دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر
سمینار کارشناسی ارشد گرایش هوش مصنوعی

عنوان:
یادگیری از صفر با شبکه‌های عمیق
Deep Zero-Shot Learning

نگارش:
سید محسن شجاعی
۹۳۲۰۷۹۷۹

استاد راهنما:
دکتر مهدیه سلیمانی

استاد ممتحن داخلی:
دکتر حمیدرضا ربیعی

بهمن ۹۴

چکیده: مسئله یادگیری از صفر [۱] به دنبال پیش‌بینی دسته‌هایی در زمان آزمون است که در زمان آموزش هیچ داده‌ای از آن‌ها مشاهده نشده است و شناسایی آن‌ها با اطلاعات جانبی صورت می‌گیرد. در یک مسئله دسته‌بندی تصاویر، یادگیری بدون برد به این صورت است که تعدادی تصویر به همراه برچسب و اطلاعات جانبی به الگوریتم داده می‌شود، در زمان آزمون اطلاعات جانبی مربوط به دسته‌های جدید و تصاویری بدون برچسب وجود دارد و هدف برچسب‌گذاری تصاویر با دسته‌های جدیدی است که اطلاعات جانبی آن‌ها داده شده. ویژگی‌های بصری و متونی که ویژگی‌های یک دسته را شرح می‌دهند، مثال‌هایی از اطلاعات جانبی مورد استفاده در این نوع مسائل هستند. در این گزارش حالت‌های مختلف تعریف مسئله یادگیری از صفر معرفی می‌شود. سپس کارهای پیشین انجام شده مورد بررسی قرار می‌گیرد. سپس یک روش پیش‌نهادی ارائه شده و نتایج آن با روش‌های پایه مقایسه خواهد شد. در نهایت راه‌کارهایی برای ادامه پژوهش بیان شده و جمع‌بندی انجام می‌شود.

واژه‌های کلیدی: یادگیری از صفر، یادگیری بازنمایی، شبکه‌های عمیق

۱ مقدمه

در حوزه یادگیری ماشین مسئله استاندارد یادگیری با نظارت به صورت‌های مختلف توسعه یافته است و به کمک این روش‌ها، یادگیری ماشین از عهده‌ی کارهای بسیار چالش‌برانگیزتری بر آمده است. بر خلاف پارادایم سنتی یادگیری با نظارت، که فرض می‌کند داده‌های فراوانی از تمام دسته‌ها برای آموزش در اختیار قرار دارد، عموم این روش‌ها به دنبال کم کردن نیاز به داده‌های برچسب‌دار در زمان آموزش هستند. *یادگیری نیمه‌نظارتی* [۲] [۱] برای استفاده کردن از حجم زیاد داده‌های بدون برچسب موجود در جریان آموزش پیشنهاد شده است. *یادگیری از تک نمونه* [۳] [۲] سعی می‌کند یک دسته را تنها بوسیله یک نمونه‌ی برچسب‌دار از آن و البته با کمک نمونه‌های برچسب‌دار از سایر دسته‌ها شناسایی کند. *انتقال یادگیری* [۴] [۳] سعی می‌کند دانش به دست آمده از داده‌های یک دامنه یا برای انجام یک وظیفه را به داده‌های دامنه‌ی دیگر یا وظیفه‌ی دیگری روی داده‌ها منتقل کند. هیچ‌کدام از این روش‌ها نیاز به داده‌های برچسب‌دار را برای دسته‌هایی که مایل به تشخیص آن هستیم را به طور کامل از بین نمی‌برد. برای دست‌یابی به چنین هدفی، مسئله یادگیری از صفر صورت‌بندی شده است [۴]. در این مسئله در حالتی که داده‌های آموزش برای بعضی از دسته‌ها هیچ نمونه‌ای در بر ندارند، به دنبال یافتن یک دسته‌بند برای آن‌ها هستیم. برای این که چنین کاری ممکن باشد فرض می‌شود که یک *توصیف* از تمامی کلاس‌ها موجود است. نیاز به حل چنین مسئله‌ای به خصوص وقتی که تعداد دسته‌ها بسیار زیاد است رخ می‌دهد. برای مثال در بینایی ماشین تعداد دسته‌ها برابر انواع اشیای موجود در جهان است و جمع‌آوری داده‌های آموزش برای همه اگر غیر ممکن نباشد به هزینه و زمان زیادی احتیاج دارد. همانطور که در [۵] نشان داده‌شده، تعداد نمونه‌های موجود برای هر دسته از قانون Zipf پیروی می‌کند و نمونه‌های فراوان برای آموزش مستقیم دسته‌بند برای همه‌ی دسته‌ها وجود ندارد. یک مثال دیگر رمزگشایی فعالیت ذهنی فرد است [۶]؛ یعنی تشخیص کلمه‌ای که فرد در مورد آن فکر یا صحبت می‌کنند بر اساس تصویری که از فعالیت مغزی او تهیه شده است. طبیعتاً در این مسئله تهیه تصویر یا سیگنال فعالیت مغزی برای تمامی کلمات لغت‌نامه ممکن نیست. یک موقعیت دیگر که توصیف مسئله یادگیری از صفر بر آن منطبق است دسته‌بندی دسته‌های نوظهور است، مانند تشخیص مدل‌های جدید محصولاتی چون خودروها که بعضی دسته‌ها در زمان آموزش اصولا وجود نداشته است. یادگیری از صفر نیز مانند بسیاری از مسائل یادگیری ماشین با توانایی‌های یادگیری در انسان ارتباط دارد و الهام از یادگیری انسان‌ها در شکل‌گیری‌اش بی‌تاثیر نبوده است. برای مثال انسان قادر است بعد از شنیدن توصیف «حیوانی مشابه اسب با راه‌راه‌های سیاه و سفید» یک گورخر را تشخیص دهد. یا تصویر یک اسکوتر را با توصیف «وسیله‌ای دو چرخ، یک کفی صاف برای ایستادن، یک میله صلیبی شکل با دو دستگیره» تطبیق خواهد داد.

در این نوشتار بر مسئله دسته‌بندی تصاویر از صفر تمرکز می‌کنیم؛ به این معنی که داده‌هایی که مایل به دسته‌بندی آن هستیم تصاویر هستند. در نتیجه در زمان آموزش تعدادی تصویر به همراه برچسب آن‌ها موجود است. دسته‌هایی که از آن‌ها در زمان آموزش نمونه موجود است را دسته‌های دیده شده یا *دسته‌های آموزش* می‌نامیم. همچنین یک نوع اطلاع جانبی هر یک از دسته‌های آموزش را وصف می‌کند؛ به این اطلاعات جانبی *توصیف* می‌گوییم. در زمان آزمون تصاویری ارائه می‌شود که به دسته‌هایی غیر از دسته‌های آموزش تعلق دارند. به این دسته‌ها با نام *دسته‌های آزمون* یا *دسته‌های دیده‌نشده* اشاره می‌کنیم. همچنین اطلاعات جانبی مربوط به این کلاس‌ها نیز در اختیار

قرار می‌گیرد. در برخی روش‌ها فرض می‌شود توصیف دسته‌های آزمون هم در زمان آموزش قابل دسترسی است. توصیف‌ها ممکن است به صورت یک بردار از ویژگی‌های بصری [۷]، عبارات زبان طبیعی [۸، ۹، ۱۰] و یا یک دسته‌بند برای آن دسته [۱۱] باشند. بردار ویژگی مرسوم‌ترین شکل توصیف کلاس است. ویژگی‌ها با توجه به نوع مسئله و گستردگی دسته‌ها تعیین می‌شوند. اکثر ویژگی‌ها، ویژگی‌های بصری هستند مانند شکل (مانند گرد یا مستطیلی)، جنس (مانند چوبی یا فلزی) و عناصر موجود در تصویر (مانند چشم، مو، پدال و نوشته). برخی ویژگی‌ها هم ممکن است مستقیما در تصویر قابل مشاهده نباشند برای مثال در یک مجموعه دادگان که دسته‌ها انواع حیوانات هستند [۱۲]، علاوه بر ویژگی‌های بصری، ویژگی‌هایی چون اهلی بودن، سریع بودن یا گوشت‌خوار بودن هم وجود دارد.

مباحث ادامه این گزارش به این صورت است: در بخش ۲ صورت‌های مختلفی مسئله یادگیری از صفر را با توجه به نوع اطلاعات جانبی مورد استفاده بیان کرده و روش‌های پیشین ارائه شده برای حل آن‌ها را مرور می‌کنیم. در بخش ۳ یک روش پیشنهادی بیان می‌شود و نتایج عملی آن در بخش ۴ ارائه و روش‌های دیگر مقایسه می‌شود. بخش ۶ به کارهای آتی، جدول زمان‌بندی پژوهش و جمع‌بندی اختصاص دارد.

۲ کارهای پیشین

روش‌های مختلف برای یادگیری از صفر بر اساس اطلاعات جانبی مورد استفاده و نحوه برقراری ارتباط بین فضای نمونه‌ها و فضای اطلاعت جانبی با یک‌دیگر تفاوت دارند. علی‌رغم این تفاوت‌ها تلاش‌هایی برای ارائه یک صورت‌بندی یک‌پارچه از این مسئله صورت گرفته است. یک نحوه مدل‌سازی یادگیری از صفر، آن طور که در [۶] بیان شده، تبدیل آن به دو زیر مسئله است. مسئله اول یادگیری یک نگاشت از مجموعه تصاویر به یک فضای میانی که توصیف دسته‌ها در آن قرار دارند و مسئله دوم یادگرفتن یک دسته‌بند که اعضای فضای میانی را به برچسب‌ها دسته‌بندی کند. در این نحوه مدل‌سازی، فضای توصیف‌ها به همراه نگاشتی یک به یک به برچسب‌ها، داده‌شده فرض می‌شود. این درحالی‌ست که بسیاری از اوقات، توصیف‌ها به صورت خام قابل استفاده نیستند. برای مثال وقتی اطلاع جانبی از نوع متن است را در نظر بگیرید، فضای متون فضایی با بعد بسیار بالاست و لازم است که خود به یک فضای میانی نگاشته شود. از آن‌جا که یادگیری نگاشت از توصیف‌ها به فضای میانی ممکن است به صورت هم‌زمان و با اشتراک بعضی پارامترها با سایر قسمت‌های مدل یادگرفته شود، لازم است یادگیری این نگاشت را هم جزء چارچوب ارائه شده در نظر بگیریم. این نحوه‌ی مدل‌سازی یک چارچوب کلی برای بسیاری از روش‌های ارائه شده در یادگیری از صفر خواهد بود. در این بخش، با توجه به فراگیری این چارچوب ابتدا توصیف رسمی و نمادگذاری برای آن ارائه می‌شود. سپس روش‌های ذیل این چارچوب را مرور کرده و در پایان سایر روش‌ها را بیان می‌کنیم.

۱،۲ نمادگذاری

تصاویر را با $x \in \mathbb{R}^d$ نشان می‌دهیم که d ابعاد داده را نشان می‌دهد. توصیف‌ها را با $c \in \mathbb{R}^a$ نمایش می‌دهیم. a ابعاد توصیف‌هاست. مجموعه دسته‌های دیده‌شده را با $\mathcal{S}$ و دسته‌های دیده‌نشده را با $\mathcal{U}$ و مجموعه کل برچسب‌ها را با $\mathcal{Y}$ نشان می‌دهیم که $\mathcal{Y} = \mathcal{U} \cup \mathcal{S}$. همچنین s و u به ترتیب تعداد هر کدام از دسته‌ها را نشان می‌دهد. c^y که $y \in \mathcal{U} \cup \mathcal{S}$ بردار توصیف دسته y را نشان می‌دهد.

فرض می‌کنیم در زمان آموزش $\{(x^i, y^i)\}_{i=1}^{N_s}$ شامل N_s تصویر از دسته‌های دیده شده به همراه برچسب موجود است. $X_s \in \mathbb{R}^{N_s \times d}$ مجموعه تصاویر، Y بردار برچسب‌ها با نمایش یکی یک [۵] است. هم‌چنین توصیف‌های هر کدام از دسته‌های آموزش، $C_s \in \mathbb{R}^{s \times a}$ نیز موجود است. X_u و C_u بطور مشابه برای دسته‌های آزمون تعریف می‌شوند. $(X)_i$ سطر iم از ماتریس X و x_i درایه‌ی iم از بردار x را نشان می‌دهد.

فضای میانی را با $\mathcal{M}$ و ضرب داخلی آن را با $\langle ., . \rangle$ نشان می‌دهیم. $\pi : \mathbb{R}^d \to \mathcal{M}$ و $\psi : \mathbb{R}^a \to \mathcal{M}$ نگاشت‌هایی از فضای تصاویر و توصیفات به این فضا هستند. یادگیری نگاشت‌های π و ψ ممکن است به صورت مستقل از هم انجام شود یا اینکه هم‌زمان یادگرفته شوند. در نهایت باید دسته‌بندی از $\mathcal{M}$ به برچسب‌ها داشته باشیم: $\phi : \mathcal{M} \to \mathcal{Y}$. در خیلی از موارد دسته‌بندی را تنها روی دسته‌های آزمون در

نظر می‌گیریم، یعنی برد ϕ تنها $\mathcal{U}$ را شامل می‌شود نه تمام برچسب‌ها را. در ساده‌ترین حالت ϕ یک دسته‌بند نزدیک‌ترین همسایه در نظر گرفته می‌شود، یعنی برچسب نمونه آزمون x با رابطه ۱ پیش‌بینی خواهد شد:

$$y^* = \arg\max_{y \in \mathcal{U}} \langle \pi(x), \psi(c^y) \rangle \quad (۱)$$

البته این انتخاب برای ϕ محدودیت‌های شناخته شده‌ای دارد. از جمله این که تمامی ابعاد از اهمیت یکسانی برخوردار هستند، درحالی‌که ممکن است بعض ویژگی‌ها قابلیت جداسازی بهتری داشته باشند.

چارچوب فوق را می‌توان به روش‌های احتمالی هم تعمیم داد، به این صورت که π و ψ به صورت توزیع‌های احتمال شرطی تغییر پیدا می‌کنند. این تعمیم به صورت دقیق‌تر در بخش ۱.۳,۲ بررسی خواهد شد.

۲,۲ کران خطا

تعریف و فرضیات یادگیری از صفر با حالت معمول دسته‌بندی متفاوت است. در نتیجه کران‌هایی که پایین بودن خطای دسته‌بندی را با استفاده تعداد محدودی نمونه ضمانت می‌کنند در اینجا قابل به کار بردن نیستند. برای ارائه کران‌های خطای دسته‌بندی از صفر فرض‌های ساده‌کننده‌ای به مسئله اضافه شده است. برای این منظور فرض می‌شود که یادگیری نگاشت ψ مستقل از π انجام شده و رابطه بین توصیف‌ها و برچسب دسته‌ها رابطه‌ای یک به یک است. با این دو فرض می‌توان $\psi(c^y)$ را *امضای* دسته‌ی y نامید.

در [۶] با فرض دودویی بودن هر بعد از امضای دسته‌ها، کرانی بر اساس فاصله همینگ [۶] میان امضای دسته‌ی صحیح و مقدار پیش‌بینی شده ارائه می‌شود. در [۱۳] از نتایج مشابه در حوزه تطبیق دامنه برای کران‌دار کردن خطا استفاده ارائه شده است و کران بر اساس تفاوت توزیع‌های داده‌های آموزش و آزمون به دست آمده است. در آن نوشتار راهی برای تخمین تفاوت این دو توزیع در حالت کلی ارائه نمی‌شود. تنها به دو حالت حدی اشاره می‌شود که در صورت یکسان بودن توزیع‌ها، کران ارائه شده همان کران مشهور VC [۱۴] خواهد بود. هم‌چنین درحالتی که امضای دسته‌ها بر هم کاملا عمود باشد کران برای احتمال خطا بزرگتر از یک شده و اطلاعاتی در بر ندارد.

۳,۲ پیش‌بینی ویژگی

همان‌طور که در بخش ۱ اشاره شد، بردار ویژگی مرسوم‌ترین نوع توصیف دسته‌هاست. نخستین کارها روی یادگیری از صفر در بینایی ماشین [۷، ۱۲]، روش پیش‌بینی مستقیم ویژگی‌ها را پیشنهاد داده‌اند. در این حالت سعی می‌شود بردار ویژگی از روی تصویر ورودی بازسازی شود. آن‌گاه از میان دسته‌های دیده نشده، دسته‌ای که بردار ویژگی‌اش بیشترین مشابهت را با بردار پیش‌بینی شده دارد به عنوان برچسب معرفی می‌شود. با ادبیات چارچوب معرفی شده، این روش این گونه توصیف می‌شود که فضای میانی $\mathcal{M}$ همان فضای بردار ویژگی در نظر گرفته شده است در نتیجه نگاشت ψ نگاشت همانی است و هدف تنها یادگرفتن نگاشت π است. اهمیت این روش‌ها از یک طرف بخاطر داده‌های بسیاری است که با فراداده‌ها [۷] و دنبالک‌ها [۸] همراه شده‌اند که به صورت بردار ویژگی قابل مدل‌سازی هستند. دلیل دیگری برای اهمیت این روش‌ها این است که در مواردی هم که توصیف‌ها از نوع بردار ویژگی نیستند، ابتدا ψ به صورت مستقل یادگرفته می‌شود و بعد از آن با در نظر گرفتن $\psi(c^x)$ بعنوان بردار ویژگی دسته‌ها، مسئله به حالت مورد بحث این بخش تبدیل خواهد شد.

اگر ویژگی‌ها دودویی باشند، این مسئله را می‌توان نوعی دسته‌بندی چند برچسبی [۹] دانست که مدت زیادی است در حوزه یادگیری ماشین مورد مطالعه قرار گرفته است [۱۵]. البته دسته‌بندی چندبرچسبی با یادگیری از صفر از طریق پیش‌بینی ویژگی تفاوت‌هایی دارد. در اولی خروجی الگوریتم یک بردار از برچسب‌هاست است که ترکیب‌های مختلف از وجود یا عدم وجود هر برچسب برای آن ممکن است، در دومی خروجی نهایتا یک برچسب از دسته‌های دیده نشده است و بردار ویژگی یک مقدار میانی برای رسیدن به این خروجی است. هم‌چنین همه ترکیب‌ها از ویژگی‌ها مجاز نیستند و تنها به تعداد دسته‌ها بردار ویژگی معتبر وجود دارد. در صورتی که ویژگی‌ها پیوسته باشند مسئله پیش‌بینی آن‌ها می‌تواند به صورت یک مسئله رگرسیون در نظر گرفته شود که برای در نظر گرفتن ارتباط ویژگی‌های مختلف باید با مدل‌های

رگرسیون ساختاریافته [۱۶] حل شود. روش‌های معمول رگرسیون مانند فرآیند گاوسی هر ویژگی را به صورت جداگانه یاد گرفته و ارتباط میان ابعاد در نظر گرفته نخواهد شد [۹]. مانند حالت دودویی این مسئله با یادگیری از صفر متفاوت است، در این مسئله به دنبال خطای کمتر در ویژگی‌های پیش‌بینی شده هستیم درحالی‌که در مسئله یادگیری از صفر این خطا اهمیتی ندارد و الگوریتم با دقت برچسب‌گذاری سنجیده می‌شود.

۱.۳,۲ روش‌های احتمالی

یکی از نخستین روش‌های پیش‌بینی ویژگی در [۱۲] ارائه شده است. فرض کنید در زمان آموزش نمونه‌های $\{(x_i, y_i)\}_{i=1}^{N_s}$ به همراه بردار ویژگی دسته‌های آموزش C_y در اختیار قرار گرفته است. در نسخه اول این روش که DAP [۱۰] نام دارد استفاده از داده‌های آزمون تنها به صورت یادگیری دسته‌بندهایی برای هر یک از ویژگی‌هاست. این یادگیری با فرض استقلال ابعاد ویژگی‌ها انجام می‌شود، یعنی $P(c|x) = \prod_{i=1}^{d} P(c_i|x)$. و هر یک از $P(c_i|x)$ها با یک رگرسیون منطقی [۱۱] روی کل داده‌ها (مستقل از برچسب آن‌ها) تخمین زده می‌شود. همچنین احتمال پیشین وقوع هر یک از ویژگی‌ها، $P(c_i)$، به صورت تجربی [۱۲] با توجه به تعداد وقوع تعیین می‌شود. رابطه بین توصیف‌ها و برچسب‌ها قطعی در نظر گرفته شده است. یعنی $P(u|c) = \frac{p(u)\mathbb{I}(c=c^u)}{p(c^u)}$ که $\mathbb{I}(t)$ وقتی که شرط t برقرار باشد برابر ۱ و در غیر این صورت صفر است. در نهایت احتمال پسین هرکدام از برچسب‌های آزمون $u \in \mathcal{U}$ از این رابطه بدست می‌آید:

$$P(u|x) = \sum_{c} P(u|c)P(c|x) = \frac{P(u)}{P(c^u)} \prod_{i=1}^{a} P(c_i^u|x) \propto \prod_{i=1}^{a} \frac{P(c_i^u|x)}{P(c_i^u)} \quad (۲)$$

مقدار صورت در این رابطه همان‌طور که گفته شد از داده‌های آموزش تخمین زده می‌شود و مخرج که احتمال پیشین رخ‌داد هر ویژگی است به صورت تجربی محاسبه می‌شود. در نسخه دیگر این روش که IAP [۱۳] نام دارد تخمین $P(c_i|x)$ تغییر داده می‌شود؛ به این صورت که ابتدا یک دسته‌بند چند دسته‌ای یعنی $P(y_k|x)$ روی داده‌ها یاد گرفته می‌شود و سپس رابطه ویژگی‌ها و برچسب‌ها به صورت قطعی مدل می‌شود:

$$P(c_i|x) = \sum_{k=1}^{s} P(y_k|x)\mathbb{I}(c_i = c_i^{y_k}) \quad (۳)$$

در نهایت در هر دو روش برچسب نهایی با تخمین MAP [۱۴] از رابطه زیر تعیین می‌شود:

$$\hat{y} = \arg\max_{u \in \mathcal{U}} P(u|x) = \arg\max_{u \in \mathcal{U}} \prod_{i=1}^{a} \frac{P(c_i^u|x)}{P(c_i^u)} \quad (۴)$$

علاوه بر این دو نسخه، این روش به حالت‌های دیگری هم توسعه داده شده است. برای مثال در [۱۷] وزن‌دهی متفاوت برای مدل‌سازی اهمیت هر کدام از ویژگی‌ها به مدل اضافه شده است. این روش دو کمبود مهم دارد، اول این که فرض استقلال میان ویژگی‌ها بسیار غیر واقعی است. برای مثال ویژگی‌های بصری خاک و صحرا وابستگی واضحی وجود دارد. مشکل دوم این است که یادگیری دسته‌بندها برای هر ویژگی بدون توجه به مراحل بعدی و نتایج سایر دسته‌بندهاست؛ درحالی‌که خروجی هر دسته‌بند در دسته‌بندی دیگری استفاده خواهد شد و معیار ارزیابی، عمل‌کرد خطای دسته‌بند دوم است، یعنی خطای پیش‌بینی ویژگی‌ها به طور مستقیم اهمیت ندارد. نویسندگان [۱۸] برای حل این مشکل پیشنهاد می‌کنند فرض یک به یک بودن نگاشت بین بردارهای ویژگی و برچسب‌ها را در نظر نگیریم. در این روش پیش‌بینی ویژگی‌ها مانند مدل DAP با رگرسیون منطقی انجام می‌شود با این تفاوت که یادگیری پارامترهای آن‌ها و ϕ به صورت مشترک انجام می‌شود. ϕ یک نگاشت خطی در نظر گرفته می‌شود: $\phi(\pi(x)) = R\pi(x)$. دو محدودیت روی مقادیر R اعمال می‌شود. یک محدویت سطری و یک محدودیت ستونی. محدودیت سطری مانع از این می‌شود که فاصله همینگ سطرها از حدی کمتر بشود. دقت کنید که در یک دسته‌بند خطی به شکل بالا، هر سطر را می‌توان مرکز ثقل نمونه‌های دسته‌ی متناظر آن سطر تعبیر کرد. در نتیجه این محدودیت تضمین می‌کند که

بردار ویژگی نماینده هر دسته با دسته‌های دیگر متفاوت باشد. محدودیت ستونی یک مقدار حداکثری برای همبستگی میان ستون‌ها در نظر می‌گیرد تا به این صورت اطلاعات تکراری در ویژگی‌ها وجود نداشته باشد. نویسندگان این مقاله استدلال می‌کنند که با این دو محدودیت باعث حذف ویژگی‌های تکراری و ویژگی‌های غیر بصری (مانند بدبو بودن) خواهد شد.

نویسندگان [۱۹] برای در نظر گرفتن ارتباط بین ویژگی‌ها و ارتباط ویژگی‌ها با برچسب نهایی روش‌های مدل‌سازی موضوع [15] را از حوزه یادگیری در متن اقتباس می‌کنند. همچنین نویسندگان [۲۰] برای این کار یک چارچوب بر اساس مدل‌های گرافی احتمال معرفی می‌کنند. در این چارچوب یک شبکه بیزی[16] برای مدل کردن این روابط در نظر گرفته می‌شود و ساختار آن که نشان‌دهنده وابستگی یا استقلال ویژگی‌ها با هم یا با برچسب است، با کمک روش‌های یادگیری ساختار [17] شناخته می‌شود.

۲.۳,۲ نگاشت‌های خطی

چند روش اخیر وجود دارد که علی‌رغم ساده بودن نتایج بهتری از روش‌های قبلی کسب کرده‌اند. در این روش‌ها نگاشت ψ همانی، دسته‌بند ϕ دسته‌بند نزدیک‌ترین همسایه و نگاشت π خطی (به صورت $\pi(x) = xW$) در نظر گرفته شده‌اند. اما معرفی توابع هزینه یا جمله‌های منظم‌سازی [18] هوشمندانه‌تر باعث شده که نتایج بهتری به دست بیاورند. یکی از این روش‌ها که در [۲۱] معرفی شده، تابع هزینه‌ای ارائه می‌دهد که هم خطای دسته‌بندی، هم خطای پیش‌بینی ویژگی‌ها را در نظر می‌گیرد. این تابع هزینه چنین شکلی دارد:

$$L(W) = \frac{1}{N_s} \sum_{n=1}^{N_s} \lambda_{r_\Delta(x_n,y_n)} \sum_{y \in \mathcal{Y}} \max(0, l(x_n, y_n, y)) \tag{۵}$$

$$l(x_n, y_n, y) = \mathbb{I}(y \neq y_n) + F(x_n, c_y; W) - F(x_n, c_{y_n}; W) \tag{۶}$$

که در آن $r_\Delta(x_n, y_n) = \sum_{y \in \mathcal{Y}} \mathbb{I}(l(x_n, y_n, y) > 0)$، F یک تابع رتبه‌بندی و λ_k یک تابع نزولی از k است. این تابع، پیش‌بینی اشتباه ویژگی‌ها را این گونه جریمه می‌کند که به ازای برچسب نادرستی که رتبه بالاتری از برچسب صحیح در دسته‌بندی دریافت کرده، جریمه‌ای متناسب با امتیاز برچسب ناصحیح در نظر گرفته می‌شود.ضریب نزولی λ_k میزان جریمه را برای برچسب‌های غلط در رتبه‌های بالا بیشتر در نظر می‌گیرد.

یک روش دیگر که در [۱۳] ارائه شده، نگاشت‌های مشابهی را استفاده می‌کند. همچنین تابع هزینه آن شکل ساده نرم ۲ را دارد. مسئله‌ی بهینه‌سازی تعریف شده به این شکل است:

$$\underset{W \in \mathbb{R}^{d \times a}}{\text{minimize}} \left\| X_s W C_s^T - Y \right\|_{Fro}^2 + \Omega(W) \tag{۷}$$

که $\Omega(W)$ یک جمله منظم‌سازی است که به این صورت تعریف می‌شود:

$$\Omega(W; X_s, C_s) = \gamma \left\| W C_s^T \right\|_{Fro}^2 + \lambda \left\| XW \right\|_{Fro}^2 + \beta \left\| W \right\|_{Fro}^2 \tag{۸}$$

γ، λ و β فراپارامترهایی هستند که اهمیت هر یک از جملات را تعیین می‌کنند. تابع هزینه فوق تنها دسته‌بندی اشتباه را جریمه می‌کند. مناسب نبودن تابع هزینه نرم ۲ برای خطای دسته‌بندی مسئله‌ای شناخته شده در یادگیری ماشین است و عملکرد خوب این تابع در این روش شاید در نگاه اول عجیب بنظر برسد. اگر در جمله منظم‌سازی تعریف شده دقت کنیم این مسئله روشن‌تر خواهد شد. علت نامناسب بودن تابع هزینه نرم ۲ این است که حتی دسته‌بندی‌های صحیح را اگر مقداری غیر از مقدار تعیین شده (معمولا یک) داشته باشند، به اندازه فاصله‌شان از این مقدار جریمه می‌کنید. اما جمله منظم‌سازی تعریف شده اصولا مانع بزرگ شدن مقدار پیش‌بینی شده خواهد شد. جمله اول در معادله (۸) را می‌توان اندازه بردار تصویر متوسط برای هر دسته دانست. جمله دوم مقدار بردار ویژگی پیش‌بینی شده برای هر دسته است و جمله سوم هم که یک جمله معمول است که پارامترهای نگاشت را کنترل می‌کند. در زمان آزمون برای نمونه x مقدار xWC_u را

محاسبه کرده و دسته‌ای که درایه‌ی متناظرش بیشترین مقدار را دارد به عنوان پیش‌بینی معرفی می‌کنیم. یک ویژگی این روش این است که با انتخاب $\beta = \gamma\lambda$ در معادله (۸) بهینه‌سازی معادله (۷) جواب بسته خواهد داشت؛ در نتیجه زمان اجرای این روش بسیار کمتر از سایر روش‌هایی است که مرور شد.

یک روش خطی دیگر که مستقیم از ویژگی‌ها استفاده نمی‌کند، کاری است که در [۲۲] معرفی شده است. این روش تنها از نام هر دسته به عنوان توصیف بهره می‌برد. در این روش نام‌ها، مستقل از اطلاعات دیگر مسئله، به بردارهایی نگاشته می‌شوند، بردارهای حاصل را می‌توان مانند بردار ویژگی در سایر مسائل به حساب آورد؛ در نتیجه این روش را ذیل عنوان پیش‌بینی ویژگی مرور می‌کنیم. این روش ابتدا برای بدست‌آوردن بردارهای مربوط به نام‌ها از مدل مشهور word2vec [۲۳] با پیش‌آموزش روی مقالات ویکی‌پدیای انگلیسی استفاده می‌کند، هم‌چنین برای ویژگی‌های تصویر از شبکه عصبی برنده چالش ILSVRC 2012، AlexNet استفاده می‌کند. * این روش نیز π را خطی و دسته‌بند ϕ را نزدیک‌ترین همسایه در نظر می‌گیرد. تابع هزینه مورد استفاده از این روش یک تابع هزینه‌ی رتبه‌بند است به این معنی که مانند [۲۱] به ازای برچسب‌هایی که امتیاز بیشتری نسبت به برچسب صحیح کسب کرده‌اند، جریمه در نظر می‌گیرد:

$$L((x_n, y_n); W) = \sum_{y \neq y_n} \max(0, \text{magrin} - x_n W c_{y_n} + x_n W c_y) \tag{۹}$$

۴,۲ یادگیری دسته‌بند

روشی ارائه شده در [۹] برای نخستین بار، از استفاده از متونی در مورد هر دسته را به عنوان توصیف در نظر گرفته و مجموعه دادگانی برای این موضوع فراهم می‌آورد. در این روش هدف یافتن یک دسته‌بند دودویی (رد یا قبول) برای هر دسته از روی توصیف متنی آن است. با توجه به این که دسته‌بند مورد نظر خطی فرض شده است، می‌توان این روش را این گونه هم تعبیر کرد که به دنبال نگاشتن متون توصیف‌کننده به فضای تصاویر است. یعنی فضای میانی فضای تصاویر در نظر گرفته شده و پس از نگاشتن توصیف‌ها به آن فضا برچسب‌ها با دسته‌بند نزدیک‌ترین همسایه مشخص می‌شود. این روش برای یافتن دسته‌بندهای دسته‌های دیده نشده، همزمان دو رویکرد رگرسیون احتمالی [۱۹] و تطبیق دامنه را استفاده می‌کند. طبق نمادگذاری تعریف شده c_u یک توصیف برای دسته دیده نشده و $\psi(c_u)$ نگاشت آن به فضای تصویر (یا به تعبیر دیگر دسته‌بند دسته‌ی مربوط به دسته‌ی c_u) است. ماتریس تطبیق دامنه W با استفاده از نمونه‌های آموزش طوری یادگرفته شده است که اگر c و x متعلق به یک دسته باشند، $c^T W x$ از آستانه‌ای مانند t بیشتر است. مسئله بهینه‌سازی تعریف شده برای محاسبه $\psi(c_u)$ به این صورت است:

$$\begin{aligned} \psi(c_u) = \arg\min_{h,\zeta_i} \{ & h^T h - \alpha c_u^T W h - \beta \ln(P(h|c_u)) + \gamma \sum \zeta_i \} \\ s.t: & -(h^T)(X_s)_i \geq \zeta_i, \quad \zeta_i \geq 0, \quad i = 1, \ldots, N_s \\ & c^T W c \geq t \end{aligned} \tag{۱۰}$$

α، β و γ فراپارامتر هستند. جمله اول در معادله (۱۰) یک عبارت منظم‌سازی است، جمله دوم مربوط به رویکرد تطبیق دامنه است که دسته‌بند تخمین زده شده را $c_y^T W$ می‌داند و جمله سوم مربوط به رگرسیون احتمالی است که دسته‌بند را به سمت عبارت حاصل از این رگرسیون سوق می‌دهد. محدودیت روی ζ_iها اجبار می‌کند که این دسته‌بند به نمونه‌ای از دسته‌های آموزش برچسب مثبت اختصاص ندهد. نویسندگان این پژوهش یک نسخه با امکان استفاده از هسته نیز از کار خود در [۲۴] ارائه می‌دهند. بر مبنای چارچوب همین پژوهش در [۲۵] از شبکه‌های عصبی برای تخمین زدن دسته‌بند برای دسته‌های دیده نشده استفاده شده است. به این صورت که یک شبکه عصبی پیش‌آموزش دیده برای استخراج ویژگی از تصاویر استفاده می‌شود تا ابعاد یا تعداد پارامترهای دسته‌بندی که لازم است تخمین زده شود کم

* استفاده از مقادیر نورون‌های لایه چگال اول شبکه‌های عصبی به عنوان ویژگی‌های بصری در بسیاری از روش‌های دیگر نیز صورت گرفته است؛ در نتیجه این قسمت جزیی از نگاشت π در نظر گرفته نمی‌شود، بلکه این مقادیر را به عنوان مجموعه تصاویر (X) تلقی می‌کنیم.

شود. آن‌گاه از یک شبکه عصبی دیگر برای نگاشت متن به برداری در فضای ویژگی‌های تصویر (یا همان دسته‌بند دسته‌ی مربوط به آن متن) استفاده می‌شود. این پژوهش همچنین دسته‌بند پیچشی [۲۰] را معرفی می‌کند. فرض کنید x'_i مقادیر ویژگی‌نگار [۲۱] iم آخرین لایه پیچشی یک CNN [۲۲] با K ویژگی‌نگار، برای ورودی x باشد. در این صورت یک دسته‌بند پیچشی مانند r در حقیقت یک صافی پیچشی [۲۳] است و ابعاد آن برابر $K \times s \times s$ است که s اندازه صافی است که یک فراپارامتر است. برچسبی که r برای نمونه x پیش‌بینی می‌کند از این رابطه به دست خواهد آمد:

$$\hat{y} = o\big(\sum_i r_i * r_i\big) \tag{۱۱}$$

که $o(.)$ یک تابع ادغام [۲۴] است که به طور معمول در شبکه‌های پیچشی مورد استفاده قرار می‌گیرد. یادگیری r به این صورت خواهد بود که رابطه بالا کمترین خطا را در پیش‌بینی برچسب روی نمونه‌های آموزش داشته باشد. خطای در نظر گرفته شده برای بدست آوردن r، آنتروپی متقابل [۲۵] است.

یک روش‌دیگر که از چنین رویکردی استفاده می‌کند در [۲۶] معرفی شده است. در این روش توصیف هردسته، پارامترهای یک دسته‌بند برای آن در نظر گرفته شده. دسته‌بند یک دسته دیده نشده بر حسب جمع وزن‌دار دسته‌های دیده شده بیان می‌شود. وزن‌های این جمع وزن‌دار از امتیاز شباهت دسته‌ها با هم بدست می‌آید. امتیاز شباهت، با استفاده از تعداد رخ‌داد همزمان نام‌های برچسب‌ها در یک مجموعه متن سنجیده می‌شود. نویسندگان این پژوهش همچنین مسئله یادگیری از صفر چندبرچسبی را معرفی می‌کنند. در این چارچوب برچسب‌ها یا ویژگی‌ها به صورت یک برچسب که یک دسته‌بند برای آن موجود است در نظرگرفته می‌شود و با استفاده از روشی که معرفی شد می‌توان از آن‌ها در به دست آوردن دسته‌بندی برای یک برچسب بی‌نمونه استفاده کرد.

۵,۲ نگاشت به فضای دسته‌های دیده شده

یک انتخاب محبوب برای فضای میانی $\mathcal{M}$ فضایی با ابعاد تعداد دسته‌های دیده شده است. در نگاشت به این فضا سعی می‌شود تصاویر یا توصیفات دسته‌های آزمون بر حسب نسبت‌هایی از دسته‌های دیده بیان شود. یکی از روش‌هایی که از چنین نگاشتی استفاده می‌کند، روشی است که نویسندگان [۲۷] در ادامه کار پیشین خود [۲۲]، که در بخش ۲.۳,۲ مرور شد، ارائه می‌دهند. در این روش بجای استفاده از مقادیر نورون‌های میانی شبکه AlexNet از خروجی آخرین لایه این شبکه، یعنی لایه‌ی softmax استفاده می‌کنند. این لایه به تعداد دسته‌های دیده شده نورون دارد و تعبیر مقادیر این لایه، امتیازی است که شبکه برای تعلق تصویر به هر دسته می‌دهد. پس بدست آوردن این نمایش برای تصویر در فضای میانی، این نمایش به فضای توصیف‌ها که بردارهای متناظر با نام دسته‌هاست نگاشته می‌شود؛ به این صورت که بردارهای نام دسته‌های دیده شده با این وزن‌ها با یکدیگر جمع شده و حاصل با استفاده از دسته‌بندی نزدیک‌ترین همسایه برچسب یک دسته دیده نشده را معین می‌کند.

یک روش اخیر معرفی شده در [۲۸] که نتایج را به طرز قابل توجهی بهبود داده است نیز از این فضا به عنوان فضای میانی استفاده می‌کند. نگاشت بردارهای ویژگی به این فضا با حل معادله زیر انجام می‌شود:

$$\psi(c) = \underset{\boldsymbol{\alpha}\in\Delta^s}{\arg\min} \left\{ \gamma \left\|\boldsymbol{\alpha}\right\|^2 + \left\| c - \sum_{y\in\mathcal{S}} c_y \alpha_y \right\|^2 \right\} \tag{۱۲}$$

که $\boldsymbol{\alpha}$ برداری به اندازه تعداد دسته‌های دیده است و هر درایه‌ی $\boldsymbol{\alpha_y}$ آن نسبت دسته $\boldsymbol{y}$ را در تشکیل دسته دیده نشده تعیین می‌کند. $\boldsymbol{\Delta^s}$ تک‌جهتی [۲۶] $\boldsymbol{s}$ بعدی است، یعنی این نگاشت در حقیقت یک بافت‌نگاره [۲۷] از دسته‌های دیده شده تولید می‌کند. نگاشت تصاویر به این فضا از یک مسئله بهینه‌سازی محدودیت‌دار بدست می‌آید. برای هر بعد از این نگاشت، یک نگاشت دیگر ویژه هر دسته دیده شده یادگرفته می‌شود که میزان حضور آن دسته را در تصویر مشخص می‌کند.

۶,۲ یادگیری نگاشت‌ها از دو دامنه

در اکثر روش‌هایی که تا کنون مرور شد، فضای میانی همان فضایی که توصیف‌ها در آن هستند در نظر گرفته می‌شد یا این‌که نگاشت از دامنه توصیف‌ها به طور مستقل از مسئله (برای مثال با استفاده از اطلاعات یک مجموعه متن) یاد گرفته می‌شد؛ چنین رویکردی دارای این ضعف آشکار است که نمایش بدست آمده برای برچسب‌ها جدا کننده [۲۸] نباشد. از میان روش‌هایی که دیدیم [۱۸، ۲۰] با پیچیده‌تر کردن ساختار دسته‌بند ϕ یا به عبارتی به هم زدن رابطه *امضا بودن* بردارهای ویژگی برای دسته‌ها سعی در حل این مشکل داشتند. یک روش اخیر [۲۹] با الهام از راه‌کارهای یادگیری نیمه‌نظارتی روشی برای یادگیری یک دسته‌بند چند دسته‌ای و نگاشت‌های برای برچسب‌ها به صورت هم‌زمان ارائه می‌دهد. در این روش دسته‌بند یادگرفته شده برای تمامی دسته‌ها (و نه تنها برای دسته‌های آزمون) است. توصیف‌هایی که برای دسته‌ها وجود دارد می‌توانند بعنوان یک مقدار پیشین [۲۹] برای نمایش برچسب‌ها در فضای میانی در نظر گرفته شوند. این پژوهش هم‌چنین تاثیر ابعاد فضای میانی را روی دقت دسته‌بندی برای دو مجموعه داده بررسی می‌کند، در ابعاد بررسی شده (بین ۲۰ تا ۱۰۰ بعد) دقت تابعی صعودی از ابعاد است. ابعاد تصاویر در این بررسی ۲۰۰۰ برای یک مجموعه داده و ۱۵۰۰ برای مجموعه داده دیگر بوده است. رویکردی مشابه در [۳۰] از همین گروه ارائه شده که چارچوب مسئله یادگیری نیمه‌نظارتی را به یادگیری از صفر تبدیل می‌کند با این تفاوت که نمایش برچسب‌ها در فضای میانی ثابت فرض می‌شوند و بردارهای مربوط به نام دسته‌ها هستند.

۷,۲ سایر روش‌ها

روشی که در [۳۱] معرفی شده و تا کنون بهترین نتایج را روی مجموعه‌دادگانی که توصیف دسته‌ها از نوع بردار ویژگی بدست آورده از رویکردی کاملا متفاوت بهره می‌برد. در این روش تنها یک دسته‌بند ساخته می‌شود که دو ورودی دارد: یک تصویر و یک توصیف و مقدار خروجی که مقداری دودویی است مشخص می‌کند که تصویر و توصیف ورودی متعلق به یک دسته هستند یا خیر. $\boldsymbol{y^{xc}}$ را به صورت یک متغیر دودویی که اگر $\boldsymbol{x}$ و $\boldsymbol{c}$ متعلق به یک دسته باشند یک و در غیر این صورت صفر است، تعریف می‌کنیم. آماره‌ی کافی برای دسته‌بند مورد نظر $\boldsymbol{P(y^{xc}|x,c)}$ است. برای تخمین این احتمال از دو متغیر نهان کمک گرفته می‌شود. این متغیرها یک زنجیر مارکف تشکیل می‌دهند که رابه (۱۳) نشان داده شده است.

$$\boldsymbol{X \leftrightarrow Z^{(x)} \leftrightarrow Y \leftrightarrow Z^{(c)} \leftrightarrow C.} \tag{۱۳}$$

با توجه به (۱۳) مشخص است که با داشتن برچسب دسته‌ها متغیرهای تصادفی تصاویر و توصیف‌ها و نمایش نهان آن‌ها از یک‌دیگر مستقل هستند در نتیجه احتمال پسینی به این صورت جدا می‌شود:

$$\boldsymbol{p(y^{(xc)}, z^{(x)}, z^{(c)} \mid x, c) = p(y^{(xc)} \mid z^{(x)}, z^{(c)}) p(z^{(x)}, z^{(c)} \mid x, c)}$$

هم‌چنین فرض می‌شود که در غیاب اطلاعی در مورد برچسب‌ها نمایش‌های نهان از هم مستقل هستند، یعنی $\boldsymbol{p(z^{(x)}, z^{(c)}) \approx p(z^{(x)}) p(z^{(c)})}$ نهایتا دسته‌بندی با میانگین‌گیری روی متغیرهای نهان مقدور خواهد بود:

$$\boldsymbol{p(y^{(xc)} \mid x, c) = \int\int p(z^{(x)}|x) p(z^{(c)}|c) p(y^{(xc)}|z^{(x)}, z^{(c)}) dz^{(x)} dz^{(c)}} \tag{۱۴}$$

در ادامه این روش از محاسبه این انتگرال صرف‌نظر شده و یک کران پایین از آن جایگزین آن شده است:

$$\boldsymbol{\log p(y^{(xc)} \mid x, c) \geq \max_{z^{(x)}, z^{(c)}} \log p(z^{(x)}|x) p(z^{(x)}|c) p(y^{(xc)}|z^{(x)}, z^{(c)})} \tag{۱۵}$$

در [۳۲] انواع نگاشت‌های مختلفی که برای برچسب‌ها با استفاده از روش‌های غیرنظارتی وجود دارد بررسی و ارزیابی می‌شود. هم‌چنین عمل‌کرد ویژگی‌های مختلف تصویر در یادگیری از صفر بررسی می‌شود. نویسندگان این پژوهش برای مقایسه این نگاشت‌ها از آن‌ها در روشی که خود معرفی کردند [۲۱] و ما آن را در بخش ۲.۳,۲ شرح دادیم استفاده کرده‌اند.

در [۳۳] برای اولین بار یادگیری از صفر به صورت یک مسئله تطبیق‌دامنه‌ی بدون نظارت از دامنه داده‌های آموزش به دامنه داده‌های آزمون مدل شده است. صورت مسئله یادگیری از صفر با مسئله تطبیق دامنه متفاوت است چرا که دامنه مقصد در مسئله یادگیری از صفر مجموعه برچسب‌هایی متفاوت از دامنه مبدا دارد. برای تبدیل این مسئله به یک مسئله تطبیق دامنه، نویسندگان پژوهش مسئله نگاشت تصاویر آزمون به فضای میانی (برای مثال فضای بردارهای ویژگی) را در نظر می‌گیرند. مسئله نگاشت به این فضا را می‌توان یک مسئله تطبیق دامنه در نظر گرفت چرا که داده‌های آموزش و آزمون هر دو باید به این فضا نگاشته شوند. این مسئله جدید یک مسئله تطبیق دامنه بدون نظارت است چرا که نمونه‌های دامنه مقصد برچسبی ندارند. برای حل این مسئله‌ی تطبیق دامنه فضای میانی، فضایی با ابعاد بالا در نظر گرفته شده است و از روش‌های نمایش تنک [۳۰] و یادگیری واژه‌نامه برای یافتن نمایش دسته‌ها استفاده شده است.

در [۸] توصیف دسته‌ها، نام برچسب‌ها در نظر گرفته شده و با استفاده از یک مدل پیش‌آموزش دیده شده، آن‌ها را به بردارهای ۵۰ـ بعدی تبدیل می‌کند. سپس تصاویر را با یک شبکه عصبی دو لایه به این فضا نگاشته و با روش نزدیک‌ترین همسایه برچسب را پیش‌بینی می‌کند. تفاوت این پژوهش با سایر پژوهش‌ها این است که دسته‌بندی نهایی را تنها روی دسته‌های دیده نشده در نظر نمی‌گیرد بلکه روی کل دسته‌ها در نظر گرفته و روشی برای تشخیص این که آیا نمونه آزمون می‌تواند به دسته‌های دیده شده تعلق داشته باشد یا نه ارائه می‌دهد.

یک راه‌حل مبتنی بر جنگل‌های تصادفی [۳۱] در [۳۴] ارائه می‌شود. این روش در مرحله تشخیص ویژگی، برای هر ویژگی یک SVM به طور مستقل یاد می‌گیرد. قسمت اصلی روش پیش‌نهادی در طراحی دسته‌بند ϕ است. این دسته‌بند با جنگل‌های تصادفی ساخته می‌شود اما از آن‌جایی که ویژگی‌های پیش‌بینی شده برای تصاویر احتمالا با بردار ویژگی توصیف کننده‌ی دسته تفاوت‌هایی دارد، اجازه پیگیری همزمان چند مسیر در جنگل تصادفی داده می‌شود تا این عدم قطعیت در نظر گرفته شود.

۳ روش ارائه شده

در [۳۵] نشان داده‌شد که خروجی لایه‌ی سافت‌مکس [۳۲] شبکه‌های عصبی اطلاعاتی بیش از تنها یک دسته‌بند یکی‌یک در خود دارد. این اطلاعات می‌تواند برای بیان یک نمونه از دسته‌های دیده‌نشده بر اساس نسبت‌هایی از دسته‌های دیده شده استفاده شود. اگر خروجی چنین لایه‌ای را به ورودی z (که خروجی لایه‌ی قبل است) برابر q باشد، خواهیم داشت:

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \qquad (۱۶)$$

در زمان آموزش شبکه مقدار $T = ۱$ در نظر گرفته می‌شود، در این حالت تابع سعی می‌کند یکی از ابعاد ورودی که مربوط به دسته‌ی پیش‌بینی شده است را یک و سایر ابعاد را صفر کند بعبارتی مقدار q_i بیشینه اختلاف زیادی با سایر q_jها خواهد داشت. با افزایش مقدار T که از آن به نام دما نیز یاد می‌شود، اختلاف پدید آمده نرم‌تر می‌شود و مقدار بعد بیشینه به سایر ابعاد نزدیک می‌شود. در روش پیشنهادی ما از چنین لایه‌ای برای نگاشت تصاویر آزمون به هستوگرام‌هایی از دسته‌های دیده‌شده استفاده می‌کنیم. به این صورت که ابتدا یک شبکه عصبی چند لایه با اتصالات چگال و فعال‌سازی [۳۳] سافت‌مکس در لایه آخر به عنوان یک دسته‌بند روی دسته‌های دیده‌شده آموزش داده می‌شود. سپس از همین شبکه بدون تغییری در وزن‌های یادگرفته شده و تنها با جایگیزین کردن فعال‌سازی لایه‌ی آخر با تابع (۱۶) با مقدار $T > ۱$ برای نگاشت تصاویر آزمون به فضای دسته‌های دیده‌شده استفاده می‌شود. برای فعال‌سازی لایه‌های پایین‌تر از واحد تصحیح خطی [۳۴] استفاده شده است. هم‌چنین بین تمامی لایه‌های میانی منظم‌سازی حذف تصادفی [۳۵] وجود دارد که احتمال حذف آن با اعتبارسنجی متقابل تعیین می‌شود.

تعداد لایه‌های شبکه و اندازه هر لایه برای هر مجموعه داده با اعتبار سنجی متقابل [۳۶] تعیین می‌شود. مقدار T نیز با اعتبار سنجی متقابل به صورت حذف بعضی از دسته‌های دیده شده از جریان آموزش و محک زدن مدل با آن‌ها به عنوان دسته‌های دیده‌نشده قابل تعیین است اما در آزمایش‌های انجام شده مقدار آن را ثابت برابر ۲۰ در نظر گرفته‌ایم.

برای نگاشت بردارهای ویژگی به هیستوگرام‌هایی از دسته‌های دیده‌شده، از مجموع عکس فاصله‌های اقلیدسی و بلوکی [۳۷] بردارهای ویژگی استفاده شده است:

$$\psi_i(c^u) = \sum_{u \neq y} \frac{1}{\|c_i^u - c_i^y\|_2 + \|c_i^u - c_i^y\|_1} \quad (۱۷)$$

در نهایت دسته‌بندی به صورت دسته‌بندی نزدیک‌ترین همسایه (با فاصله اقلیدسی) انجام می‌شود.

۴ نتایج پیاده‌سازی

روش ارائه شده در بخش ۳ روی سه مجموعه داده‌ی استاندارد در ادبیات یادگیری از صفر یعنی آوا [۳۸] [۱۲] ، سان [۳۹] [۳۶] و پاسکال و یاهو با ویژگی [۴۰] [۷] آزمایش شده است. ویژگی‌های این سه مجوعه داده در جدول ۱,۴ آمده است.

جدول ۱,۴: *مشخصات مجموعه دادگان مورد استفاده*

مجموعه داده	تعداد تصاویر	تعداد دسته‌های آموزش	تعداد دسته‌های آزمون	ابعاد ویژگی
AwA	۳۰۴۷۵	۴۰	۱۰	۸۵
aPascal/Yahoo	۱۵۳۳۹	۲۰	۱۲	۶۴
SUN	۱۴۳۴۰	۷۰۷	۱۰	۱۰۲

نتیجه معیار دقت دسته‌بندی چنددسته‌ای [۴۱] روی این سه مجموعه داده برای روش پیشنهادی و روش‌های پایه در جدول ۲,۴ آمده است.

جدول ۲,۴: *نتایج بر اساس معیار دقت دسته‌بندی (درصد)، به صورت میانگین $\pm$ انحراف معیار. نتایج مربوط به سایر روش‌ها از [۲۸] گزارش شده است.*

روش	AwA	aPY	SUN
IAP [۱۲]	۴۴/۵	۳۸/۱۶	۸۲
DAP [۱۲]	۵۳/۲	–	۷۲
ESZSL [۱۳]	۶۲/۸۵	۲۷/۳	۶۵/۷
پیشنهادی	۶۳/۷۵ $\pm$ ۰/۶۷	۳۹/۴ $\pm$ ۰/۶۳	۶۸/۴

۵ کارهای آتی

در ادامه مسیر تحقیق ایده‌ی مطرح شده برای بدست آوردن نگاشت تصاویر در فضای میانی تکمیل خواهد شد تا نزدیکی نگاشت‌های بدست آمده به نگاشت‌های مربوط به توصیف‌ها در جریان یادگیری لحاظ شود. موضوع دیگری که دنبال خواهد شد استفاده از بردارهای ویژگی برای هر تصویر آموزش (در مقابل یک بردار ویژگی برای هر دسته) و وجود چندین توصیف برای یک دسته آزمون است؛ چرا که

این اطلاعات برای اکثر مجموعه داده‌های واقعی در دسترس قرار دارد. علاوه بر این، به استفاده از توصیف‌های متنی و هم‌چنین اطلاعات داده‌های بدون برچسب نیز خواهیم پرداخت. دو زیر بخش آتی به کارهای آتی در این دو حوزه اختصاص دارد.

۱,۵ توصیف از نوع متن

با توجه به اینکه متون توصیف‌هایی با قابلیت دسترسی بیشتر نسبت به بردارهای ویژگی هستند و جمع‌آوری آن‌ها از منابعی مانند دائره‌المعارف‌ها بدون هزینه و دخالت نیروی انسانی امکان‌پذیر است، این نوع توصیف گزینه‌ی بسیاری مناسب‌تری برای یادگیری از صفر در مقیاس بزرگ است. از طرفی متن‌ها دارای ساختار خاصی هستند و برخلاف بردارهای ویژگی، استفاده مستقیم از آن‌ها امکان‌پذیر نخواهد بود. یکی از کارهای آتی تلاش برای یافتن نگاشتی‌ست که ویژگی‌های بیان شده در متن را به ویژگی‌های بصری استخراج شده از تصاویر مربوط کند. با توجه به شباهت این کار به ترجمه متن به زبانی دیگر، از مدل‌های موفق ارائه شده برای ترجمه‌ی خودکار که مبتنی شبکه‌های عصبی بازگردنده [۴۲] برای این کار استفاده خواهد شد. هم‌چنین با توجه به اینکه استخراج ویژگی از تصاویر با استفاده از شبکه‌های پیش‌آموزش دیده نتایج بسیار بهتری به دنبال دارد، برای ساخت آموزش دادن یک شبکه برای استخراج ویژگی از متن با استفاده از متون فراوان موجود روی اینترنت (مانند ویکی‌پدیا [۴۳]) تلاش خواهد شد.

۲,۵ یادگیری نیمه‌نظارتی

بعضی از روش‌های اخیر [۳۳، ۲۹، ۳۰] این فرض که داده‌های آزمون در زمان آموزش نیز موجود هستند و تنها برچسب ندارند را اضافه کرده‌اند و البته این فرض در اکثر کاربردهای واقعی برقرار است. با این فرض امکان استفاده از اطلاعاتی که در نمونه‌های بدون برچسب در مورد ساختار و توزیع داده‌ها وجود دارد، فراهم می‌شود. یک رویکرد برای استفاده از داده‌های آزمون، حل هم‌زمان یک مسئله دسته‌بندی روی داده‌های آموزش و یک مسئله خوشه‌بندی روی داده‌های آزمون است که در [۳۰] به کار گرفته شده است، اما بنظر می‌رسد تابع هزینه معرفی شده برای مدل کردن مسئله فوق نزدیکی خوشه‌ها به توصیف‌های دسته‌های آزمون را در نظر نمی‌گیرد که وارد کردن آن به تابع هزینه می‌تواند در بهبود نتایج موثر باشد.

خلاصه‌ای از مراحل و میزان پیشرفت پروژه در جدول ۳,۵ آمده است.

جدول ۳,۵: *جدول زمان‌بندی*

عنوان فعالیت	مدت زمان لازم	درصد پیشرفت	زمان اتمام
مطالعه و بررسی روش‌های موجود و راه‌کارهای قابل استفاده	۳ ماه	۱۰۰	شهریور ۹۴
آزمایش روش‌های موجود بر روی مجموعه داده‌های معرفی شده در مقالات و مقایسه آن‌ها	۲ ماه	۱۰۰	آبان ۹۴
بررسی و یافتن کاستی‌های روش‌های موجود	۱ ماه	۶۰	آبان ۹۴
پیشنهاد و پیاده‌سازی و ارزیابی روش جدید	۴ ماه	۲۰	اسفند ۹۴
ارزیابی روش نهایی و مقایسه با روش‌های دیگر	۲ ماه	۰	اردیبهشت ۹۵
نگارش پایان‌نامه	۲ ماه	۰	تیر ۹۵

۶ جمع‌بندی

در این گزارش مسئله یادگیری از صفر به همراه نسخه‌های مختف آن و یک چارچوب کلی برای مسئله یادگیری از صفر معرفی شد. سپس به معرفی روش‌های ارائه شده برای حل این مسئله پرداختیم. با توجه به جدید بودن این مسئله و اینکه اکثر روش‌هایی که مرور شد در چندماه اخیر ارائه شده‌اند، تقسیم‌بندی استانداری از روش‌ها صورت نگرفته است. در این گزارش سعی شد برای روش‌ها یک تقسیم‌بندی بر اساس انتخاب فضای میانی، نوع نگاشت‌ها به این فضا و نوع دسته‌بند مورد استفاده ارائه شود. برخی از روش‌های پیشین مطرح در جدول ۴,۶ به طور خلاصه ذکر شده‌اند. در بخش ۳ فضای دسته‌های آموزش را به عنوان فضای میانی در نظر گرفته و روشی برای تخمین وزن‌ها در این فضا ارائه دادیم. در نهایت راه‌کاری آتی و جدول زمان‌بندی ادامه‌ی کار در بخش ۵ ارائه شد.

مراجع

[1] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[2] E. G. Miller, *Learning from one example in machine vision by sharing probability densities*. Ph.D. thesis, MIT, 2002.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol.22, pp.1345–1359, 2010.

[4] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *National Conference on Artificial Intelligence (AAAI)*, pp.646–651, 2008.

[5] R. Salakhutdinov, A. Torralba, and J. Tenenbaum, "Learning to share visual appearance for multiclass object detection," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pp.1481–1488, 2011.

[6] M. Palatucci, G. Hinton, D. Pomerleau, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in Neural Information Processing Systems (NIPS) 22*, pp.1410–1418, 2009.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing Objects by Their Attributes," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pp.1778–1785, 2009.

[8] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in Neural Information Processing Systems (NIPS) 26*, pp.935–943, 2013.

[9] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *Computer Vision (ICCV), IEEE Conference on*, pp.2584–2591, 2013.

[10] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *International Conference on Learning Representations (ICLR)*, 2014.

[11] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing Category-Level Attributes for Discriminative Visual Recognition," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pp.771–778, 2013.

[12] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pp.951–958, 2009.

[13] B. Romera-Paredes and P. H. S. Torr, " An Embarrassingly Simple Approach to Zero-shot Learning," *Journal of Machine Learning Research*, vol.37, 2015.

[14] V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[15] G. Tsoumakas and Katakis, "Multi Label Classification: An Overview," *International Journal of Data Warehousing and Mining*, vol.3, no.3, pp.1–13, 2007.

[16] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learnin*. New York: Springer, 2009.

[17] M. Suzuki, H. Sato, S. Oyama, and M. Kurihara, "Transfer learning based on the observation probability of each attribute," in *Systems, Man and Cybernetics (SMC), IEEE International Conference on*, pp.3627–3631, 2014.

جدول ۴,۶: *مقایسه مهم‌ترین روش‌های ارائه شده برای یادگیری از صفر*

نام روش	سال ارائه	نوع توصیف قابل استفاده	مزایا و معایب
DAP [۱۲]	۲۰۰۹	بردار ویژگی	+ارائه یک چارچوب نظام‌مند + امکان تعویض برخی قسمت‌ها مانند نوع دسته‌بند مورد استفاده ـ مدل نکردن ارتباط میان ویژگی‌ها ـ در نظر گرفتن خطای دسته‌بندی در آموزش
ESZSL [۱۳]	۲۰۱۵	بردار ویژگی	+ درنظرگرفتن خطای دسته‌بندر در آموزش + دارای جواب بسته و پیاده‌سازی یک خطی + سرعت آموزش و آزمون بالا ـ در نظر نگرفتن ارتباط بین ویژگی‌ها ـ محدود بودن رابطه به روابط خطی
COSTA [۲۶]	۲۰۱۴	برچسب‌های دیگر	+عدم نیاز به توصیف کلاس تهیه شده توسط انسان + امکان انجام یادگیری از صفر چند برچسبی ـ تنها امکان استفاده از اطلاع جانبی قابل دسته‌بندی ـ عدم امکان استفاده از ویژگی‌های غیر دودویی
SSE [۲۸]	۲۰۱۵	بردار ویژگی	+ امکان طبیعی استفاده از ویژگی‌ها با مقدار حقیقی + ارائه یک روش عمومی برای بیان دسته‌های آزمون بر حسب دسته‌های آموزش ـ مسئله بهینه‌سازی نسبتا زمان‌بر ـ الزاما یکسان در نظر گرفتن توزیع داده‌های آموزش و آزمون
تشخیص هم‌دسته بودن توصیف و تصویر [۳۱]	۲۰۱۵	انواع مختلف	+ امکان طبیعی استفاده از انواع ویژگی‌ها + پارامترهای مستقل از تعداد دسته‌ها ـ استنتاج سنگین که به اجبار تخمین زده می‌شود
یادگیری از صفر نیمه‌نظارتی با یادگیری نمایش برچسب‌ها [۲۹]	۲۰۱۵	بردار ویژگی یا بدون توصیف	+ یادگیری نمایش برچسب‌ها طوری که متمایزکننده‌ی دسته‌ها شود + دسته‌بندی روی تمام دسته‌های آموزش و آزمون + امکان دسته‌بندی حتی بدون توصیف با یادگیری توصیف‌ها
پیش‌بینی دسته‌بند از متن توصیفی [۲۵]	۲۰۱۵	متن	+ معرفی دسته‌بند پیچشی ـ استخراج ویژگی‌های نه چندان خوب از متن ـ جمع‌آوری متون مناسب ممکن است هزینه‌بر باشد
DeViSE [۲۲]	۲۰۱۴	نام دسته‌ها	+ عدم نیاز به تهیه توصیف توسط انسان + بهره‌گیری از پیش‌آموزش روی داده‌های فراوان ـ عدم دسته‌بندی دقیق برای دسته‌های نزدیک به هم

[18] D. Mahajan, S. Sellamanickam, and V. Nair, "A joint learning framework for attribute models and object descriptions," in *Computer Vision (ICCV), IEEE International Conference on*, pp.1227–1234, 2011.

[19] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *Computer Vision (ECCV), European Conference on*, vol.6315, pp.127–140, 2010.

[20] X. Wang and Q. Ji, "A unified probabilistic approach modeling relationships between attributes and objects," in *Computer Vision (ICCV), IEEE International Conference on*, pp.2120–2127, 2013.

[21] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pp.819–826, 2013.

[22] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A Deep Visual-Semantic Embedding Model," in *Advances in Neural Information Processing Systems (NIPS) 26*, pp.2121–2129, 2013.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NIPS) 26*, pp.3111–3119, 2013.

[24] M. Elhoseiny, A. Elgammal, and B. Saleh, "Tell and Predict: Kernel Classifier Prediction for Unseen Visual Classes from Unstructured Text Descriptions," *arXiv preprint arXiv:1506.08529*, 2015.

[25] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions," *arXiv preprint arXiv:1506.00511*, 2015.

[26] T. Mensink, E. Gavves, and C. Snoek, "Costa: Co-occurrence statistics for zero-shot classification," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pp.2441–2448, 2014.

[27] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *International Conference on Learning Representations*, 2014.

[28] Z. Zhang and V. Saligrama, "Zero-Shot Learning via Semantic Similarity Embedding," in *Computer Vision (ICCV), IEEE Conference on*, 2015.

[29] D. Schuurmans and A. B. Tg, "Semi-Supervised Zero-Shot Classification with Label Representation Learning," in *Computer Vision (ICCV), IEEE Conference on*, 2015.

[30] X. Li and Y. Guo, "Max-margin zero-shot learning for multi-class classification," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp.626–634, 2015.

[31] Z. Zhang and V. Saligrama, "Classifying Unseen Instances by Learning Class-Independent Similarity Functions," *arXiv preprint arXiv:1511.04512*, 2015.

[32] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of Output Embeddings for Fine-Grained Image Classification," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2015.

[33] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised Domain Adaptation for Zero-Shot Learning," in *Computer Vision (ICCV), IEEE Conference on*, pp.2927–2936, 2015.

[34] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Advances in Neural Information Processing Systems (NIPS) 27*, pp.3464–3472, 2014.

[35] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling The Knowledge in a Neural Network," in *NIPS Deep Learning Workshop*, 2014.

[36] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *International Journal of Computer Vision*, vol.108, no.1-2, pp.59–81, 2014.

۷ واژه‌نامه

[۱] Zero-Shot Learning
[۲] Semi-supervised learning
[۳] One-shot learning
[۴] Transfer Learning
[۵] One-Hot Encoding
[۶] Hamming
[۷] Meta-data
[۸] Tag
[۹] Mulit-label Classification
[۱۰] Direct Attribute Prediction
[۱۱] Logistic Regression
[۱۲] Emprical
[۱۳] Indirect Attribute Prediction
[۱۴] Maximum a Posteriori
[۱۵] Topic Modeling
[۱۶] Baysian Network
[۱۷] Structure Learning
[۱۸] Regularization Term
[۱۹] Probabilistic Regression
[۲۰] Convolutional
[۲۱] Feature Map
[۲۲] convolutional Neural Network
[۲۳] Convolotional Filter
[۲۴] Pooling
[۲۵] Cross-Entropy
[۲۶] Simplex
[۲۷] Histogram
[۲۸] Discriminative
[۲۹] prior
[۳۰] Sparse Coding
[۳۱] Random Forest
[۳۲] softmax
[۳۳] Activation
[۳۴] Rectified Linear Unit (ReLU)
[۳۵] Dropout
[۳۶] Cross Validation
[۳۷] Manhattan Distance
[۳۸] Animal With Attributs (AwA)
[۳۹] SUN Attribute
[۴۰] Attribute Pascal/Yahoo
[۴۱] Mulit-class accuracy
[۴۲] Recurrent
[۴۳] http://www.Wikipedia.org