین از منابع متنی نامرتبط برای ارتباط دادن اشیاء جدید به مفاهیم دیده شده در داده های جفت استفاده می‌شود.

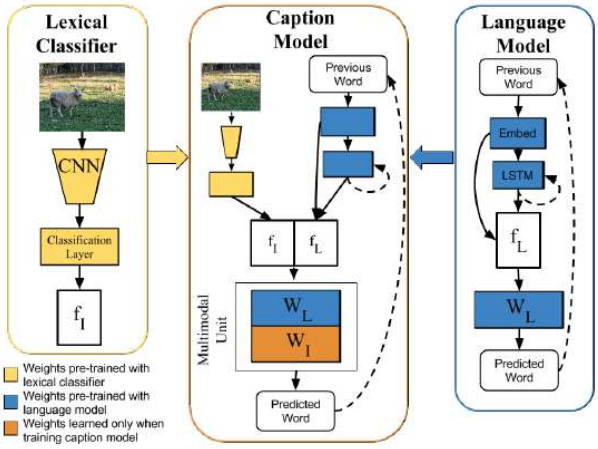

Figure ۲: DCC از یک دسته‌بندی‌گر لغت‌نامه‌ای، که پیکسل ها را به مفاهیم بصری می‌نگارد و فقط با داده های تصویری غیر جفت آموزش داده شده، و یک مدل زبانی که ساختار زبان طبیعی را یاد میگیرد و روی داده های متنی جفت نشده آموزش داده شده، تشکیل شده است. لایه چندگانه DCC، دسته‌بندی‌گر لغت‌نامه‌ای، و مدل زبانی را ترکیب می‌کند و روی داده های جفت عکس_جمله آموزش داده می‌شود.

۲ DCC

روش ما از سه گام تشکیل شده است: ۱) آموزش یک دسته‌بندی‌گر لغت‌نامه‌ای و یک مدل زبان ژرف، ۲) ترکیب دسته‌بندی‌گر لغت‌نامه‌ای و مدل زبان برای ایجاد یک مدل توصیف‌گر، و در نهایت، ۳) انتقال اطلاعات از کلماتی که در داده های جفت شده ظاهر شده اند به کلماتی که در این داده ها دیده نشده‌اند.

[۴] Recurrent
[۵] Convolutional
[۶] Lexical Classifier
[۷] Language Model
[۸] Multimodal Layer

۱.۲ دسته‌بندی‌گر لغت‌نامه‌ای ژرف

دسته‌بندی‌گر لغت‌نامه‌ای (شکل ۲ چپ) یک CNN[۹] است که تصاویر را به مفاهیم معنایی می‌نگارد. برای آموزش دسته‌بندی‌گر، ابتدا مفاهیمی را که در داده های جفت متداول هستند را استخراج می‌کنیم. به این شکل که ابتدا نوع هر کلمه (فعل، اسم، ...) را پیدا می‌کنیم [۳] و سپس متداول‌ترین اسمها، صفتها و فعلها را انتخاب می‌کنیم. هیچ اصلاحی در مفاهیم ذخیره شده انجام نمی‌دهیم و بنابراین بعضی از مفاهیم مانند "use"، صرفاً بصری نیستند. علاوه بر مفاهیم متداول در داده‌های جفت، دسته‌بندی‌گر همچنین با اشیایی که خارج از داده‌های جفت‌شده هستند و میخواهیم توصیفشان کنیم نیز آموزش میبیند. دسته بندی گر با تنظیم[۱۰] کردن یک CNN که با بخش training از مجموعه داده ILSVRC-2012 پیش آموزش داده شده، آموزش داده می‌شود. در توصیف تصاویر، مفاهیم بصری بسیاری بر توصیف اثر می‌گذارند. برای مثال در جمله "An alpaca stands in the green grass"، مفاهیم بصری "alpaca"، "stands"، "grass" و "green" نقش دارند. ما برای نسبت دادن چند برچسب به یک عکس، از cross-entropy sig-moid loss استفاده می‌کنیم. بردار ویژگی عکس که همان خروجی دسته‌بندی‌گر لغت‌نامه‌ای است را با f_I نمایش میدهیم، که هر درایه آن بیانگر احتمال حضور یک مفهوم در عکس است.

۲.۲ مدل زبانی

مدل زبانی (شکل ۲ سمت راست)، ساختار جمله را تنها با استفاده از منابع متنی جفت نشده یاد می‌گیرد و شامل یک لایه نشاندن[۱۱] که نمایش one-hot-vector[۱۲] کلمات را به یک فضای با بعد پایین‌تر می‌نگارد، یک LSTM و یک لایه پیش‌بینی کلمه است. مدل زبانی آموزش می‌بیند که که با استفاده از کلمات قبلی جمله، کلمه بعدی را پیش‌بینی کند. در هر گام، کلمه قبلی ورودی LSTM است که ساختار بازگشتی (وابستگی به قبل) زبان را یاد می‌گیرد. کلمه نشانده شده و خروج LSTM به یکدیگر الحاق میشوند تا بردار ویژگی زبانی، f_L را تشکیل دهند. f_L ورودی یک لایه ضرب داخلی است که خروجی‌اش کلمه بعدی در جمله تولید شده است. در زمان آموزش، همواره کلمه درست به عنوان ورودی به بخش زبانی داده می‌شود اما در زمان تست، کلمه قبلی پیش‌بینی شده توسط خود بخش زبانی به عنوان ورودی به آن داده می‌شود. همچنین با اضافه کردن این محدودیت که مدل نتواند یک کلمه را دو بار پشت سر هم پیش‌بینی کند، نتایج بهبود پیدا می‌کند.

۳.۲ مدل توصیف‌گر

مدل توصیف‌گر دسته بندی گر لغتنامه ای و مدل زبانی را ترکیب میکند تا یک مدل مشترک برای توصیف عکس را یاد بگیرد. همان‌طور که در عکس ۲(وسط) نشان داده شده، واحد چندگانه در مدل توصیف‌گر بردار ویژگی تصویر، f_I، و بردار ویژگی زبانی، f_L را ترکیب میکند. لایه چندگانه مورد استفاده، یک انتقال آفین از این دو بردار ویژگی است:

$$p_w = softmax(f_I W_I + f_L W_L + b)$$

که W_I، W_L و b ماتریس های وزن های یاد گرفته شده هستند و p_w توزیع احتمال روی کلمات است. به طور شهودی، وزنهای W_I یاد میگیرند که مجموعه ای از کلمات را که احتمال بیشتری برای حضور در عکس ــ به شرط وجود عناصر بصری که توسط دسته بندی گر لغت نامه ای داده شده اند ــ دارند را پیش بینی کند. از طرف دیگر، W_L ساختار دنباله ای زبان را یاد میگیرد؛ با یاد گرفتن پیش بینی کلمه بعد با استفاده از کلمات قبل در یک دنباله داده شده. با جمع کردن $f_I W_I$ و $f_L W_L$ ، لایه چندگانه اطلاعات بصری به دست آمده از دسته بندی گر لغتنامه ای را با اطلاعات ساختاری زبان که از مدل زبانی یاد گرفته شده اند ترکیب میکند، که یک جمله منسجم راجع به عکس به دست آورد. هر دوی مدل توصیف‌گر و مدل زبانی آموزش میبینند که دنباله ای از کلمات را پیش بینی کنند،

[۹] Convolutional Neural Network

[۱۰] Fine-tuning

[۱۱] Embedding

[۱۲] در این نمایش به هر یک از کلمات یک اندیس نسبت داده میشود وسپس برداری به کلمه نسبت داده میشود که تمام درایه های آن صفر است به جز درایه اندیس متناظر

در حالی که دسته بندی گر لغتنامه ای آموزش میبیند که مجموعه ای از عناصر بصری احتمالی داخل یک عکس را پیش بینی کند. در نتیجه، وزن های W_L، که ویژگیهای زبانی را به کلمه پیش بینی شده مینگارند، در زمان آموزش مدل زبانی یاد گرفته می شوند، اما وزن های W_I نه. وزن های داخل W_L با استفاده از داده های متنی جفت نشده آموزش داده می شوند؛ قبل از تنظیم شدن با داده های جفت عکس‌ـجمله. اما W_I فقط با داده های عکس‌ـجمله آموزش داده می شود. با اینکه در این‌جا از یک لایه چندگانه خطی استفاده میشود، نتایج به دست آمده با نتایج به دست آمده توسط سایر مدلها که شامل یک لایه غیر خطی برای پیش بینی کلمه هستند، قابل مقایسه است. مدل توصیف‌گر به گونه ای طراحی شده است که انتقال وزن های یادگرفته شده مربوط به کلمات موجود در داده های جفت، به وزن های مربوط به کلمات غایب در این مجموعه ها، به آسانی انجام شود. در ابتدا، با استفاده از دسته بندی گر لغتنامه ای، ویژگیهای عکس، دارای مفهوم مشخصی می شوند. بنابراین می توان به سادگی این ویژگی ها را گسترش داد تا شامل اشیاء جدید شوند و وزنهای مربوط به اشیاء مشخصی را تغییر داد. سپس، با یادگیری ویژگیهای زبانی با استفاده از داده های متنی جفت نشده، مطمئن میشویم که مدل یک نشاندن خوب برای کلمات غایب پیدا میکند. در انتها، با استفاده از یک لایه چندگانه خطی، وابستگی بین ویژگیهای زبانی و ویژگیهای عکس و کلمه پیش بینی شده، به سادگی به دست می آید.

۴.۲ انتقال اطلاعات بین اشیاء

انتقال مستقیم

اولین روش انتقال وزن بین اشیاء که به بررسی آن میپردازیم، به طور مستقیم، وزنهای داخل W_I، W_L و b را از اشیاء دیده شده در داده های جفت به اشیاء دیده نشده انتقال میدهد. به طور شهودی، روش انتقال مستقیم فرض میکند که یک کلمه جدید، مانند کلمه ای که از نظر معنایی به آن نزدیک است، توصیف می شود. برای مثال کلمه جدید "alpaca" که از نظر معنایی به کلمه شناخته شده "sheep" نزدیک است را در نظر بگیرید. فرض کنید v_a و v_s نشان دهنده اندیس های کلمات "alpaca" و "sheep" در مجموعه کلمات ما باشند. با داشتن بردارهای ویژگی عکس و ویژگی زبانی، f_I و f_L، احتمال پیش بینی کلمه "sheep" متناسب است با:

$$f_I W_I[:, v_s] + f_L W_L[:, v_s] + b[v_s]$$

برای ساختن جمله با "alpaca" به همان شکل که جمله های شامل "sheep" ساخته می شوند، ابتدا وزنهای $W_I[:, v_s]$ ، $W_L[:, v_s]$ و $b[v_s]$ (نشان داده شده با قرمز در شکل ۳) را به طور مستقیم به وزنهای $W_I[:, v_a]$ ، $W_L[:, v_a]$ و $b[v_a]$ انتقال میدهیم. علاوه بر این، انتظار داریم پیش بینی کلمه "sheep" وابستگی زیادی به احتمال حضور یک "sheep" در تصویر داشته باشد. به عبارت دیگر، انتظار داریم $W_I[:, c_s]$ نقش تعیین کننده ای در بخشی از خروجی دسته بندی گر لغتنامه ای که مربوط به "sheep" است داشته باشد. از طرف دیگر، $W_I[:, c_a]$ باید در خروجی مربوط به کلمه "alpaca" تاثیر گذار باشد. برای ایجاد این شرایط قرار میدهیم: $W_I[r_a, c_a] = W_I[r_s, c_s]$ که r_a و r_s نشان دهنده اندیس های "sheep" و "alpaca" در بردار ویژگیهای عکس هستند. در نهایت، انتظار نداریم که خروجی کلمه "alpaca" به حضور یک "sheep" در تصویر وابسته باشد و یا برعکس. پس قرار میدهیم $W_I[r_s, c_a] = W_I[r_a, c_s] = 0$

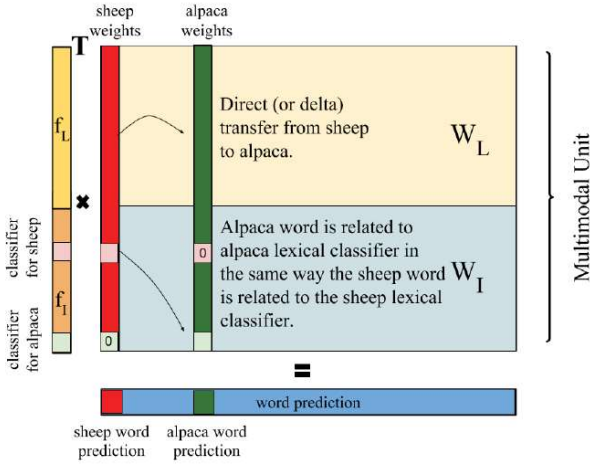

Figure ۳: شیوه انتقال اطلاعات از اشیاء دیده شده در منابع جفت عکس‌ـجمله به اشیاء دیده نشده. برای جزییات بیشتر بخش ۳,۴ را ببینید.

انتقال دلتا

به جای انتقال مستقیم وزن ها،میتوانیم تغییر وزنها در طول آموزش با داده های جفت شده را انتقال دهیم. دوباره، انتقال کلمه "sheep" به کلمه "alpaca" را درنظر بگیرید. Δ_L را به این صورت تعریف میکنیم:

$$\Delta_L = W_{L-caption}[:, v_s] - W_{L-language}[:, v_s]$$

که $W_{L-caption}$ وزنهای یاد گرفته شده در زمان آموزش با عکس و جمله هستند و $W_{L-language}$ وزنهای یاد گرفته شده در مدل زبانی هستند. وزنهای مربوط به کلمه جدید "alpaca" به این صورت اصلاح میشوند:

$$W_{L-caption}[:, v_a] = W_{L-language}[:, v_a] + \Delta_L$$

انتقال دلتا میتواند بهتر عمل کند، زیرا برخلاف انتقال مستقیم، اطلاعات مربوط به وزنهای داخل W_L در طول انتقال از بین نمیرود. هنگامی که برای انتقال وزنهای W_L، از انتقال دلتا استفاده میکنیم نیز، همچنان از انتقال مستقیم برای انتقال در W_I استفاده میکنیم.

تشخیص شباهت بین مفاهیم

تعیین اینکه کدام کلمات در داده های جفت عکس‌ـجمله از نظر معنایی به کلمات خارج از داده های جفت شباهت دارند، یکی از مراحل مهم برای انتقال است. ما شباهت معنایی را با استفاده از مدل word2vec[۱۴] پیدا میکنیم. که با British National Corpus(BNC)، UkWaC و Wikipedia آموزش داده شده اند و فاصله بین دو کلمه (در فضای نشانده شده) با کسینوس زاویه بین آنها تعریف می شود.

۳ چارچوب عملی

۱.۳ مجموعه داده ها

برای اینکه در عمل برآوردی از عملکرد روشمان به دست آوریم، زیر مجموعه ای از مجموعه آموزش[۱۳] MSCOCO میسازیم که فاقد هرگونه جفت عکس جمله است که حداقل یکی از هشت شیء خاص را توصیف میکنند. برای اینکه مطمئن شویم که اشیاء حذف شده شبیه حداقل یکی از اشیاء حذف نشده هستند، ۸۰ شیء مربوط به چالش مرزبندی MSCOCO[۱۴] را با استفاده از نشاندن word2vec خوشه بندی[۱۵] میکنیم و از هر خوشه یک شیء را حذف میکنیم. برای این کار، این کلمات انتخاب شدند: "bottle"،"bus"، "couch" ، "microwave"،"pizza" ، "racket"، "suitcase"، "zebra".

به طور تصادفی ۵۰ درصد از مجموعه ارزیابی[۱۶] MSCOCO را برای ارزیابی انتخاب میکنیم و ۵۰ درصد دیگر را برای تست باقی میگذاریم. مفاهیم بصری داخل هر عکس را با استفاده از پنج موضوع اصلی قطعی[۱۷] اشاره شده که در مجموعه داده MSCOCO در اختیار قرار داده شده است، برچسب گذاری میکنیم. اگر یکی از توضیحات قطعی مربوط به یک عکس، به یک شیء اشاره کنند، آن عکس یک مورد دارای آن شیء به حساب میاید. علاوه بر این، عملکرد DCC را هنگامی که شبکه های زبانی و دسته بندی گر بر روی مجموعه داده های جداگانه آموزش داده میشوند نیز بررسی میکنیم. به طور خاص، ۶۴۲ شیء از مجموعه داده تشخیص اشیاء ImageNet را، که در MSCOCO وجود ندارند، اما در مجموعه داده های متنی به دست آمده از صفحات وب وجود دارند، را انتخاب میکنیم. هیچ اصلاحی روی این مفاهیم به دست آمده انجام نمیدهیم و در نتیجه برخی از کلاسهای اشیاء شامل تعداد کم و برخی دیگر شامل تعداد زیادی عکس هستند. از 75 درصد عکسهای هر کلاس برای آموزش دسته بندی گر لغتنامه ای و از بقیه برای ارزیابی استفاده میکنیم. دقت کنید که هیچ توصیفی از این دسته ها نداریم.

[۱۳] Training set
[۱۴] MSCOCO segmentation challenge
[۱۵] cluster
[۱۶] Validation set
[۱۷] Ground truth caption annotations

۲.۳ آموزش دسته بندی گر لغتنامه ای

ما هر دوی MSCOCO و ImageNet را به عنوان داده های عکسهای برچسب گذاری شده، برای آموزش دسته بندی گر لغتنامه ای در نظر میگیریم. برای تمام اشیاء داخل داده های جفت عکس‌ـ جمله، از عکسهای COCO که با ۴۷۱ مفاهیم بصری برچسب گذاری شده اند، برای آموزش دسته بندی گر لغتنامه ای استفاده میکنیم. برای هشت شیءی که در داده های جفت ظاهر نمیشوند هم آموزش با MSCOCO و هم آموزش با ImageNet را در نظر میگیریم. دسته بندی گر لغتنامه ای با تنظیم یک شبکه پیچشی ژرف (VGG-16 layer)[35]، که بر روی زیرمجموعه آموزش مربوط به تشخیص اشیاء ILSVRC-2012 از مجموعه ImageNet آموزش داده شده، آموزش میبینند.

۳.۳ آموزش مدل زبانی

سه منبع مختلف برای داده های متنی جفت نشده، برای آموزش مدل زبانی در نظر میگیریم: ۱)توصیفات داخل مجموعه آموزش MSCOCO ۲)متن های مربوط به توصیف عکس شامل سایر مجموعه داده های جفت عکس‌ـ توصیف: Flickr1M[5]، Flickr30k[21]، Pascal1k[8] و ImageCLEF-2012 . این منبع شامل جملات MSCOCO نیست. ۳)متن های غیر مرتبط از صفحات وب که شامل ۶۰ میلیون جمله از British National Cor-pus (BNC)، UkWaC و Wikipedia است.

۴.۳ آموزش مدل توصیف‌گر

پس از آموزش دسته بندی گر لغتنامه ای و مدل زبانی، وزنهای لایه چندگانه مدل توصیف‌گر با داده های جفت عکس‌ـ جمله آموزش داده میشوند. برای روش انتقال مستقیم، تنها وزنهای داخل مدل چندگانه (W_I و W_L) را یاد میگیریم و سایر وزنها را ثابت نگه میداریم. برای روش انتقال دلتا، اگر وزنهای داخل W_L، که با مدل زبانی پیش آموزش داده شده اند، از مقادیر اصلیشان زیاد تغییر کنند، انتقال خوب عمل نمیکند. در نتیجه، ابتدا وزنهای داخل W_L را ثابت نگه میداریم و فقط W_I را یاد میگیریم و سپس W_I و W_L را همزمان آموزش میدهیم.

Lexical classifier	Language model	B-1	METEOR	F1
MSCOCO	MSCOCO	64.40	21.00	39.78
Imagenet	MSCOCO	64.00	20.71	33.60
Imagenet	CaptionTxt	64.79	20.66	35.53
Imagenet	WebCorpus	64.85	20.66	34.94

جدول ۱: DCC بدون انتقال را با DCC با انتقال دلتا (ΔT) ، DCC با انتقال مستقیم (DT) و یک مدل قدرتمند دیگر برای توصیف عکس (LRCN) مقایسه میکنیم. توانایی مدل برای برای به کارگیری کلمات جدید در جمله را با امتیاز F1 اندازه میگیریم. همچنین امتیازهای Bleu-1 و METEOR را گزارش میکنیم که بیانگر کیفیت کلی جمله هستند. DCC با موفقیت کلمات جدید را به کار میگیرد و کیفیت جملات آن به نسبت بالاتر است. (مقادیر به درصد هستند).

۵.۳ محک ها

برای اینکه روش انتقال را ارزیابی کنیم، باید یک محک که نشان دهد یک جمله تولید شده شامل یک شیء جدید هست یا نه را انتخاب کنیم. محک های متداول توصیف عکس عبارت اند از BLEU[2] و METEOR[4] که کلیت معنا و فصیح بودن جمله را میسنجند. با این وجود برای خیلی از اشیاء جدید میتوان بدون اشاره به آنها به امتیازات خوبی در این دو رسید. (برای مثال پسر در حال تنیس بازی کردن در شکل ۴) برای ارزیابی دقیق توانایی مدل در به دست آوردن مجموعه کلمات جدید، امتیاز F1 را نیز درنظر میگیریم. امتیاز F1، غلطهای مثبت (هنگامی که کلمه ای در یک جمله ظاهر میشود که نباید بشود)، غلطهای منفی (هنگامی که کلمه ای در جمله ظاهر نمیشود و باید بشود) و درست های مثبت(هنگامی که کلمه ای در جمله ظاهر میشود و باید بشود) را در نظر میگیرد. یک جمله تولید شده را مثبت (مثال مثبتی از یک کلمه خاص) میگیریم اگر شامل حداقل یک اشاره به یک کلمه کنار گذاشته شده داشته باشد و جملات قطعی را مثبت میگیریم، اگر به یک کلمه در یکی از توصیفهای قطعی آن اشاره شده باشد. مدلها با

	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	average
Pair Supervision	23.20	72.07	50.60	39.48	77.07	38.52	46.50	91.02	54.81
DT	4.63	29.79	45.87	28.09	64.59	52.24	13.16	79.88	39.78

جدول ۲: مقایسه امتیاز F1 برای انتقال مستقیم مدل DCC (DT) ویک مدلی که با مثالهای جفت عکس‌ـجمله برای همه اشیاء آموزش داده شده است.

استفاده از Caffe آموزش داده میشوند.

۴ نتایج

همان‌طور که در شکل ۴ نشان داده شده، DCC میتواند کلمات جدید را به شکل درست وارد توصیف کند.

انتقال مستقیم در مقایسه با انتقال دلتا

جدول ۱ میانگین امتیاز F1 در بین هشت شیء کنار گذاشته شده را در حالت انتقال مستقیم و غیر مستقیم مقایسه میکند. علاوه بر این LRCN([5]) را نیز با مجموعه داده کنار گذاشته شده MSCOCO آموزش میدهیم. نتایج نشان میدهد که مدل ما در حالت بدون انتقال، با LRCN قابل مقایسه است و با انتقال، طبق هر محکی بهتر عمل میکند. همان‌طور که امتیاز F1 در جدول ۱ نشان میدهد، هر دوی انتقال دلتا و انتقال مستقیم میتوانند کلمات جدیدی را وارد دامنه لغات خود کنند. همچنین امتیاز BLEU-1 که بیانگر میزان هم پوشانی کلمات تولید شده وکلمات جمله مرجع است را نیز گزارش میکنیم. با اندازه گیری امتیاز METEOR، از فصاحت جملاتمان پس از اضافه شدن کلمات جدید، اطمینان حاصل میکنیم. DCC همواره امتیاز METEOR را افزایش میدهد که این نشان دهنده افزایش کیفیت کلی جملات با DCC است. روش انتقال مستقیم، همه امتیازات را به مقدار بیشتری نسبت به انتقال دلتا افزایش میدهد. یک نکته مهم این است که امتیازهای BLEU و METEOR برای اشیاء داخل مجموعه داده کنار گذاشته شده ها، کاهش پیدا نمیکند. برای روشن شدن اینکه مدل ما برروی کدام کلمات بهتر کار میکند، امتیاز F1 را برای تمام اشیاء جدول ۲ گزارش میکنیم و با مدلی که با جفت عکس‌ـجمله برای هشت شیء کنار گذاشته شده، آموزش داده شده مقایسه میکنیم. برای تمام اشیاء، DCC میتواند جملاتی تولید کند که شامل آن شیء هستند.

Lexical classifier	Language model	B-1	METEOR	F1
MSCOCO	MSCOCO	64.40	21.00	39.78
Imagenet	MSCOCO	64.00	20.71	33.60
Imagenet	CaptionTxt	64.79	20.66	35.53
Imagenet	WebCorpus	64.85	20.66	34.94

شکل ۴: مقایسه اثر پیش آموزش دسته بندی گر لغتنامه ای و مدل زبانی با مجموعه داده های غیر جفت عکس و متن. همان‌طور که انتظار میرود بهترین نتیجه زمانی به دست می آید که از داده های داخل دامنه MSCOCO برای آموزش هر دوی دسته بندی گر لغتنامه ای و مدل زبانی استفاده شود. با این وجود، نتایج آموزش با منابع غیر مرتبط نیز قابل مقایسه است. (مقادیر به درصد هستند.)

۵ جمع‌بندی

در این مقاله توصیف‌گر ترکیبی ژرف (DCC) را معرفی کردیم که میتواند برای توصیف اشیاء جدید که در منابع فعلی توصیفی حضور ندارند به کار رود. نتایج کمی و کیفی ما نشان دهنده توانایی این مدل برای استفاده از کلمات جدید در توصیف عکس با استفاده موثر از مجموعه داده های مربوط به عکس و داده های متنی غیر جفت است. با ترکیب داده های منابع مختلف و انتقال اطلاعات بین مفاهیمی که از نظر معنایی نزدیک هستند، DCC با تولید توصیفهایی غنی که محدود به در دسترس بودن منابع جفت عکس‌ـجمله نیستند، مدلهای فعلی توصیف عکس را بهبود میبخشد.

شکل ۵: مقایسه توصیفهای تولید شده توسط مدل بدون انتقال، DCC با آموزش داخل دامنه، (MSCOCO) با آموزش خارج از دامنه، (ImageNet, WebCorpus) و یک مدل آموزش داده شده با جفت عکس‌ـ جمله برای تمام اشیاء. DCC MSCOCO میتواند جملاتی مشابه حالت آموزش دیده شده با جفت عکس‌ـ جمله تولید کند.

References

[1] S. Hochreiter and J. Schmidhuber. “Long short-term memory”. In: *Neural computation* (1997).

[2] W.-J. Zhu. “BLEU: a method for automatic evaluation of machine translation”. In: *ACL* (2002).

[3] Y. Singer. “Feature-rich part-of-speech tagging with a cyclic dependency network”. In: *NAACL* (2003).

[4] S. Banerjee and A. Lavie. “An automatic metric for mt evaluation with improved correlation with human judgments”. In: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (2005).

[5] M. J. Huiskes and M. S. Lew. “The mir flickr retrieval evaluation”. In: *Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval, New York, NY, USA, 2008. ACM* (2008).

[6] L. Davis. “Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos”. In: *CVPR* (2009).

[7] L. Fei-Fei. “What does classifying more than 10,000 image categories tell us?” In: *ECCV* (2010).

[8] J. Hockenmaier. “Collecting image annotations using amazon’s mechanical turk”. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk* (2010).

[9] B. Schiele. “What helps Where - and Why? Semantic Relatedness for Knowledge Transfer”. In: *CVPR* (2010).

[10] D. L. Chen and W. B. Dolan. “Collecting highly parallel data for paraphrase evaluation”. In: *ACL* (2011).

[11] D. Parikh and K. Grauman. “Relative attributes”. In: *ICCV* (2011).

[12] B. Thomee and A. Popescu. “Overview of the imageclef 2012 flickr photo annotation and retrieval task”. In: *CLEF (Online Working Notes/Labs/Workshop)* (2012).

[13] T. Berg. “Babytalk: Understanding and generating simple image descriptions”. In: *TPAMI* (2013).

[14] J. Dean. “Efficient estimation of word representations in vector space”. In: *ICLR Workshop* (2013).

[15] “Devise: A deep visual-semantic embedding model”. In: *NIPS* (2013).

[16] A. Ng. “Zero-shot learning through cross-modal transfer”. In: *NIPS* (2013).

[17] K. Saenko. “Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition”. In: *ICCV* (2013).

[18] S. Guadarrama. “Generating natural-language video descriptions using text-mined knowledge”. In: *AAAI* (2013).

[19] H. Nickisch C. Lampert and S. Harmeling. “Attributebased classification for zero-shot visual object categorization”. In: *TPAMI* (2014).

[20] T. Darrell. “Caffe: Convolutional architecture for fast feature embedding”. In: *In Proceedings of the ACM International Conference on Multimedia* (2014).

[21] J. Hockenmaier. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *TACL* (2014).

[22] “Imagenet large scale visual recognition challenge”. In: *IJCV* (2014).

[23] R. J. Mooney. “Integrating language and vision to generate natural language descriptions of videos in the wild”. In: *COLING* (2014).

[24] K. Saenko. “LSDA: Large scale detection through adaptation”. In: *NIPS* (2014).

[25] C. L. Zitnick. “Microsoft coco: Common objects in context”. In: *ECCV* (2014).

[26] M. Rohrbach A. Rohrbach and B. Schiele. “The long-short story of movie description”. In: *GCPR* (2015).

[27] T. Darrell. “Long-term recurrent convolutional networks for visual recognition and description”. In: *CVPR* (2015).

[28] A. Dick. “Image captioning with an intermediate attributes layer”. In: *arXiv preprint arXiv:1506.01144* (2015).

[29] D. Erhan. “Show and tell: A neural image caption generator”. In: *CVPR* (2015).

[30] A. Karpathy and L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. In: *CVPR* (2015).

[31] M. Mitchell. “Language models for image captioning: The quirks and what works”. In: *ACL* (2015).

[32] R. Salakhuditnov R. Kiros and R. S. Zemel. “Unifying visual-semantic embeddings with multimodal neural language models”. In: *TACL* (2015).

[33] K. Saenko. “Sequence to sequence - video to text”. In: *ICCV* (2015).

[34] K. Saenko. “Translating videos to natural language using deep recurrent neural networks”. In: *NAACL* (2015).

[35] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *ICLR* (2015).

[36] A. L. Yuille. “Learning like a child: Fast novel visual concept learning from sentence descriptions of images”. In: *ICCV* (2015).

[37] G. Zweig. “From captions to visual concepts and back”. In: *CVPR* (2015).