
بسم الله الرحمن الرحيم

دانشگاه شهید بهشتی

پژوهشکده فضای مجازی

ارائه روشی برای تشخیص جعل تصویر چهره بر پایه یادگیری بدون نمونه

پایان‌نامه کارشناسی‌ارشد فناوری اطلاعات گرایش چند رسانه‌ای

سید مصطفی حسینی کشکوئیه

استاد راهنما

دکتر احمد محمودی ازناوه

۱۴۰۰

دانشگاه شهید بهشتی
پژوهشکده فضای مجازی

پایان‌نامه کارشناسی ارشد – آقای سید مصطفی حسینی کشکوئیه

تحت عنوان

ارائه روشی برای تشخیص جعل تصویر چهره بر پایه یادگیری بدون نمونه

در تاریخ توسط کمیته‌ی تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت:

۱ - استاد راهنمای پایان‌نامه	دکتر
۳- استاد داور	دکتر
۴- استاد داور	دکتر
سرپرست تحصیلات تکمیلی دانشکده	دکتر

تشکر و قدردانی

پروردگار منان را سپاسگزارم.

فهرست مطالب

عنوان **صفحه**

فهرست تصاویر

فهرست جداول

چکیده

جمع آوری داده و پرچسب زدن داده‌ها از مهم‌ترین مراحل پیش‌پردازش در یادگیری ماشین است.اما همواره جمع‌آوری به سادگی نیست و ما به داده‌های برچسب‌خورده دسترسی نداریم. عدم دسترسی به دادگان مناسب، هزینه زیاد جمع آوری دادگان و مشاهده نمونه‌های جدید از عوامل حرکت به سمت یادگیری با نمونه کم، یک نمونه و در نهایت بدون نمونه بوده‌اند.

در سیستم‌های تشخیص چهره، یکی از ارکان اساسی برای سلامت و کارایی سیستم توانایی جلوگیری از حملات مختلف و تشخیص جعل تصویر است. راهکارها و روش‌های بسیاری در طول سالیان مختلف ارائه شده است؛ اما نکته حائز اهمیت تغییر کردن و به‌روز شدن حملات است که باعث می‌شود این کشمکش بین حملات و سیستم‌های تشخیص حملات و جعل تصاویر همواره وجود داشته باشد.

از روش‌های پر کاربردی که برای تولید تصاویر جعلی و انجام حملات به سیستم‌های تشخیص چهره استفاده می‌شود روش‌های تخاصمی و استفاده از شبکه‌های متولد متخاصم است. جلوگیری از این حملات و نیازمند روش‌های هوشمند و به‌روزی است به همین به سراغ استفاده از روش یادگیری بدون نمونه رفته‌ایم تا بتوانیم با استفاده از آن از حملات جلوگیری کنیم و جعل تصویر چهره را تشخیص دهیم.

واژه‌های کلیدی: ۱ـ یادگیری ماشین ۲ـ جعل تصویر چهره ۳ـ شبکه‌های مولد تخاصمی ۴ـ یادگیری بدون نمونه

فصل اول

مقدمه

تشخیص چهره یکی از روش‌های احراز هویت بیومتریک[1] است که در قدیم توسط سیستم‌های امنیتی پیشرفته انجام می‌شد. در روش‌های بیومتریک از اثر انگشت یا کل دست، عنبیه چشم، صدا، چهره و غیره استفاده می‌شود که متداول‌ترین آن استفاده از چهره فرد است. امروزه با توجه به پیشرفت تکنولوژی و همه‌گیر شدن آن تشخیص چهره را در گوشی‌های هوشمند همراه می‌توان یافت.

همه گیری این تکنولوژی باعث پیشرفت‌های فراوان آن نیز شده است و امروزه گوشی‌های همراه می‌توانند به سرعت و با دقت بالا تشخیص چهره را انجام دهند و تقریبا به بهترین شکل اینکار را انجام می‌دهند. اما به امنیت این سیستم‌ها به اندازه کافی توجه نشده‌است. به این سیستم‌ها حملات متعددی صورت می‌گیرد که می‌توانند باعث خرابی سیستم، جعل هویت و یا گرفتن دسترسی کامل توسط حمله کننده و هکر شود. این سیستم‌ها به طور عمده

[1] biometric authentication

امروزه بر اساس شبکه‌های عصبی عمیق یا DNN[1] و شبکه‌های عصبی کانولوشن یا CNN[2] ساخته می‌شوند. در مقالات زیادی انواع حملات به این نوع شبکه‌ها بررسی شده‌اند.

تشخیص حملات و مقابله با آن‌ها یکی از مواردی است که باید به آن توجه ویژه‌ای داشت. یکی از مشکلات اساسی در حملات و جعل تصاویر صورت گرفته دسته بندی حملات و تشخیص نوع حمله است که به آسانی توسط انسان امکان‌پذیر نیست.

۱-۱ صورت مسئله

تشخیص و دسته‌بندی حملات به سیستم‌های تشخیص چهره و جعل تصاویر چهره از بزرگ‌ترین معزلات توسعه دهندگان این سیستم‌هاست. تضمین امنیت سیستم برای جلوگیری از دور زدن سیستم، گرفتن دسترسی‌های بیشتر و یا احراز هویت به جای فرد دیگری توسط هکرها از مواردی است که توسعه دهندگان باید مد نظر داشته باشند. با توجه به اینکه به طور عمده حملات در دسته‌های مشخصی قرار می‌گیرند و ویژگی‌هایی دارند که می‌توان آن‌ها را دسته‌بندی کرد استفاده از روش‌های دسته‌بندی به شیوه‌های مختلف می‌تواند کارآمد و کمک‌کننده باشد.

برای دسته‌بندی باید بتوانیم ویژگی‌های حملات مختلف را تشخیص دهیم. تشخیص این ویژگی‌ها به سادگی تشخیص ویژگی‌های تصاویر حیوانات مختلف برای دسته‌بندی آن‌ها نیست و ویژگی‌ها پیچیده‌تر و نیازمند سیستم‌های تشخیص دقیق‌تری هستند. بعد از تشخیص ویژگی‌های مورد نظر می‌توان حمله را دسته‌بندی کرد و متناسب با نوع حمله با آن مقابله کرد.

۱-۲ اهداف پژوهش

تجربه نشان داده است که یادگیری بدون نمونه برای دسته‌بندی و تشخیص نمونه‌هایی که کمتر تا کنون دیده شده‌اند یا اصلا دیده نشده‌اند، موفق عمل کرده و توانسته دسته‌بندی را به نحو احسنت انجام دهد. استفاده از

[1] Deep Neural Networks
[2] Convolutional Neural Network

یادگیری بدون نمونه در مواردی که پیدا کردن ویژگی‌های مشترک مانند حیوانات و اشیاء پیرامون به سادگی نیست و نمی‌توان فهرستی از ویژگی‌های مشترک را به آسانی تهیه کرد؛ موفق عمل کرده است. هدف پژوهش استفاده از دادگان‌های موجود و پر کاربرد در حوزه جعل تصویر چهره و یادگیری بدون نمونه برای ارائه راهکاری که قابلیت تشخیص و دسته‌بندی حملات مختلف به سیستم‌های تشخیص چهره را با دقتی نزدیک به دقت تشخیص چهره دارد، است.

۱-۳ نوآوری پژوهش
۱-۴ ساختار پایان‌نامه

در فصل دوم، به بررسی اجمالی درباره یادگیری بدون نمونه و انواع آن می‌پردازیم و همچنین حملات و جعل‌های مختلف تصویر چهره را بررسی می‌کنیم. در فصل سوم، کارهای مرتبط به این پژوهش بررسی می‌شوند و مزایا و معایب آن‌ها به طور مختصر بررسی میگردد. در فصل چهارم، به تفضیل روش پیشنهادی و ایده اصلی بیان شده است. نحوه پیاده‌سازی و ساختار مدل بررسی شده‌است. در فصل ششم نتایج و ارزیابی‌های مرتبط بررسی شده و میزان کارایی و نتایج به‌دست آمده مورد تحلیل قرار گرفته‌است و در نهایت در فصل هفتم جمع‌بندی پایانی صورت گرفته‌است.

فصل دوم

پیش‌زمینه

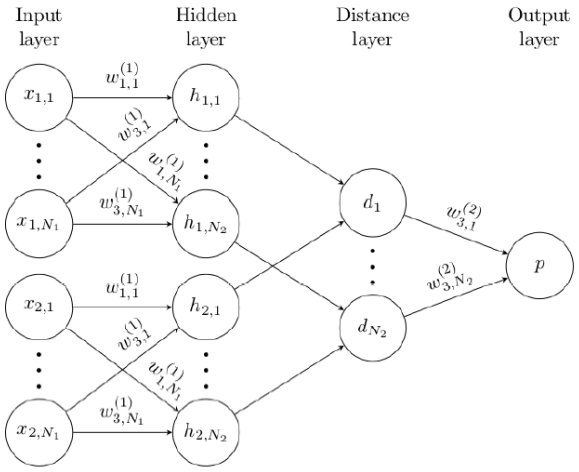
۲-۱ یادگیری با یک نمونه

در یادگیری به صورت سنتی، باید داده‌ها با دو دسته آموزش و آزمون تقسیم‌بندی شوند و برای هر کلاس جدید برای دسته‌بندی باید به تعداد کافی داده وجود داشته باشد تا مدل بتواند دسته‌بندی مناسبی انجام دهد. مانند مدل کانولوشن ساده‌ای که با آن می‌توان سگ و گربه را از هم تفکیک داد. این مدل برای هر داده سگ و گربه باید تعداد قابل توجهی داده دیده باشد تا بتواند دسته‌بندی را به درستی انجام دهد. حال برای دسته‌بندی یک حیوان دیگر توسط این مدل باید داده‌های جدیدی به مدل بدهیم تا مدل بتواند آن نوع حیوان را نیز دسته‌بندی کند.

با بهبود ساختار شبکه یادگیرنده می‌توان نیاز فرآیند یادگیری به داده‌های زیاد را کاهش داد تا مدل بتواند با داده‌های کم نیز دسته‌بندی را انجام دهد. ارائه راهکارها و روش‌هایی در کنار استفاده از شبکه‌های عصبی این امکان را به وجود آورد تا مدل فقط با دیدن یک نمونه از یک کلاس داده در فرآیند آموزش بتواند داده جدید را

دسته‌بندی کند. این روش با نام یادگیری با یک نمونه شناخته می‌شود.

در این روش در فرآیند آموزش مدل به حداقل یک داده از یک کلاس داده نیاز دارد. برای جلوگیری از بیش‌برازش[1] در فرآیند آموزش از روش‌هایی مانند ایجاد اعوجاج در داده‌ها برای تولید داده جدید در حالات مختلف استفاده می‌شود تا مدل بتواند دسته‌بندی بهتری انجام دهد. فرآیند یادگیری با یک نمونه در مواردی پر کاربردتر است که ما تمامی کلاس‌های داده را از قبل داشته‌باشیم؛ برای مثال: در مبحث پردازش زبان طبیعی انسان، ساختار هر زبان مشخص است و زبان‌ها از یکدیگر با ساختارهای مشخص قابل تفکیک هستند و هر زبان تعداد ثابتی قواعد و واژگان دارد. این قابلیت به ما کمک می‌کند تا بتوانیم از مدل یادگیری با یک نمونه برای کاربرد پردازش زبان طبیعی استفاده کنیم.



شکل ۲ - ۱: شبکه ساده سیامی برای دسته بند دودویی[۱]

شبکه سیامی[2] برای تشخیص زبان سیامی یکی از زبان‌های آسیایی است و ساختار مشخص و واژگان ثابتی را دارد ایجاد شد.این شبکه به صورت یک شبکه عصبی دو قلو است. برای دسته‌بندی این شبکه دوتایی‌های مشابه و مخالف را ایجاد می‌کند. یعنی با استفاده از شباهت‌ها و تفاوت‌های موجود درون کاراکترهای زبان آن‌ها را دسته‌بندی می‌کند در انتها هنگام دریافت نمونه جدید شبکه شباهت و تفاوت‌های این نمونه با نمونه‌های موجود

[1] over fitting
[2] Siamese network

را تشخیص می‌دهد و آن را در یک دسته قرار می‌دهد.

برای محاسبه تفاوت و شباهت یک معیار فاصله وزن دار L۱ بین بردارهای ویژگی استخراج شده از دو شبکه دو قلو استفاده کرده و بر اساس آن نرخ یادگیری شبکه و نحوه دسته‌بندی واژگان را مشخص کرده‌است. در شکل ۲-۱ یک نمونه ساده از این شبکه برای دسته بندی دودویی را مشاهده می‌کنید.

شبکه سیامی یکی از شبکه‌های پر کاربرد در یادگیری عمیق است که در سیستم‌های تشخیص چهره نیز کاربرد دارد. در ادامه بررسی حملات یک مدل حمله به این شبکه نیز بررسی خواهد شد. به دلیل اینکه پایه کار و روش پیشنهادی یادگیری بدون نمونه است، در این بخش به طور مفصل آن را همراه با چند مثال بررسی می‌کنیم.

۲-۲ یادگیری بدون نمونه

در یادگیری بدون نمونه ما با دو نوع داده سروکار داریم. داده‌های دیده‌شده که برای آموزش مدل از آن‌ها استفاده می‌کنیم و داده‌های دیده‌نشده که مدل آموزش دیده شده باید بتواند آن‌ها را دسته‌بندی کند. داده‌های دیده‌شده، داده‌هایی برچسب خورده هستند و یک سری ویژگی‌های مشخص دارند. مدل آموزش دیده بر اساس این داده‌ها باید بتواند داده‌های دیده‌نشده را بر اساس ویژگی‌های استخراج کرده و ویژگی‌های داده‌های قبل، داده‌های جدید را دسته‌بندی کند.

فرآیند یادگیری و بررسی مدل، دو فرآیند جدا از هم هستند و داده‌های دیده‌شده و دیده‌نشده با یکدیگر اشتراکی ندارند. (به این مدل، مدل مجموعه باز[1] گفته می‌شود.) بر این اساس می‌توان از روش یادگیری انتقالی برای یادگیری مدل‌های بر پایه یادگیری بدون نمونه استفاده کرد. در یادگیری انتقالی مدل یکبار با استفاده از داده‌های مناسب وزن‌دهی و مقداردهی شده است، حال کافی است مدل آماده را برای تشخیص داده‌های جدید یا در روش‌های دیگر برای فعالیت‌های دیگر به کار گرفت و نیاز نیست تا دوباره از ابتدا فرآیند یادگیری را از سر بگیریم. فرآیند یادگیری انتقالی مانند نوزاد انسان است که در ابتدا درباره محیط اطراف اطلاعات کمی دارد اما

[1] open-set

کم کم از محیط یاد می‌گیرد و از تجربیات قبلی در یادگیری‌های بعدی نیز استفاده می‌کند.

چند حالت کلی رایج وجود دارد که استفاده از یادگیری بدون نمونه در این موارد به ما کمک می‌کند:

- وقتی فضای هدف به قدری بزرگ هست که همواره نمی‌توان تمامی حالات ممکن را پوشش داد؛ برای مثال: در تشخیص اشیا اطراف همواره اشیا جدید وجود دارد که در یک دسته‌بندی جدا قرار می‌گیرد.
- ممکن است نمونه‌های کلاس هدف کم‌یاب باشند و به سادگی به آن‌ها دسترسی نداشته باشیم؛ برای مثال: در تشخیص گونه‌های مختلف یک گیاه ممکن است همواره گونه‌های جدیدی پیدا شوند که تا به حال دیده نشده‌اند و یا گونه‌هایی باشند که دسترسی به آن‌ها دشوار است.
- ممکن است نمونه‌های کلاس هدف در طول زمان دچار تغییر شوند؛ مانند: نمونه‌های برندهای مختلف پوشاک
- ممکن است داده وجود داشته‌باشد اما فرآیند برچسب گذاری هزینه‌بر باشد.

با توجه به اینکه داده‌های کلاس هدف را تا کنون ندیده‌ایم، نیاز به یک سری اطلاعات جانبی[1] داریم تا بتوانیم ارتباطی بین داده‌های آموزش و داده‌های هدف پیدا کنیم. اطلاعات جانبی باید در فضای ویژگی مرتبط با نمونه‌ها باشند تا بتوانیم به عنوان اطلاعات مناسب از آن‌ها استفاده کنیم؛ به طور مثال: در مسئله تشخیص حیوانات در صورتی که مدل ما اسب و ببر را دیده باشد می‌تواند آن‌ها را دسته‌بندی کند. حال اگر یک نمونه تصویر گورخر به مدل نشان دهیم با توجه به اینکه تاکنون مدل آن را ندیده‌است باید بتوانیم با روشی به مدل بفهمانیم تا بتواند آن را نیز دسته‌بندی کند. گورخر جثه‌ای شبیه به اسب و طرحی شبیه به ببر دارد؛ این‌ها اطلاعات جانبی است که با استفاده از آن‌ها می‌توانیم بین داده‌های کلاس آموزش و هدف ارتباط برقرار کنیم. نمونه استفاده از اطلاعات جانبی در دسته‌بندی را می‌توانید در شکل ۲ - ۲ مشاهده کنید.

[1] auxiliary information

اطلاعات جانبی مورد استفاده به طور معمول یکسری اطلاعات معنایی هستند. این اطلاعات کمکی در یک فضای جدید (فضای معنایی[۱]) شامل هر دو دسته کلاس‌های دیده‌شده و دیده‌نشده می‌شوند. فضای معنایی نیز مانند فضای ویژگی یک فضای چند بعدی است. در فضای معنایی هر کلاس با یک توصیف برداری شناخته می‌شود که به این توصیف برداری، ضابطه کلاس[۲] گویند.

شکل ۲ - ۲: نحوه دسته بندی تصاویر توسط مدل ارائه شده[۲]

به طور کلی مقالات حوزه یادگیری بدون نمونه به دو طریق بر اساس تفاوت‌های فضای معنایی و بر اساس تفاوت روش‌های استفاده شده بررسی شده‌اند.فضاهای استفاده شده در دو دسته مهندسی شده[۳] و فضاهای یادگرفته شده [۴] و روش‌ها در دو دسته بر پایه دسته‌بند[۵] و بر پایه نمونه[۶] بررسی می‌شوند.[۶]

[1] semantic space
[2] class prototype
[3] Engineered
[4] Learned
[5] classifier-based
[6] instance-base

۲-۲-۱ فضاهای مهندسی شده

فضاهای مهندسی شده توسط متخصصان طراحی و متناسب با کاربرد مورد نظر استفاده شده‌اند. نمونه‌های پر کاربرد آن‌ها می‌توان به فضای مشخصه[۱]، لغوی[۲] و متنی_کلیدواژه‌ها [۳] اشاره کرد که هر یک را متخصر توضیح می‌دهیم.

فضای مشخصه

بر اساس یک سری مشخصه‌ها ساخته شده‌اند. مشخصه‌ها ویژگی‌های سطح بالایی هستند که برای انسان یک معنا و مفهومی را تداعی می‌کنند؛ مانند: رنگ پوست، نوع زیست و محیط زندگی. این ویژگی‌ها را نمی‌توان همانند ویژگی‌های سطح پایین مانند شکل و فرم کلی بدن با استفاده از شبکه‌های عصبی کانولوشنی یاد گرفت. کاربرد این مشخصه‌ها در انتقال یادگیری است. زیرا ویژگی‌های سطح پایین برای انتقال یادگیری در یادگیری بدون نمونه ارزش چندانی ندارند. این نوع از فضا به چند دسته تقسیم بندی می‌شود: دودویی، پیوسته و نسبی.

در مقاله Lampert۲۰۱۴ [۲] برای دسته‌بندی حیوانات یک پایگاه داده با فضای مشخصه پایگاه داده[۴] معرفی شده است. که شامل بیش از ۳۰۰۰۰ تصویر حیوانات در ۵۰ دسته مختلف و ۸۵ ویژگی معنایی است.

فضای لغوی

محموعه لغاتی هستند که بر پایه برچسب کلاس‌ها و داده‌هایی اند که اطلاعات معنایی دارند. مجموعه لغات استفاده شده می‌تواند WordNet باشد و برای ایجاد رابطه می‌توان روابط خواهر_برادی، والد_فرزندی و یا هر رابطه قابل تعریف دیگری استفاده کرد.

---

[1] Attribute
[2] lexical
[3] text-keywords
[4] Animals with Attributes-AwA

فضای متن_ کلیدواژه‌ها

این فضا بر اساس کلیدواژه‌ها ساخته شده‌اند. کلید واژه‌ها را می‌توان از هر منبعی یا سایتی مانند ویکی‌پدیا استخراج کرد. این کلیدواژه‌ها از توصیفاتی که درباره کلاس‌های مختلف وجود دارند استخراج می‌شوند.

به طور خلاصه فضاهای مهندسی شده توسط افراد متخصص طراحی می‌شوند و می‌توانند انعطاف پذیر باشند و متناسب با فضای معنایی و قوانین دلخواه تولید شوند؛ اما نکته منفی در مورد آن‌ها این است که وابسته به یک متخصص هستیم و تلاش انسانی زیادی برای تولید آن‌ها باید صورت گیرد.

۲-۲-۲ فضاهای یادگرفته شده

این فضاها بر خلاف مدل قبلی توسط ماشین یاد گرفته شده‌اند. در این فضاها هر بعد به تنهایی نمی‌تواند بیانگر یک معنا و مفهوم باشد و ابعاد مختلف در کنار هم معنا پیدا می‌کنند. مدل‌های ماشینی استفاده شده برای یادگیری این فضاها می‌توانند مدل‌های از قبل آموزش دیده باشند یا می‌توان از پایه یک مدل را طراحی کرد. سه فضای پر کاربرد می‌توان به جایگذاری برچسب[1]، جایگذاری متن[2] و نمایش تصویر[3] اشاره کرد.

فضای جایگذاری برچسب

این فضا بر اساس روش جایگذاری کلمه در پردازش زبان طبیعی ساخته شده‌است. در این روش، کلمات در فضایی از اعداد حقیقی جایگذاری می‌شوند و در نتیجه یک بردار در فضای جایگذاری به ازای هر کلمه ایجاد خواهد شد. این فضا حاوی اطلاعات معنایی است و کلماتی که حاوی معنایی نزدیک و مشابه اند در نزدیکی یکدیگر قرار می‌گیرند. در یادگیری بدون نمونه برچسب هر کلاس یک کلمه است که می‌توان از این روش استفاده کرد.

[1] Label-embedding
[2] Text-embedding
[3] Image-representation

فضای جابگذاری متن

این فضا شبیه به فضای متن_ کلیدواژه در فضاهای مهندسی شده‌است با این تفاوت که در این فضا ماشین ارتباطات و ضوابط بین کلاس‌ها و متون را پیدا خواهد کرد.

فضای نمایش تصویر

در این فضا برای هر کلاس تعدادی تصویر به عنوان نمونه انتخاب خواهد شد. این نمونه‌ها به یک ماشین، داده خواهد شد تا ارتباط بین تصاویر و کلاس‌های مورد نظر را پیدا کند و از آن طریق بردارهای خروجی را یافته و ضابطه کلاس‌ها را تشکیل دهد. برای مقابله با حملات و جعل تصاویر چهره ما نیاز به تولید چنین فضاهایی هستیم.

به طور خلاصه فضاهای یادگرفته شده نیازی به یک انسان متخصص ندارند و می‌توانند ویژگی‌هایی را تولید کنند که شاید از دید متخصص پنهان بنمان که این می‌تواند یک برتری نسبت به فضای مهندسی شده باشد. از طرفی، این نکته که هر بعد این فضاها به تنهایی یک معنا و مفهوم مستقل ندارد می‌تواند یک نقطه ضعف در مقابل فضاهای مهندسی شده باشد.

۲-۲-۳ روش‌های بر پایه دسته‌بند

هدف و تمرکز در این روش‌ها این است که از یک دسته‌بند برای تشخیص کلاس‌های دیده‌نشده استفاده کنیم. روش این دسته‌بندها یک در مقابل بقیه[1]، است که در این روش در هر تصمیم گیری برای کلاس دیده نشده یک مسئله دسته‌بندی دودویی وجود خواهد داشت. در حقیقت این روش‌ها یک سری دسته‌بند دودویی در کنار هم هستند. روش‌های ساخت دسته‌بندها در ادامه آورده شده‌است. این روش‌ها در سه دسته مبتنی بر تطابق[2]، میتنی بر ارتباط[3] و مبتنی بر ترکیب[4] بررسی می‌شوند.

[1] one versus rest
[2] Correspondence
[3] Relationship
[4] Combination

روش مبتنی بر تطابق

هدف، ساخت دسته‌بند با یافتن شباهت‌های موجود بین دسته‌بند یک در مقابل بقیه هر کلاس و ضابطه آن کلاس است. ضابطه هر کلاس یک نمایش و توصیفی از کلاس است و یافتن شباهت و تطابق بین آن و دسته‌بندی مربوط به آن کلاس به تولید دسته‌بند اصلی منجر می‌شود. برای داده‌های دیده نشده با استفاده از ضابطه کلاس و تابع تطابق یافت شده یک دسته‌بند ساخته می‌شود و کلاس جدید دسته‌بندی می‌شود.

نتطه قوت این روش ارتباطات بر پایه تابع تطابق است که به راحتی قابل یافتن است اما نقطه ضعف این روش این است که این ارتباطات را به طور صریح و واضح بیان نمی‌کند.

میتنی بر ارتباط

هدف، ساخت دسته‌بند بر اساس ارتباط بین کلاس‌هاست. برای یافتن ارتباط میان کلاس‌های دیده‌شده و دیده‌نشده می‌توان از ارتباط بین ضابطه کلاس‌ها و یا هر ارتباط مناسب دیگری استفاده کرد. برای کلاس‌های دیده‌نشده با استفاده از ارتباط موجود و دسته‌بند کلاس‌های دیده‌شده یک دسته بند برای هر کلاس دیده‌نشده ساخته می‌شود.

از دسته‌بندی کلاس‌های دیده‌شده می‌توان در مسائل دیگر نیز استفاده کرد و هزینه آموزش مدل را کاهش داد؛ اما، روابط بین کلاس‌ها از فضای معنایی به فضای ویژگی به طور مستقیم متنقل می‌شود که حل مسئله سازگاری از فضای معنایی به فضای ویژگی مشکل است.

مبتنی بر ترکیب

در این روش هر کلاس را ترکیبی از چند عنصر در نظر می‌گیریم که این عناصر در فضای معنایی وجود دارند. در حقیقت برای استفاده باید از فضاهای دودویی استفاده کرد که فضاهای معمول استفاده شده فضاهای مهندسی شده هستند. روش ساخت دسته‌بند برای کلاس‌های دیده‌نشده بدین صورت است که برای هر یک از عناصر سازنده کلاس‌ها دسته‌بند ساخته شده‌است سپس از آن دسته‌بندها برای ساخت دسته‌بند کلاس جدید استفاده می‌کنیم.

همانند روش قبل می‌توان از دسته‌بندی‌های مورد استفاده در مسائل دیگر نیز استفاده کرد؛ اما، بهینه سازی مسئله دو مرحله‌ای آموزش دسته‌بند مشخصه و استنتاج از مشخصه به کلاس دشوار است.

۲-۲-۴ روش‌های بر پایه نمونه

هدف در این روش‌ها ایجاد نمونه برچسب‌خورده برای کلاس‌های دیده نشده است. با استفاده از این نمونه‌ها دسته‌بند اصلی آموزش می‌بیند. با توجه به تفاوت در نحوه و منبع تولید نمونه، این روش‌ها به سه زیر دسته مبتنی بر تصویر کردن[1]، میتنی بر قرض نمونه[2] و مبتنی بر سنتز[3] تقسیم می‌شوند.

مبتنی بر تصویر کردن

تولید نمونه برای کلاس دیده‌نشده از طریق تصویرکردن نمونه‌های موجود در فضای ویژگی و ضابطه‌های موجود در فضای معنایی، به یک فضای مشترک انجام می‌شود. در فضای ویژگی با استفاده از دسته‌بند می‌توان داده‌های آزمایش را دسته‌بندی کرد. این ویژگی‌ها فقط برای داده‌های آزمایشی موجود هستند. فضای معنایی شامل ضابطه‌هایی است که هم شامل داده‌های جدید و هم شامل داده‌های آزمایش می‌شوند. می‌بایست یک ارتباط بین فضای ویژگی و فضای معنایی پیدا کنیم و با انتقال هر دو فضا و تصویر کردن در یک فضای سوم از ضابطه‌های موجود به عنوان نمونه کلاس‌های دیده نشده استفاده کنیم. با توجه به اینکه با این روش برای هر کلاس دیده‌نشده یک نمونه تولید می‌شود و فرآیند دسته‌بندی دشوار می‌شود می‌توان از روش‌های ناپارامتری مانند روش KNN استفاده کرد تا بتوان دسته‌بندی بهتری داشت.

انتخاب تابع نگاشت انعطاف پذیر است و با توجه به شرایط مسئله می‌توان دادگان مناسب انتخاب کرد؛ اما، چون برای هر کلاس دیده‌نشده یک نمونه برچسب زده وجود دارد ناچار به استفاده از روش‌های ناپارامتری می‌شویم.

---

[1] Projection
[2] Instance-borrowing
[3] Synthesizing

مبتنی بر قرض نمونه

در این روش بر اساس شباهت‌های بین کلاس‌های دیده شده و دیده نشده برای کلاس‌های دیده نشده از کلاس‌های دیده‌شده نمونه قرض می‌گیریم؛ برای مثال: اگر تا کنون کلاس یوزپلنگ را ندیده‌ایم می‌توانیم از کلاس پلنگ و ببر نمونه قرض بگیریم و برای آموزش کلاس یوزپلنگ از این نمونه‌ها استفاده کنیم. این نمونه‌ها به طور کامل مانند نمونه اصلی نیستند اما به دلیل شباهت‌هایی که دارند می‌توانند برای استفاده مناسب باشند.

به دلیل گستردگی نمونه‌های قرض داده‌شده می‌توان مدل‌های دسته‌بندی نظارت شده مختلفی را استفاده کرد؛ اما، چون نمونه‌ها همان نمونه‌های کلاس‌های دیده شده هستند باعث می‌شود تا دقت پایینی در دسته‌بندی داشته باشیم.

مبتنی بر سنتز

در این روش برای هر کلاس دیده‌نشده یک سری نمونه برچسب‌خورده می‌سازیم و با استفاده از آن‌ها کلاس‌های دیده نشده را آموزش می‌دهیم. در حقیقت پس از ساخت نمونه برای همه کلاس‌ها مسئله به یک مسئله یادگیری ماشین نظارت شده می‌شود. برای تولید داده مصنوعی روش‌های منعددی وجود دارد؛ برای مثال: اگر فرض کنیم که کلاس‌ها از یک توزیع خاص تبعیت می‌کنند با استفاده از حدس پارامترهای آن توزیع برای کلاس‌های دیده‌نشده می‌توان داده‌های مصنوعی تولید کرد.

از دیگر روش‌ها می‌توان به شبکه‌های مولد متخاصمی یا GAN[1] اشاره کرد. این شبکه‌ها از دو شبکه ساخته شده‌اند که یکی وظیفه تولید داده و دیگری وظیف صحت سنجی داده را دارد تا داده تولید شده به داده اصلی شبیه تر باشد. این شبکه‌ها در تولید داده مصنوعی از روش‌های دیگر از عملکرد و دقت بالاتری برخوردار هستند. استفاده از شبکه‌های GAN، یکی از روش‌های مرسوم برای تولید تصاویر جعلی برای دور زدن فرآیند احراز هویت سیستم‌های تشخیص چهره و گرفتن دسترسی از سیستم است.

[1] Generative Adverserial Networks

به دلیل گستردگی نمونه‌های تولید شده می‌توان مدل‌های دسته‌بندی نظارت شده مختلفی را استفاده کرد؛ اما، چون نمونه‌های تولید شده به طور معمول از توزیع نرمال پیروی می‌کنند. می‌تواند باعث سو گیری مدل تولید شده شود.

در خارج از شرایط آزمایشگاهی کلاس‌های دیده‌شده و دیده‌نشده در ترکیب با هم هستند و به این سادگی نمی‌توان آن‌ها را جدا از هم در نظر گرفت به همین دلیل میحث یادگیری بدون نمونه تعمیم یافته [۱] مطرح می‌شود. در حل این نوع مسائل روش‌های یاد شده لزوما نتیجه مناسبی نمی‌دهند و باید به دنبال روش‌های بهتر بود.

۲-۲-۵ یک مثال

روش‌ها و فضاهای یادگیری در یادگیری بدون نمونه در این فصل بررسی شد در ادامه یک مثال از یادگیری بدون نمونه برای دسته‌بندی حیوانات را برای فهم بهتر بررسی می‌کنیم.

مدل‌های مختلف که بر پایه روش یادگیری بدون نمونه ارائه شده‌اند می‌توانند ترکیبی از روش‌ها و فضاها را داشته باشند و یا فقط متکی بر یکی از آن‌ها طراحی شده باشند. به عنوان مثال در مقاله [۲] از فضای مشخصه استفاده کرده‌است و سعی کرده با ارائه دو روش برای یادگیری حیوانات مختلف را دسته‌بندی کند. مقاله از سه روش برای دسته‌بندی نام برده‌است و آن‌ها را با یکدیگر مقایسه می‌کند. دسته بندی چندکلاسه مسطح[۲]، پیش بینی مشخصه به طور مستقیم[۳] و پیش بینی مشخصه به طور غیر مستقیم [۴] به طور شهودی این سه روش در شکل ۲-۳ نشان داده شده‌اند. به دلیل این‌که همواره نمی‌توان پایگاه داده‌ای کامل و برچسب خورده داشت نیازمند راهکارهایی هستیم تا تلاش انسان در جمع‌آوری و دسته‌بندی داده‌ها را کم کنیم.

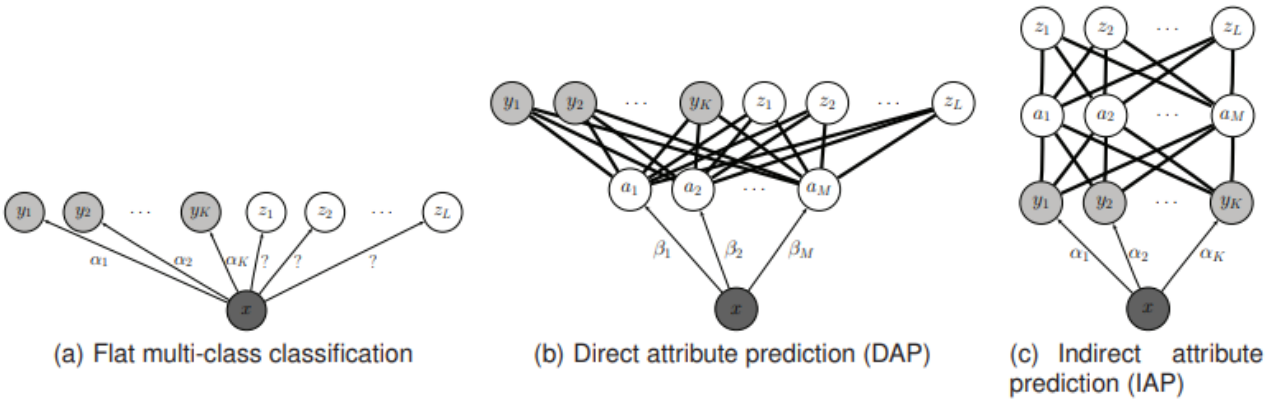
در شکل ۲-۳ تصویر الف، نشان دهنده روش اول است. در این روش ماشین یک بردار ثابت را یاد میگیرد. x ورودی مسئله، y برچسب داده‌ها آزمایش و z برچسب داده‌هایی است که نیاز داریم آن‌ها را بیابیم. چون فرآیند

[1] Generalized zero-shot learning
[2] flat multi-class classification
[3] direct attribute prediction
[4] indirect attribute prediction

یادگیری y هیچ تاثیری بر فرآیند یادگیری z ندارد پس از یادگیری‌های قبلی برای پیش بینی نمی‌توان استفاده کرد و این روش یک روش یادگیری معمولی است که قبلا نیز استفاده می‌شد و برای یادگیری z نیازمند داده‌های ورودی هستیم که با توجه به اینکه به دنبال یادگیری بدون نمونه هستیم این روش مناسب نیست.

راهکار ارائه شده برای یادگیری استفاده از ویژگی‌های معنایی سطح بالایی هستند که هرکدام برای انسان معنای خاصی دارند. این مشخصه‌ها، ویژگی‌های قابل نام‌گذاری هستند؛ مانند: رنگ، شکل، طرح بدن و عادات غذایی. استفاده از این مشخصه‌ها به ما کمک می‌کند تا نقش انسان را در فرآیند یادگیری کمتر کنیم و نیاز کمتری به ویژگی‌های سطح پایین تصویر داشته باشیم.

این مشخصه‌ها را می‌توان همراه با تصاویر یا همراه با دسته‌بندی‌های تصاویر درنظر گرفت و این ویژگی باعث شده است تا فرآیند دسته‌بندی بر اساس مشخصه‌ها[1] توسط مقاله معرفی شود. در این فرآیند با توجه به اینکه دسته‌های آموزش و آزمایش جدا از هم هستند و به یکدیگر وابسته نیستند؛ می‌توان با استفاده از این صفات و انتقال یادگیری[2]، بین داده‌های آموزشی، داده‌های آزمایشی را دسته‌بندی کرد. در شکل ۲-۲ می‌توانید نحوه دسته‌بندی بر اساس صفات را مشاهده کنید. با توجه به توضیحات ارائه شده به سراغ توضیح دو روش دیگر می‌رویم.



شکل ۲-۳: روش‌های دسته‌بندی [۲]

[1] attribute-based classification
[2] transfer learning

در روش پیش‌بینی مشخصه به صورت مستقیم[1] یا DAP در مرحله آموزش کلاس خروجی هر نمونه (x) برای لایه مشخصه‌ها نیز به طور قطع یک برچسب (y) مشخص می‌کند. متعاقبا می‌توان از هر روش با نظارتی استفاده کرد و پارامترهای هر مشخصه β_m را پیدا کرد. این بدین معنی است که فرآیند یادگیری اولیه برای یادگیری مشخصه‌ها متناسب با هر نمونه و دسته‌بندی مشخصه‌ها و نمونه‌ها را می‌توان با یک روش یادگیری نظارتی ساده انجام داد. حال با استفاده از یادگیری انتقالی و استفاده از شبکه‌ای که از پیش نمونه‌ها و مشخصه‌ها آن دسته‌بندی شده‌اند برای نمونه‌های جدید که از پیش دیده نشده‌اند (داده‌های آزمون) دسته‌بندی جدیدی را فقط بر اساس لایه مشخصه‌ها وزن‌ها یا پارامترهایی که شبکه تا به حال دارد و به لایه مشخصه‌ها اختصاص داده‌است، استفاده کرد و داده جدید را دسته‌بندی کرد.

در روش پیش‌بینی مشخصه به صورت غیر مستقیم[2] یا IAP همانند روش قبل از مشخصه‌ها برای انتقال دانش بین دسته‌ها استفاده می‌کند اما اینبار مشخصه‌ها بین دولایه از برچسب‌ها قرار دارند. یک لایه برچسب‌هایی که در لایه آموزش اختصاص می‌یابند (y) و یک لایه برچسب‌هایی که باید به داده‌های جدید اختصاص پیدا کنند (z) . در مرحله آموزش پارامترهای لایه مشخصه‌ها توسط برچسب‌های آموزش مقدار دهی می‌شوند همانند یک مسئله دسته‌بندی چندکلاسه که می‌توان با یک روش نظارتی ساده نیز یادگیری را انجام داد. در مرحله آزمایش با استفاده از مقادیری که به لایه مشخصه‌ها اختصاص داده‌شده دسته‌بندی داده‌های آزمایش را صورت می‌گیرد. استفاده از لایه مشخصه‌ها به برای دسته‌بندی داده‌های آزمایش این امکان را می‌دهد تا بتوانیم عملیاتی مشابه تنظیم[3] را انجام دهیم و فقط ترکیبات با معنی از مشخصه‌ها را ایجاد کنیم.

روش‌های یاد شده یک سری استراتژی کلی محسوب می‌شوند که می‌توانند با ترکیبی از روش‌های موجود مانند: یادگیری با نظارت یا رگرسیون‌ها بر روی مشخصه‌ها تصاویر یا دسته‌های تصاویر با استفاده از پارامترهای

[1] direct attribute prediction
[2] indirect attribute prediction
[3] regularization

پیش‌بینی شده انجام شوند. در این مقاله از روش توزیع‌های احتمالاتی استفاده شده و مشخصه‌های استفاده شده را مشخصه‌هایی بله یا خیر در نظر گرفته‌است. برای مشخصه‌هایی که بله یا خیر نیستند می‌توان به جای دسته‌بندی از رگرسیون استفاده کرد.

در روش DAP برای دسته‌بندی کردن تصاویر آزمون، احتمال قطعی به دست خواهد آمد زیرا در مرحله آموزش لایه مشخصه‌ها مقادیر مناسب را اتخاذ می‌کنند و از همین مقادیر برای تصاویر آزمون نیز استفاده می‌شود. به زبان ریاضی معادلات زیر صادق است.

$$p(z|x) = \sum_{a\in(0,1)^M} p(z|a)p(a|x)\frac{p(z)}{p(a^a)}\prod_{m=1}^{M} p(a_m^z|x). \quad (۱-۲)$$

$$f(x) = \underset{l=1,\dots,L}{\operatorname{argmax}} p(z=l|x) = \underset{l=1,\dots,L}{\operatorname{argmax}} \prod_{m=1}^{M} \frac{p(a_m^{z_l}|x)}{p(a_m^{z_l})}. \quad (۲-۲)$$

در روش IAP لایه مشخصه‌ها نقش یک تنظیم‌کننده را ایفا می‌کند پس به طور قطع نمی‌توان برای تشخیص لایه آزمون استفاده کرد و می‌بایست با استفاده از روابط احتمال یک احتمال میانی را حساب کرد و پس از آن از رابطه۲ - ۲ استفاده کرد.

$$p(a_m|x) = \sum_{k=1}^{K} p(a_m|y_k)p(y_k|x). \quad (۳-۲)$$

سه بانک اطلاعاتی استفاده شده است. در ادامه هر یک را توضیح می‌دهیم.۲ - ۱

Animal with Attributes

این بانک اطلاعاتی به عنوان بانک اصلی استفاده شده است. این بانک شامل ۵۰ کلاس حیوانات و ۸۵ کلاس مشخصه‌های معنایی است؛ از این تعداد ۴۰ کلاس به عنوان داده‌های آموزشی و ۱۰ کلاس به عنوان داده‌های

آزمایش استفاده شده است. تقسیم بندی به صورت تصادفی نیست اما سعی شده است تا توزیع داده‌ها در هر دو دسته آزمایش و آزمون به طور مناسبی صورت شود.

در کل ۳۰۴۷۵ عکس در این بانک داده وجود دارد که تعداد تصاویر برای دسته‌های مختلف حیوانات متفاوت می‌باشد. برای تسریع در محاسبات یک سری از ویژگی‌های تصاویر مانند: جنبه‌های رنگ، بافت و شکل، هیستوگرام رنگ و سایر ویژگی‌های مهم تصویر نیز به پایگاه داده اضافه شده‌و برای آموزش از روش 5 fold cross-validation استفاده شده‌است.

aPascal-aYahoo

این بانک شامل دو دسته داده یکی داده‌های بانک داده PASCAL و دیگری داده‌هایی که از موتور جستجوی Yahoo استخراج شده است. ۲۰ کلاس داده در PASCAL و ۱۲ کلاس داده در Yahoo و دسته‌بندی داده‌ها در هر یک با دیگری متفاوت است پس می‌توان از داده‌های PASCAL برای آموزش و از داده‌های Yahoo برای آزمون استفاده کرد. هر تصویر ۶۴ مشخصه‌ها دودویی را شامل می‌شود و همانند بانک داده قبلی برای تسریع در محاسبات یک سری از ویژگی‌های تصاویر از قبل محاسبه شده‌اند.

SUN Attributes

این بانک زیر مجموعه‌ای از بانک داده SUN Database که شامل ۷۱۷ کلاس داده و هر تصویر شامل ۱۰۲ مشخصه دودویی است. این مشخصه‌ها شامل توضیفات صحنه، شرایط نورپردازی، مواد داخل تصویر و ... است.

جدول ۲ - ۱: بانک‌های اطلاعاتی[۲]

Dataset	AwA	aP/aY	SUN
# Images	۳۰۴۷۵	۱۵۳۳۹	۱۴۳۴۰
# Classes	۵۰	۳۲	۷۱۷
# Attributes	۸۵	۶۴	۱۰۲
Annotation Level (real- binary) or valued	class per both	image per binary	image per binary
Type Annotation			

برای ارزیابی از ماشین بردار پشتیبان[1] یا SVM استفاده شده‌است. برای روش مستقیم از یک SVM غیر خطی و برای روش غیرمستقیم از نوع one-versus-rest استفاده شده‌است. نتایج به دست آمده را در جداول۲-۲ و۲-۳ مشاهده می‌کنید.

جدول ۲-۲: تقسیم بندی پیش‌فرض داده‌ها [۲]

rnd	CT-H	CT-cc	IAP	DAP	method
۰.۱۰	30.8 ± 0.2	30.7 ± 0.2	۲.۴۲	۴.۴۱	MC acc.
۰.۵۰	۴.۷۳	۴.۷۳	۰.۸۰	۴.۸۱	classAUC
۰.۵۰	–	–	۱.۷۲	۸.۷۲	attrAUC

جدول ۲-۳: تقسیم بندی داده‌ها به صورت تصادفی[۲]

md	CT-H	CT-cc	LAP	DAP	method
۰.۱۰	27.3 ± 4.0	27.7 ± 4.3	34.1 ± 5.1	37.1 ± 3.9	MC acc.
۰.۵۰	72.8 ± 3.1	72.4 ± 2.7	76.3 ± 5.5	80.4 ± 3.1	classAUC
۰.۵۰	–	–	69.7 ± 3.8	70.7 ± 3.5	attrAUC

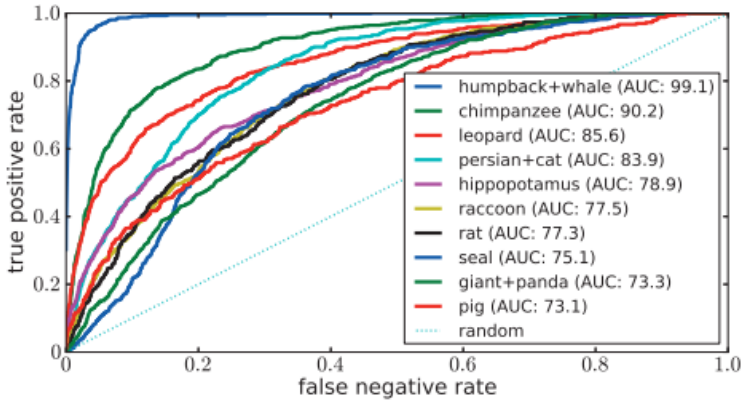
در روش مستقیم برای بدست آوردن مقادیر مناسب کرنل ازSVM منحنی ROC[2] و میانگین سطح زیر منحنی[3] برای صفات و روش 5 fold cross-validation استفاده شده‌است.

منحنی یاد شده، یک نمودار برای نمایش توانایی ارزیابی یک سیستم دسته‌بندی دودویی محسوب می‌شود که آستانه تشخیص آن نیز متغیر است. که با ترسیم نسبت نرخ مثبت صحیح[4] که به اختصار TPR نامیده می‌شود برحسب نرخ مثبت کاذب[5] با نام اختصاری ،FPR ایجاد می‌شود.

در روش غیر مستقیم مراحل مانند روش قبل است با این تفاوت که میانگین سطح زیر منحنی بر روی پیش‌بینی کلاس‌ها استفاده شده‌است.

همان‌طور که در شکل۲-۴ مشاهده می‌کنید. دقت روش در تشخیص برخی از دسته‌ها مانند نهنگ‌های کوهان‌دار دقت بالایی همانند روش‌های یادگیری با نظارت دارد؛ اما در مورد دسته‌بندی‌هایی مانند خوک‌هاو

[1] Support Vector Machine
[2] Receiver Operation Characteristic
[3] AUC
[4] True Positive Rate
[5] False Positive Rate



شکل ۲-۴: منحنی]ROC۲[

پانداهای بزرگ دقت خیلی پایینی دارد. یکی از دلایل این اتفاق شباهت‌هایی است که در این دسته‌ها رخ می‌دهد؛ به طور مثال: ظاهر پانداهای بزرگ، خوک‌ها و اسب‌های آبی.

به طور مشابه بر روی دو پایگاه داده دیگر نیز بررسی‌ها انجام شده‌است و نتایج را در جداول زیر مشاهده می‌کنید.

جدول ۲-۴: تقسیم بندی پیش‌فرض داده‌ها برای]aPascal-aYahoo۲[

rnd	CT-H	IAP	DAP-C	DAP-I	method
۳.۸	16.7 ± 0.5	۹.۱۶	۱.۱۹	۸.۱۶	MC acc.
۰.۵۰	۲.۶۴	۴.۷۵	۵.۷۶	۹.۷۶	classAUC
۰.۵۰	–	۱.۷۳	۷.۷۳	۶.۷۰	attrAUC

جدول ۲-۵: تقسیم بندی داده‌ها برای sub]attributes۲[

rnd	CT-H	IAP	DAP-C	DAP-I	method
۴.۱	12.9 ± 1.3	18.0 ± 1.5	22.2 ± 1.6	18.1 ± 1.2	MC acc.
۲.۶	32.6 ± 2.0	41.1 ± 2.1	46.6 ± 1.7	40.2 ± 2.1	level2 acc.
۳.۳۳	74.2 ± 2.0	82.1 ± 2.5	85.7 ± 2.1	74.2 ± 4.0	level1 acc.
۰.۵۰	77.1 ± 0.0	87.9 ± 0.7	92.3 ± 0.7	90.5 ± 0.7	class mAUC
۰.۵۰	–	82.7 ± 0.8	83.9 ± 0.8	82.0 ± 0.6	attrAUC

۲-۳ شبکه‌های GAN و روش‌های حمله

در یادگیری ماشین دو دسته مدل وجود دارند، دسته اول مدل‌های جداساز[1] و دسته دوم مدل‌های مولد می‌باشند. در مدل‌های جداساز هدف دسته‌بندی و متمایز ساختن کلاس‌هاست، درحالیکه در مدل‌های مولد توزیع داده‌ها محاسبه و تخمین زده می‌شود، سپس با استفاده از پارامترهای بدست آمده برای توزیع کلاس‌ها دسته‌بندی می‌شوند. علاوه بر دسته‌بندی، می‌توان برای هر کلاس و توزیع داده جدید تولید کرد.

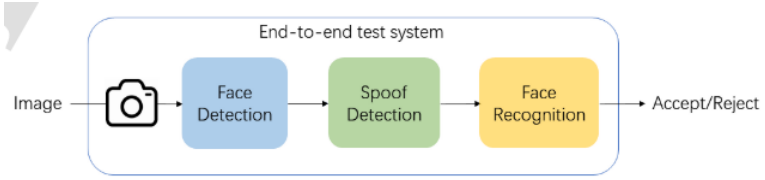
GAN یک شبکه مولد است که می‌توان از آن برای تولید تصاویر ساختگی که در عین واقعی و طبیعی بودن هرگز وجود نداشته‌اند استفاده کرد. این شبکه شامل دو بخش است، یک بخش مولد و دیگری جداساز. بخش مولد تصویر را تولید و بخش جداساز آن را تشخیص می‌دهد که واقعی است یا غیر واقعی و در صورت تشخیص غیر واقعی دوباره وزن‌های شبکه مولد به‌روزرسانی شده و تصویر جدیدی تولید می‌شود و این کار تا جایی ادامه پیدا می‌کند که شبکه جداساز به طور تقریباً برابر تشخیص واقعی و غیر واقعی بدهد و در دسته‌بندی تصویر تولیدی عاجز شود. [۷]

حملات به سیستم‌های تشخیص چهره در دو دسته کلی تقسیم بندی می‌شوند: حملات به ساختار شبکه اصلی (حملات درب پشتی[2]) و حملات با استفاده از روش‌های جعل تصویر

به طور کلی حملات درب پشتی بیشتر در دسته دحمله به ساختار سیستم‌ها و شبکه‌ای قرار می‌گیرند که مدل‌ما بر روی آن در حال اجراست. مانند سیستم‌های ابری که سرویس‌های یادگیری ماشین یا یادگیری عمیق ارائه می‌کنند و حملات جعل تصویر حملات دقیق تری به خود ساختار شبکه‌ی یادگیری مدل محسوب می‌شوند.

حملات جعل تصویر به طور عمده در دسته‌های replay، ۳D print، makeup mask و partial قرار می‌گیرند. در هر کدام از این حملات شبکه‌های GAN می‌توانند برای تولید تصویر استفاده شوند.

[1] Discriminative
[2] backdoor attacks

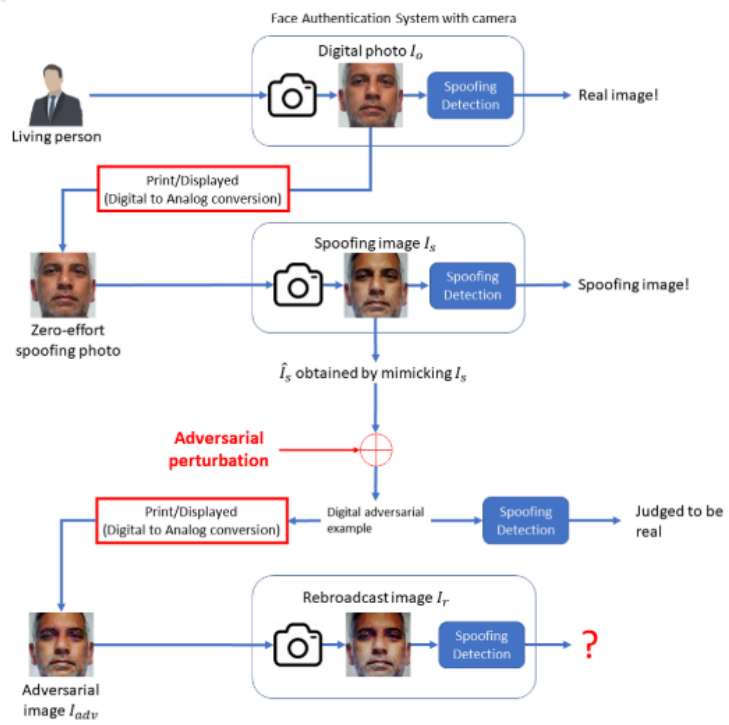


شکل ۲ - ۵: ساختار شبکه مورد استفاده برای یک سیستم تشخیص چهره[۳]

برای مثال در مقاله (Zhang۲۰۲۰) [۳] یک نوع از این حملات مورد بررسی قرار گرفته‌است. این حمله در بدترین سناریو یعنی زمانی که هکر از ساختار کلی شبکه با خبر است و هیچ دسترسی به سرور و محیط مدل و دادگان ندارد در نظر گرفته شده است.

در این روش هکر باید یک مدل مشابه با مدلی که قرار است به آن حمله صورت بگیرد درست کند و تصویر مد نظر را به مدل دهد خروجی آن را از مرحله تشخیص جعل بگیرد و با استفاده از روش های GAN جوری تصویر را تغییر دهد تا در مرحله بعد بتواند با تصویر مورد نظر سیستم را دور زده و دسترسی بگیرد. نکته حائز اهمیت در این سیستم‌ها این است یک فرآیند تبدیل داده از دیجیتال به آنالوگ و برعکس آن در فرآیند تصویر برداری از فرد و استفاده از تصویر گرفته شده در مدل است و این میتواند برای تولید داده توسط شبکه‌های GAN ایجاد مشکل کند زیرا باید بتواند بعد از تغییر از دیجیتال به آنالوگ و برعکس آن نیز ویژگی‌های مورد نیاز برای دور زدن سیستم را داشته باشد.

همانطور که در تصویر ۲ - ۵ مشاهده می‌کنید شبکه از سه بخش تشخیص چهره، تشخیص جعل و شناسایی چهره تشکیل شده‌است. نکته حائز اهمیت این است که تصویر دستکاری شده باید از مرحله تشخیص چهره عبور کند، پس باید ساختار چهره در حین دستکاری حفظ شود. در تصویر ۲ - ۶ مشاهده می‌کنید در این روش تصویر می‌بایست دوبار از فریاند تبدیل دیجیتال به آنالوک و برعکس عبور کند و می‌توان آن را نقص روش در نظر گرفت. با این حال این روش می‌تواند موفقیت‌آمیز باشد و مدل کلی را دور بزند و به هکر اجازه ورود بدهد.



شکل ۲ - ۶: ساختار حمله مورد استفاده برای یک سیستم تشخیص چهره[۳]

حملات درب پشتی را می‌توان در چند دسته قرار داد: [۸]:

- راه‌اندازی با استفاده از یک ورودی خاص (یک تصویر یا یک طرح خاص)
- نخریب فرآیند آموزش و دستکاری داده‌ها (مثلا در MLaas[1])
- انجام یک رفتار مخرب مانند: ناتوانی در دسته‌بندی تصویر ورودی یا کاهش دقت کلی مدل

حملات درب پشتی میتوانند به صورت white-box یا black-box مبنی بر اینکه هکر درباره ساختار شبکه اطلاعات دارند و میتواند آن را تغییر دهد یا خیر. به طور مثال در مقاله (Guo۲۰۲۱) [۸] مدلی از حمله معرفی شده‌است که در آن فرض بر این است که هکر به تمامی فرآیند تسلط کامل دارد و با استفاده از یک تصویر و تغییر در ساختار دادگان تصاویر می‌تواند یک جعل هویت همگانی[2] ایجاد کند و پس از آن به جای هر فردی احراز هویت شود.

[1] Machine Learning as a Service
[2] universal impersonate

این حمله نسبت به حملات مشابه قبلی از شدت بیشتری برخوردار است زیرا در حملات قبلی روی کاربر خاصی تمرکز می‌شد اما در این حمله تمرکز روی همه است و نکته دیگر درباره این حمله این است که این حمله بر روی شبکه سیامی تست شده است که ویژگی مهم این شبکه، مجموعه باز بودن فرآیند آموزش و آزمایش است. این شبکه یک ساختار دو قلو دارد و یک جفت تصویر را با هم مقایسه می‌کند برای ایجاد یک جعل هویت همگانی باید کاری کرد که شبکه در همه حالت درست کار کند مگر وقتی که تصویری را به عنوان ورودی بگیرد که هکر به آن داده است در این صورت باید همواره درست باشد و احراز هویت انجام شود.

برای این کار لازم است تا دادگان تصاویر تخریب شود و به ازای هر نمونه که چند تصویر وجود دارد با احتمال (نه خیلی کم که تصویر هکر شناخته نشود و نه زیاد که سیستم دچار اختلال شود.) تصاویر با تصویر هکر جایگزین شوند و برچسب ۱ مبنی بر مطابقت داشتن بگیرند. پس از آموزش شبکه با دادگان جدید هکر می‌تواند با ارسال تصویر خود به سیستم تشخیص چهره دسترسی لازم از سیستم را بگیرد. البته ممکن است نیاز باشد چندباری تلاش کند تا جواب صحیح بگیرد. (به دلیل کم بودن تعداد تصاویر هکر در مقایسه با تصاویر دادگان). برای توضیحات بیشتر مبنی بر نحوه ساخت دادگان و دادگان‌های استفاده شده و سایر موارد مقاله اصلی [۸] را مطالعه کنید. هدف از آوردن این مقاله در این بخش آشنایی بیشتر با انواع حملاتی بود که به سیستم‌های تشخیص چهره صورت میگیرد.

۲-۴ جمع‌بندی

در بخش اول یادگیری با یک نمونه بررسی شد و شبکه سیامی که یکی از شبکه‌های پر کاربرد در تشخیص چهره است معرفی شد. در بخش دوم با یک مثال و بررسی‌های کلی در مورد یادگیری بدون نمونه سعی کردیم تا به طور کلی با این روش آشنا شویم. کاربرد روش بدون یادگیری بسیار گسترده‌است و از کاربردهای آن می‌توان به استفاده در تشخیص حرکت در ویدیو و جلوگیری از جعل هویت اشاره کرد. در بخش سوم کمی شبکه GAN را توضیح

دادیم و چند مورد از حملات به سیستم‌های تشخیص چهره را توضیح دادیم. در ادامه و در فصل بعد کارهای مرتبط انجام شده برای جلوگیری و تشخیص جعل تصویر چهره و حملات به این سیستم‌ها را بررسی می‌کنیم.

فصل سوم

کارهای مرتبط

جلوگیری از حملات و جعل تشخیص چهره نیازمند استفاده از مدل‌های مناسب به همراه دادگان مناسب است. در این فصل با بررسی چند کار مرتبط به معرفی چند دادگان که مورد استفاده در فرآیند آموزش و تست مدل‌ها هستند می‌پردازیم.

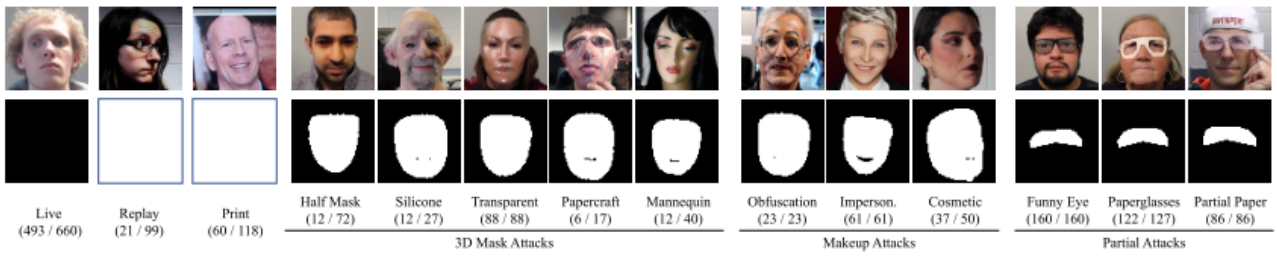
۳-۱ یادگیری بدون نمونه با درخت تصمیم

در مقاله (Liu۲۰۱۹) [۴] یک روش جدید بر پایه شبکه درختی عمیق یا DTN[1] معرفی شده‌است. هدف این روش دسته‌بندی حملات شناخته‌شده در زیردسته‌های معنایی[2] از طریق یادگیری بدون نظارت و یادگیری ویژگی‌ها به صورت سلسله مراتبی است.

[1] Deep Tree Network
[2] semantic sub-groups

درخت از دو بخش تشکیل شده است؛ نودهای داخلی که ازیک واحد کانولوشن باقی‌مانده یا CRU[1] به همراه یک واحد مسیریابی یا TRU[2] و برگ‌ها از یک واحد CRU به همراه یک واحد یادگیری با نظارت ویژگی‌ها یا SFL[3] تشکیل شده‌اند.

در بخش داخلی یادگیری بدون نظارت بر اساس یک معادله انجام می‌شود که هدف این معادله حداکثر کردن فاصله بین میانگین داده‌ها سمت شاخه سمت چپ با شاخه سمت راست است با این شرط که میانگین کل داده‌ها صفر باقی بماند. در حقیقت در نهایت به دنبال حداکثر کردن پراکندگی[4] داده‌هاست. روشی که الهام گرفته از PCA است. در برگ‌ها یادگیری با نظارت در دو شاخه انجام می‌شود، یک شاخه برای دسته‌بندی دودویی و شاخه دیگر برای عبور دادن نتیجه نود داخلی درخت از یک ماسک تولید شده از فرآیند توجه[5] برای تشخیص محل جعل در تصویر. در تصویر ۳-۱ می‌توانید حملات و ماسک‌های مربوطه را مشاهده کنید.



شکل ۳-۱: حملات و ماسک‌های توجه[۴]

برای اینکه بتواند دسته‌بندی مناسبی ارائه دهد نیاز به یک دادگان کامل است. به همین دلیل دادگان SiW-M که مدل توسعه داده شده از دادگان SiW است معرفی شد که شمال حملات بیشتری باشد. همچنین برای مقایسه نتیجه از دادگان‌های دیگری نیز استفاده شده‌است که در تصویر ۳-۲ مشاهده می‌کنید.

از ویژگی‌های این روش استفاده از داده زنده (تصویر واقعی) و تصویر جعل شده به طور همزمان در مدل در فرآیند یادگیری و تست است که کمتر در روش‌های قبلی مورد استفاده قرار می‌گرفته‌است. نتایج دقیق و

[1] Convolutional Residual Unit
[2] Tree Routing Unit
[3] Supervised Feature Learning
[4] variance
[5] attention mask

Dataset	Year	Num. of subj./vid.	Face variations			Spoof attack types					Total num. of spoof types
			pose	expression	lighting	replay	print	3D mask	makeup	partial	
CASIA-FASD [50]	2012	50/600	Frontal	No	No	1	2	0	0	0	3
Replay-Attack [15]	2012	50/1, 200	Frontal	No	Yes	1	1	0	0	0	2
HKBU-MARs [30]	2016	35/1, 008	Frontal	No	Yes	0	0	2	0	0	2
Oulu-NPU [9]	2017	55/5, 940	Frontal	No	No	1	1	0	0	0	2
SiW [32]	2018	165/4, 620	$[-90°, 90°]$	Yes	Yes	1	1	0	0	0	2
SiW-M	2019	493/1, 630	$[-90°, 90°]$	Yes	Yes	1	1	5	3	3	13

شکل ۳-۲: مقایسه دادگان مورد استفاده برای یادگیری مدل در برابر حملات جعل تصویر[۴]

تصویرسازی‌های حین و پس از فرآیند آموزش در مقاله قابل مشاهده‌است. نکته حائز اهمیت این مقاله استفاده از روش درختی بر پایه یادگیری بدون نمونه و معرفی و استفاده از دادگان غنی است که کاربرد زیادی در کار ما می‌تواند داشته باشد.

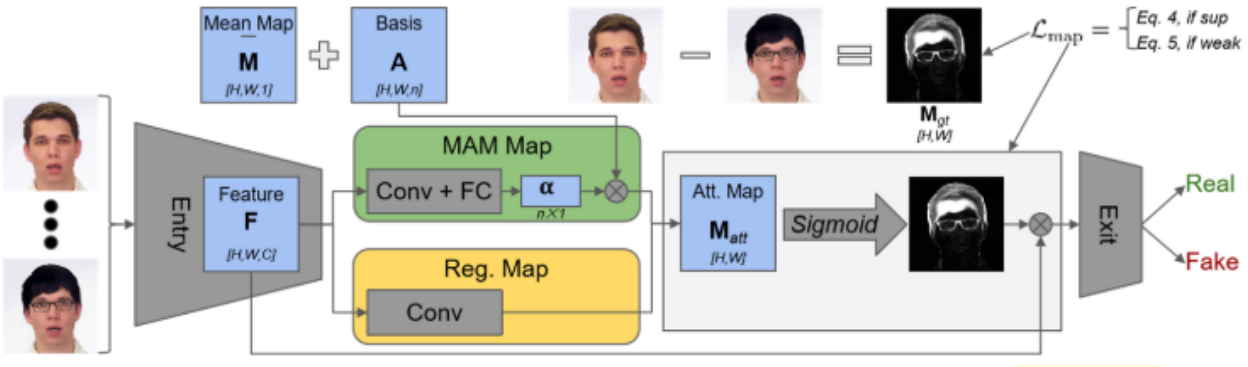
۳-۲ یادگیری بدون نمونه با مکانیزم توجه

در مقاله (Dang۲۰۱۹) [۵] از ترکیب یادگیری بدون نمونه با استفاده از فرآیند توجه استفاده شده که حائز اهمیت است. مکانیزم توجه به طور مثال در مدل‌های ترجمه متن ایجاد زیرنویس برای تصویر یا ویدیو کاربرد به سزایی دارد. انواع مکانیزم توجه وجود دارد به طور مثال در مقاله (Xu۲۰۱۵) [۹] از این مکانیزم برای تولید زیرنویس مناسب برای تصویر استفاده شده‌است. این مقاله بر روی حملات دستکاری دیجیتالی تصویر تمرکز دارد و میٰخواهد تا با استفاده از یک نگاشت توجه مناسب همانند مقاله (Liu۲۰۱۹) [۴] برای تشخیص ناحیه مورد جعل استفاده کند.

همانطور که در تصویر ۳-۳ مشاهده می‌کنید مدل شبکه از دو بخش برای مقایسه تولید شده‌است. یک بخش که از یک یا چند لایه کانولوشن تشکیل شده و بخش دیگر که از ترکیب یک یا چند لایه کانولوشن به همراه لایه تمام متصل و دو نگاشت یکی نگاشت میانگین و دیگری مجموعه‌ای از نگاشت‌های پایه تشکیل شده است. در انتها خروجی هر کدام تحت عنوان نگاشت توجه به لایه sigmoid برای تصمیم‌گیری داده می‌شود.

برای محاسبه نگاشت پایه یک عملیات PCA بر روی صد ماسک دستکاری شده از FaceApp انجام شده

است و ده مولفه اول آن به عنوان نگاشت پایه و میانگین این مولفه‌ها به عنوان نگاشت میانگین در نظر گرفته می‌شود.



شکل ۳-۳: ساختار شبکه مورد استفاده برای یک سیستم تشخیص چهره[۵]

فرآیند یادگیری در سه حالت انجام شده‌است. در حالت اول برای هر حمله ماسک متناظر وجود دارد. (به طور کامل با نظارت) در حالت دوم برای برخی از داده‌ها ماسک متناظر وجود ندارد (نیمه نظارتی) و در حالت سوم هیچ ماسک متناظری وجود ندارد و فرآیند توجه باید خودش اطلاعات را به طور خودکار فرا بگیرد. (بدون ناظر)

برای اینکه یادگیری همه جانبه باشد و انواع مختلفی از حملات را در بر گیرد یک دادگان از مجموع چند دادگان دیگر تولید شده‌است که مقایسه آن را در تصویر ۳-۴ با دیگر دادگان استفاده شده در مقاله می‌توانید مشاهده کنید.

Dataset	Year	# Still images		# Video clips		# Fake types				Pose variation
		Real	Fake	Real	Fake	Id. swap	Exp. swap	Attr. mani.	Entire syn.	
Zhou *et al.* [61]	2018	2,010	2,010	-	-	2	-	-	-	Unknown
Yang *et al.* [58]	2018	241	252	49	49	1	-	-	-	Unknown
Deepfake [29]	2018	-	-	-	620	1	-	-	-	Unknown
FaceForensics++ [42]	2019	-	-	1,000	3,000	2	1	-	-	$[-30°, 30°]$
FakeSpotter [52]	2019	6,000	5,000	-	-	-	-	-	2	Unknown
DFFD (our)	2019	58,703	240,336	1,000	3,000	2	1	28 + 40	2	$[-90°, 90°]$

شکل ۳-۴: ساختار شبکه مورد استفاده برای یک سیستم تشخیص چهره[۵]

نکته حائز اهمیت در این مقاله استفاده از مکانیزم توجه به همراه یادگیری بدون نمونه بود که برای کارهای زیادی به جز جلوگیری از جعل تصویر چهره می‌توان استفاده کرد.

۳-۳ جمع‌بندی

در این بخش به معرفی چند کار مرتبط پرداختیم. روش‌های استفاده شده در این مقالات مانند: مکانیزم توجه به همراه یادگیری بدون نمونه، استفاده از درخت تصمیم و استفاده از شبکه‌های GAN برای تولید داده و دادگان معرفی شده توسط هر کدام را به طور مختصر مورد بررسی قرار دادیم. در ادامه و در فصل آینده درباره روش پیشنهادی و نقاط صعف و قوت این روش و روش‌های پیشین بحث خواهیم کرد.

فصل چهارم

روش پیشنهادی

فصل پنجم

پیاده‌سازی

فصل ششم

ارزیابی کارایی

فصل هفتم

جمع‌بندی و نتیجه‌گیری

مراجع

[1] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition," tech. rep., 2015.

[2] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorizationa," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 453–465, mar 2014.

[3] B. Zhang, B. Tondi, and M. Barni, "Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability," *Computer Vision and Image Understanding*, vol. 197-198, p. 102988, 2020.

[4] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 4675–4684, apr 2019.

[5] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5780–5789, oct 2019.

[6] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A Survey of Zero-Shot Learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, pp. 1–37, jan 2019.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *Communications of the ACM*, vol. 63, pp. 139–144, jun 2014.

[8] W. Guo, B. Tondi, and M. Barni, "A Master Key backdoor for universal impersonation attack against DNN-based face verification," *Pattern Recognition Letters*, vol. 144, pp. 61–67, 2021.

[9] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *32nd International Conference on Machine Learning, ICML 2015*, vol. 3, pp. 2048–2057, 2015.