دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد

گرایش هوش مصنوعی

عنوان:

یادگیری صفرضرب با شبکه‌های ژرف

نگارش:

سیدمحسن شجاعی

استاد راهنما:

دکتر مهدیه سلیمانی

تابستان ۱۳۹۵


بسم الله الرحمن الرحیم

چکیده

در برخی از مسائل دسته‌بندی، ممکن است داده‌ی برچسب‌دار برای تمامی دسته‌های موجود در مسئله در دسترس نباشد. برای حل چنین مسائلی، یادگیری صفرضرب از اطلاعات جانبی توصیف کننده‌ی دسته‌ها استفاده می‌کند تا برای آن‌ها دسته‌بند بسازد. به طور خاص در مسئله دسته‌بندی صفرضرب تصاویر، زمانی که دسته‌بندی دسته‌های نوظور یا دسته‌های بسیار شبیه به هم مطرح باشد، جمع‌آوری نمونه برای تمام دسته‌ها امکان‌پذیر نخواهد بود. در این حالت از بردارهای ویژگی یا متون و کلمات توصیف‌کننده‌ی دسته‌ها برای ساختن دسته‌بند برای آنها استفاده می‌شود. در این پژوهش، روش‌هایی ارائه می‌کنیم که علاوه بر این اطلاعات، از اطلاعات بدون نظارت موجود در ساختار فضای تصاویر نیز برای دسته‌بندی تصاویر استفاده کند. با توجه به موفقیت‌های اخیر شبکه‌های عصبی ژرف در زمینه‌ی بینایی ماشین، یک نمایش غنی از تصاویر با استفاده از این شبکه‌ها قابل بدست آوردن است. این نمایش حاوی اطلاعات بدون نظارتی است که قابلیت جداسازی نمونه‌های دسته‌های متفاوت را دارد. در بعضی از روش‌های پیشنهادی از این اطلاعات برای بهبود یادگیری نگاشت از تصاویر به یک فضای میانی، که ممکن است فضای توصیف دسته‌ها یا فضای هیستوگرام‌هایی از دسته‌های دیده شده باشد، با شبکه‌های ژرف بهره می‌بریم. در یک روش پیشنهادی دیگر، با استفاده از این اطلاعات یک نگاشت خطی از فضای توصیف‌ها به فضای تصاویر پیدا می‌کنیم، به گونه‌ای که هر توصیف مربوط به دسته‌های آموزش به مرکز نمونه‌های دسته‌ی مربوط به خود نگاشته شود و توصیف مربوط به دسته‌های آزمون به نزدیکی خوشه‌ای از نمونه‌های آزمون. نشان داده خواهد شد که این روش، می‌تواند مشکل جابجایی دامنه که باعث تضعیف عملکرد روش‌های یادگیری صفرضرب می‌شود را کاهش دهد. کارایی روش پیشنهادی با آزمایشات عملی بر روی چهار مجموعه دادگان مرسوم برای مسئله یادگیری صفرضرب سنجیده می‌شود که در سه مورد از این چهار مجموعه، به دقت دسته‌بندی بالاتری نسبت به روش‌های پیشگام دست می‌یابد.

کلیدواژه‌ها: یادگیری صفرضرب، انتقال یادگیری، یادگیری نیمه‌نظارتی، شبکه‌های ژرف

فهرست مطالب

[۱] Bi-Linear

[۲] Max Margin

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱

مقدمه

در حوزه یادگیری ماشین، مسئله‌ی استاندارد یادگیری با نظارت، به صورت‌های مختلف توسعه یافته است و باعث ایجاد روش‌هایی با تعاریف و فرض‌های گوناگون شده است. به کمک این روش‌ها، یادگیری ماشین از عهده‌ی حل مسائل چالش‌برانگیزتری برآمده است. بر خلاف الگوی 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*[۱] [۲] برای استفاده کردن از حجم زیاد داده‌های بدون برچسب موجود در جریان آموزش پیشنهاد شده است. *یادگیری تک‌ضرب*[۲] [۳] سعی می‌کند بعضی دسته‌ها را تنها بوسیله یک نمونه‌ی برچسب‌دار از آن دسته و البته با کمک نمونه‌های برچسب‌دار از سایر دسته‌ها شناسایی کند. *انتقال یادگیری*[۳] [۴] سعی می‌کند دانش به دست آمده از داده‌های یک دامنه (یا دانش یادگرفته شده برای انجام یک وظیفه) را به داده‌های دامنه‌ی دیگر (یا انجام وظیفه‌ی دیگری روی داده‌ها) منتقل کند. هیچ‌کدام از این روش‌ها نیاز به داده‌های برچسب‌دار را برای دسته‌هایی که مایل به تشخیص آن هستیم، به طور کامل از بین نمی‌برد. برای دست‌یابی به چنین هدفی، مسئله *یادگیری صفرضرب* صورت‌بندی شده است [۵]. در این مسئله برای برخی از دسته‌هایی که به دنبال یافتن یک دسته‌بند برای آن‌ها هستیم، هیچ نمونه‌ای در زمان آموزش موجود نیست؛ در عوض فرض می‌شود که یک *توصیف* یا *امضا* از تمامی دسته‌ها 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

[۱] Semi-supervised Learning

[۲] One-shot Learning

[۳] Transfer Learning

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 [۶] نشان داده‌شده، تعداد نمونه‌های موجود برای دسته‌ها 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 [۷]؛ 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 تعریف مسئله یادگیری صفرضرب بر آن منطبق است دسته‌بندی در حالت وجود دسته‌های نوظهور است، مانند تشخیص مدل‌های جدید محصولاتی چون خودروها که بعضی دسته‌ها در زمان آموزش اصولا وجود نداشته است. یادگیری صفرضرب 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 در تصویر 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 یادگیری صفرضرب در دسته‌بندی تصاویر تمرکز می‌کنیم. 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 هر یک از دسته‌ها با نوعی اطلاعات جانبی توصیف می‌شوند؛ 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 دسته‌ها نیز در اختیار قرار می‌گیرد. در برخی روش‌ها فرض می‌شود که توصیف دسته‌های آزمون نیز در زمان آموزش قابل دسترسی است. توصیف‌ها ممکن است به صورت یک بردار از صفت‌ها[۴] [۸]، عبارات زبان طبیعی [۹، ۱۰، ۱۱] و یا یک دسته‌بند برای آن دسته [۱۲] باشند. بردار صفت مرسوم‌ترین شکل توصیف دسته است. صفت‌ها با توجه به نوع مسئله و دسته‌های موجود تعیین می‌شوند. اکثر صفت‌ها، صفت‌های بصری هستند که برای نمونه جهت توصیف شکل (مانند گرد یا مستطیلی)، جنس (مانند چوبی یا فلزی) و عناصر موجود در تصویر (مانند چشم، مو، پدال و نوشته) به کار می‌روند. برخی صفت‌ها هم ممکن است مستقیما در تصویر قابل مشاهده نباشند برای مثال در یک مجموعه دادگان که دسته‌ها انواع حیوانات هستند [۱۳]، علاوه بر صفت‌های بصری، صفت‌هایی چون اهلی بودن، سریع بودن یا گوشت‌خوار بودن هم وجود دارد.

[۴] Attribute

اکثر روش‌های بکار گرفته شده در یادگیری صفرضرب، با یادگیری نگاشتی از تصاویر و توصیف‌ها به یک فضای مشترک و سپس استفاده از یک معیار مانند ضرب داخلی برای سنجش شباهت تصاویر و توصیف‌ها به یکدیگر عمل می‌کنند. در نهایت برچسب تعلق گرفته به هر نمونه، برچسبی است که توصیف آن بیشترین شباهت را به تصویر داراست. در کارهای پیشین توجه اندکی به ساختار فضای تصاویر و نحوه‌ی قرارگیری نمونه‌ها در آن شده است. از طرفی پیشرفت‌های اخیر در زمینه بینایی ماشین با استفاده از شبکه‌های ژرف [۱۴] این امکان را فراهم کرده که نمایشی با قابلیت تمایز بسیار از تصاویر بدست آید و دسته‌های بصری مختلف در فضای این ویژگی‌ها به نحو مناسبی از یکدیگر جدا باشند. همان‌طور که در بخش ۴-۸ نشان داده خواهد شد، در این فضای ویژگی نمونه‌های دسته‌های مختلف تشکیل خوشه‌های جدا از هم می‌دهند و در نتیجه ساختار این فضا می‌تواند حاوی اطلاعات مفیدی برای دسته‌بندی تصاویر باشد. در روش‌های پیشنهادی سعی می‌کنیم چارچوبی برای استفاده از این اطلاعات بدون نظارت که صرفا از تصاویر استخراج می‌شوند در مسئله یادگیری صفرضرب ارائه کنیم.

ساختار ادامه‌ی این نوشتار به این صورت است: فصل ۲ به مرور روش‌های پیشین اختصاص دارد که در آن ابتدا یک چارچوب کلی برای روش‌های یادگیری صفرضرب معرفی می‌شوند و سپس روش‌ها با توجه به چارچوب ارائه شده دسته‌بندی و مرور می‌شوند. فصل ۳ به بیان روش‌های پیشنهادی اختصاص دارد که در آن ابتدا یک شبکه عصبی ژرف چندوظیفه‌ای برای یادگیری نیمه‌نظارتی در پیش‌بینی توصیف از تصویر پیشنهاد می‌شود. این شبکه از دقت دسته‌بندی صفرضرب بالاتری نسبت به سایر روش‌های پیش‌بینی صفت برخوردار است. هم‌چنین در این فصل یک تابع مطابقت میان توصیف‌ها و تصاویر پیشنهاد می‌شود و سپس یک روش ساده برای استفاده از این تابع مطابقت با استفاده از خوشه‌بندی تصاویر ارائه می‌شود. سپس برای رفع نقص‌های این روش، روشی برای خوشه‌بندی و یادگیری نگاشت به فضای مشترک به صورت توام پیشنهاد می‌شود. در فصل ۴ نتایج آزمایشات عملی برای سنجش روش‌های پیشنهادی به همراه تحلیلی برای عمل‌کرد آن‌ها ارائه می‌شود و در نهایت در بخش ۵ به جمع‌بندی و راه‌کارهای آتی پرداخته خواهد شد.

فصل ۲

روش‌های پیشین

در این فصل ابتدا یک چارچوب کلی برای روش‌های مورد استفاده در یادگیری صفرضرب توصیف می‌شود. سپس روش‌های موجود طبق این چارچوب دسته‌بندی شده و مرور خواهند شد.

از نظر تاریخی، پیش از تعریف و بیان رسمی مسئله یادگیری صفرضرب، استفاده از اشتراک و تمایز برخی صفت‌ها میان دسته‌های مختلف در بینایی ماشین مورد بررسی قرار گرفته است [۱۵، ۱۶، ۱۷]، اما این روش‌ها به شناسایی دسته‌های کاملا جدید از روی این صفت‌ها توجه نشان نداده‌اند. مسئله‌ی یادگیری تک‌ضرب هم یک مسئله نزدیک به یادگیری صفرضرب است که پیش‌تر مورد بررسی بوده است [۳]. در حقیقت می‌توان یادگیری تک‌ضرب را حالت خاصی از یادگیری صفرضرب در نظر گرفت که در آن توصیف دسته‌های دیده نشده به صورت یک نمونه از آن دسته ارائه شده است [۵]. پدیده شروع سرد[۱] در سامانه‌های توصیه‌گر[۲] را نیز می‌توان از حالت‌های خاص یادگیری صفرضرب در نظر گرفت که در آن برای یک کاربر یا مورد جدید پیشنهاد صورت می‌گیرد.

بیان مسئله یادگیری صفرضرب به طور رسمی برای اولین بار در [۵] صورت گرفت. در آن‌جا دو دیدگاه کلی برای حل مسئله یادگیری صفرضرب بیان می‌شود. یک روش که دیدگاه فضای ورودی[۳] نامیده می‌شود، سعی در مدل کردن نگاشتی با دو ورودی دارد. یک ورودی نمونه‌ها و دیگری توصیف دسته‌ها است و امتیازی مبنی بر مطابقت آن‌ها با یکدیگر تولید می‌کند، یعنی برای نمونه‌ها و توصیف‌های مربوط به یک دسته امتیاز بالا و برای نمونه‌ها و توصیفاتی که متعلق به دسته‌ی

[۱] Cold Start

[۲] Recommender System

[۳] input space view

یکسانی نیستند مقادیر کوچکی تولید می‌کند. با تخمین زدن چنین نگاشتی روی داده‌های آموزش، دسته‌بندی نمونه‌های آزمون در دسته‌هایی که تا کنون نمونه‌ای نداشته‌اند ممکن خواهد شد. به این صورت که هر نمونه با توصیف دسته‌های مختلف به این تابع داده شده و متعلق به دسته‌ای که امتیاز بیشتری بگیرد، پیش‌بینی خواهد شد. در روش دیگر که دیدگاه فضای مدل[۴] نام دارد، مدل مربوط به هر دسته (برای مثال پارامترهای دسته‌بند مربوط به آن)، به عنوان تابعی از توصیف آن دسته در نظر گرفته می‌شود.

ما در این فصل از دسته‌بندی دیگری برای مرور روش‌های پیشین استفاده می‌کنیم. برای این کار ابتدا تعریف دقیق مسئله با استفاده از نمادگذاری تعریف شده صورت می‌گیرد. پس از آن معرفی یک چارچوب کلی برای انجام یادگیری صفرضرب لازم است که دو دیدگاه فوق نیز در این چارچوب قابل بیان هستند.

۲-۱ نمادگذاری

برای این که تعریف مسئله و توصیف روش‌های پیشین به صورت دقیق ممکن باشد، در ابتدای یک نمادگذاری برای مسئله ارائه می‌دهیم و از آن برای بیان مرور روش‌های پیشین و بیان روش پیشنهادی در فصل آینده استفاده خواهیم کرد.

برای ماتریس X، $X_{(i)}$ سطر $i-$م آن و $\|X\|_{Fro}$ نرم فروبنیوس آن را نشان می‌دهد. هم‌چنین برای بردار $\mathbf{x}$، x_i درایه‌ی $i-$م را نشان می‌دهد. ضرب داخلی با نماد $\langle .,. \rangle$ نشان داده شده است. $diag(\mathbf{x})$ یک ماتریس قطری را نشان می‌دهد که بردار $\mathbf{x}$ روی قطر اصلی آن قرار داده شده است. $\mathbf{1}$ یک بردار تمام یک و $\mathbf{1}_k$ یک بردار که عنصر $k-$م آن یک و سایر عناصر آن صفر است را نشان می‌دهند.

تصاویر را با $\mathbf{x} \in \mathbb{R}^d$ نشان می‌دهیم که d ابعاد داده را نشان می‌دهد. توصیف‌ها را با $\mathbf{c} \in \mathbb{R}^a$ نمایش می‌دهیم که 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 $\mathcal{Y} = \mathcal{U} \cup \mathcal{S}$. تعداد دسته‌های آموزش را با n_s و تعداد دسته‌های آزمون را با n_u نشان می‌دهیم. هم‌چنین $\mathbf{c_y}$ که در آن $y \in \mathcal{U} \cup \mathcal{S}$ بردار توصیف دسته y را نشان می‌دهد.

فرض می‌کنیم در زمان آموزش $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_s}$ شامل N_s تصویر از دسته‌های دیده شده به همراه برچسب موجود است. $X_s \in \mathbb{R}^{d \times N_s}$ ماتریس مجموعه تصاویر و Y_s ماتریس برچسب‌های داده‌های آموزش با کدگذاری یکی‌یک [۵]

[۴] model space view

[۵] One-Hot Encoding

است. هم‌چنین توصیف‌های هر 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. $X = [X_s; X_u]$ ماتریس ویژگی تمام نمونه‌ها، اعم از آموزش و آزمون است.

۲-۲ تعریف مسئله

در مسئله دسته‌بندی تصاویر به صورت صفرضرب، فرض می‌شود N_s تصویر آموزش به همراه برچسب‌هایشان، یعنی $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_s}$ موجود است. این تصاویر متعلق به دسته‌های موجود در $\mathcal{S}$ هستند، به عبارت دقیق‌تر

$$(\mathbf{y}_i)_j = 0 \quad \forall n_s < j, \tag{۱-۲}$$

هدف در مسئله پیش‌بینی برچسب‌های $\{(\mathbf{y^*}_i)\}_{i=N_s+1}^{N_s+N_u}$ برای نمونه‌های آزمون $\{(\mathbf{x}_i)\}_{i=N_s+1}^{N_s+N_u}$ است. به صورتی که تفاوت $\{(\mathbf{y_i^*})\}_{i=N_s+1}^{N_s+N_u}$ با برچسب‌های صحیح کمینه شود. به عبارت دیگر هدف مسئله کمینه کردن تابع زیر است:

$$\min_{\mathbf{y}^*} \sum_{i=N_s+1}^{N_s+N_u} \mathbb{1}(\mathbf{y_i^*} \neq \mathbf{y_i}). \tag{۲-۲}$$

در اکثر مواقع فرض ساده‌کننده‌ی جدا بودن دسته‌های آزمون و آموزش نیز در مسئله وجود دارد به این معنا که:

$$(\mathbf{y_i})_j = 0 \quad \forall j \leqslant n_s. \tag{۳-۲}$$

برای این که حل چنین مسئله‌ای امکان‌پذیر باشد، دسته‌های دیده نشده باید به وسیله‌ای مشخص و از یکدیگر متمایز شوند. در مسئله یادگیری صفرضرب، برای این هدف از توصیف‌های C_s و C_u استفاده می‌شود. به همین علت از بردار توصیف هر دسته با عنوان *امضای*[۶] آن دسته نیز یاد می‌شود.

اشاره این نکته نیز می‌تواند مفید باشد که تعریف مسئله یادگیری تک‌ضرب کاملا مشابه تعریف ارائه شده در بالاست و تنها با نوع توصیف مورد استفاده از مسئله یادگیری صفرضرب متمایز می‌شود. در مسئله یادگیری تک‌ضرب امضای هر دسته دیده نشده یک (یا تعداد اندکی) نمونه از آن دسته هستند و امضای یک دسته‌ی دیده شده تمام نمونه‌های موجود از آن. به این علت همان‌طور که در ابتدای فصل عنوان شد می‌توان مسئله یادگیری تک‌ضرب را که مسئله‌ای قدیمی‌تر از یادگیری صفرضرب است، در حقیقت یک حالت خاص از یادگیری صفرضرب دانست.

می‌توان گفت که هر روش برای یادگیری صفرضرب از سه قسمت تشکیل شده است که ممکن است به صورت مستقل یا همزمان انجام شوند؛ این سه قسمت عبارتند از:

[۶] Signature

۱. یادگرفتن نگاشتی از فضای تصاویر به فضای مشترک که آن را با $\phi : \mathbb{R}^d \to \mathcal{M}$ نشان می‌دهیم.

۲. نگاشت توصیف دسته‌ها به فضای مشترک که آن را با $\theta : \mathbb{R}^a \to \mathcal{M}$ نشان می‌دهیم.

۳. ارائه روشی برای تعیین مشابهت در این فضای مشترک و اختصاص برچسب به تصاویر. (برای مثال یک ضرب داخلی یا عکس فاصله در فضای $\mathcal{M}$).

چارچوبی که در ادامه می‌آید بر این اساس استوار است که تصاویر و توصیفات آن‌ها به یک فضای مشترک نگاشته می‌شوند. اگر بخواهیم دسته‌بندی ارائه شده در [۵] را که در ابتدای فصل بیان شد در این چارچوب توصیف کنیم، در دیدگاه فضای ورودی، فضای مشترک فضایی است که نگاشت شباهت سنجی، ضرب داخلی آن فضاست و در دیدگاه فضای مدل، فضای مشترک فضای دسته‌بندها خواهد بود.

۲-۳ 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 از 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 θ مستقل از ϕ انجام شده و رابطه بین توصیف‌ها و برچسب دسته‌ها رابطه‌ای یک به یک است. با این دو فرض می‌توان $\theta(\mathbf{c_y})$ را *امضای دسته‌ی* y نامید.

در [۷] با فرض دودویی بودن هر بعد از امضای دسته‌ها، کرانی بر اساس فاصله همینگ [۷] میان امضای دسته‌ی صحیح و مقدار پیش‌بینی شده ارائه می‌شود. در [۱۸] از نتایج مشابه در حوزه تطبیق دامنه برای کران‌دار کردن خطا استفاده 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 [۱۹] 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.

[۷] Hamming

۲-۴ پیش‌بینی صفت

این دسته از روش‌ها عموما به حالتی از مسئله یادگیری صفرضرب تعلق دارند که توصیف دسته‌ها از نوع بردار صفت باشد. در این حالت فضای مشترک همان فضای صفت‌ها در نظر گرفته می‌شود. به عبارت دیگر نگاشت θ نگاشت همانی فرض شده و یادگرفته نخواهد شد. روش‌های اولیه ارائه شده برای یادگیری صفرضرب از نوع پیش‌بینی صفت[۸] بوده‌اند و پس از آن هم قسمت قابل توجهی از روش‌ها در این دسته جای می‌گیرند که در ادامه آن‌ها را به تفصیل مرور می‌کنیم.

۲-۴-۱ پیش‌بینی صفت مستقیم و غیر مستقیم

در [۷] از چند رگرسیون لجستیک[۹] مستقل برای پیش‌بینی‌های صفت‌های دودویی از تصاویر fMRI استفاده شده و سپس دسته‌بندی با دسته‌بند نزدیک‌ترین همسایه بر اساس نزدیکی بردار صفت پیش‌بینی شده و امضای دسته‌های آزمون صورت می‌پذیرد.

در [۱۳] با فرض این که صفت‌ها به صورت مستقل از یکدیگر قابل پیش‌بینی هستند دو دیدگاه برای این کار ارائه می‌کند. پیش‌بینی صفت مستقیم[۱۰] (DAP) و پیش‌بینی صفت غیرمستقیم[۱۱] (IAP). مدل گرافی مورد استفاده در این دو دیدگاه در تصویر ۲-۱ آمده است. در پیش‌بینی صفت مستقیم برچسب‌ها به شرط دانستن صفت‌های درون تصویر، از تصویر مستقل هستند. در این روش برای هر یک صفت‌ها یک دسته‌بند یاد گرفته می‌شود. با توجه به این که صفت‌ها برای تصاویر آزمون معین هستند این کار با استفاده از یک دسته‌بند احتمالی برای هر صفت قابل انجام است. در نهایت احتمال تعلق هر یک از برچسب‌های $u \in \mathcal{U}$ با استفاده از رابطه زیر بدست خواهد آمد

$$P(u|\mathbf{x}) = \sum_{\mathbf{c}\in\{۰,۱\}^a} P(u|\mathbf{x})p(\mathbf{c}|\mathbf{x}). \qquad (۴-۲)$$

با توجه به فرض استقلال صفت داریم $P(\mathbf{c}|\mathbf{x}) = \prod_{n=۱}^{a} P(\mathbf{c}_m|\mathbf{x})$. برای محاسبه جمله $P(u|\mathbf{c})$ از قانون بیز استفاده می‌کنیم:

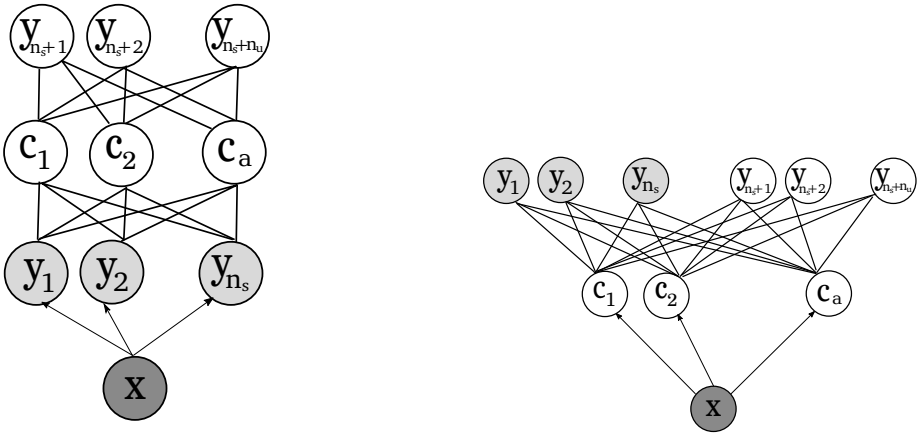
$$P(\mathbf{u}|\mathbf{c}) = \frac{P(u)P(\mathbf{c}|u)}{P(\mathbf{c_u})} = \frac{P(u)\mathbb{1}(c = \mathbf{c_u})}{P(\mathbf{c_u})},$$

[۸] Attribute Prediction

[۹] Logistic Regression

[۱۰] Direct Attribute Prediction

[۱۱] Indirect Attribute Prediction



(ب) (آ)

شکل ۲-۱: مدل گرافی پیش‌بینی ویژگی مستقیم (آ) و غیر مستقیم (ب). رئوس با سایه‌ی روشن رئوسی هستند که در زمان آموزش رویت شده هستند و رئوس با سایه‌ی تیره همواره رویت شده‌اند. رئوس بدون سایه مربوط به متغیرهایی است که باید استنتاج در مورد آن‌ها انجام شود. یال‌های ضخیم‌تر روابط ثابت را نشان می‌دهند که جزو داده‌های آموزش هستند و یال‌های نازک‌تر روابطی را که باید کشف شوند. x یک تصویر است، متغیرهای دودویی $y_1, \dots y_{n_s}$ تعلق یا عدم تعلق تصویر به دسته‌های دیده شده و بصورت مشابه $y_{n_s+1}, \dots y_{n_s+n_u}$ تعلق یا عدم تعلق به دسته‌های دیده نشده را نشان می‌دهند. $c_1, \dots, c_a$ ویژگی‌های توصیف‌کننده دسته‌ها هستند. آ) در مدل پیش‌بینی ویژگی مستقیم رابطه میان برچسب‌ها و ویژگی‌ها ثابت فرض می‌شود و هدف استنتاج ویژگی از روی تصاویر است. بعد از آن با استفاده از رابطه از پیش تعیین شده برچسب‌ها با ویژگی‌ها، برچسب تعیین می‌شود. ب) در مدل پیش‌بینی ویژگی غیر مستقیم، یک دسته‌بند چنددسته‌ای روی دسته‌های آموزش یادگرفته می‌شود و با توجه به وقوع یا عدم وقوع هر یک از ویژگی‌ها در این دسته‌ها رابطه‌ی ثابتی میان دسته‌های دیده شده $y_1, \dots y_{n_s}$ و ویژگی‌ها فرض می‌شود. هم‌چنین رابطه ویژگی‌ها با دسته‌های دیده نشده $y_{n_s+1}, \dots y_{n_s+n_u}$ رابطه امضا بودن است و دانسته فرض می‌شود [۱۳].

و با جایگذاری آن در رابطه (۲-۴) خواهیم داشت:

$$P(u|\mathbf{x}) = \frac{P(u)}{P(\mathbf{c_u})} \prod_{n=1}^{a} P(\mathbf{c_u}_n|\mathbf{x}). \qquad (۲-۵)$$

در نهایت برچسبی که احتمال فوق را بیشینه کند، پیش‌بینی مربوط به تصویر x خواهد بود.

در روش پیش‌بینی صفت غیر مستقیم، IAP تخمین $P(c_i|\mathbf{x})$ تغییر داده می‌شود؛ به این صورت که ابتدا یک دسته‌بند چند دسته‌ای یعنی $P(y_k|\mathbf{x})$ روی داده‌ها یاد گرفته می‌شود و سپس رابطه صفت‌ها و برچسب‌ها به صورت قطعی مدل

می‌شود:

$$P(\mathbf{c}_i|\mathbf{x}) = \sum_{k=1}^{n_u} P(y_k|\mathbf{x})\mathbb{I}(\mathbf{c}_i = \mathbf{c}_{\mathbf{y_k}i}). \quad (۶-۲)$$

در نهایت در هر دو روش برچسب نهایی با تخمین MAP[۱۲] از رابطه زیر تعیین می‌شود:

$$\hat{y} = \arg\max_{u\in\mathcal{U}} P(u|\mathbf{x}) = \arg\max_{u\in\mathcal{U}} \prod_{i=1}^{a} \frac{P(\mathbf{c}_{\mathbf{u}i}|\mathbf{x})}{P(\mathbf{c}_{\mathbf{u}i})} \quad (۷-۲)$$

روش ارائه شده در [۲۰] مشابه همین روش است با این تفاوت که احتمال مشاهده هر کدام صفت‌ها را هم در محاسبه دخیل می‌کند تا با وزن‌های متفاوت با توجه به اهمیتشان در دسته‌بندی نقش داشته باشند. ضعف بزرگ این روش‌ها فرض مستقل بودن صفت‌ها از یکدیگر است؛ چرا که این فرض در مسائل واقعی معمولا برقرار نیست. برای مثال زمانی که صفت آبزی بودن برای یک موجود در نظر گرفته می‌شود احتمال صفت پرواز کردن برای آن بسیار کاهش می‌یابد.

۲-۴-۲ مدل‌سازی احتمالی روابط بین صفت‌ها

تا کنون تعدادی مدل گرافی برای در نظر گرفتن وابستگی‌های میان صفت‌ها معرفی شده‌است. نویسندگان [۲۱] برای در نظر گرفتن ارتباط بین خود صفت‌ها و ارتباط صفت‌ها با برچسب نهایی روش‌های مدل‌سازی موضوع[۱۳] را از حوزه یادگیری در متن اقتباس کرده‌اند. همچنین نویسندگان [۲۲] برای این کار یک چارچوب بر اساس مدل‌های گرافی احتمالی معرفی می‌کنند. در این چارچوب شبکه بیزی[۱۴] برای مدل کردن این روابط در نظر گرفته می‌شود و ساختار آن که نشان‌دهنده وابستگی یا استقلال صفت‌ها با هم یا با برچسب است، با کمک روش‌های یادگیری ساختار[۱۵] شناخته می‌شود.

۲-۵ نگاشت به فضای توصیف‌ها

در برخی موارد توصیف‌های داده شده از جنسی غیر از صفت هستند ولی فضای مشترک همان فضای توصیف‌ها در نظر گرفته می‌شود و سعی می‌شود تصاویر به این فضا نگاشته شوند. روش ConSE[۱۶] [۱۱] از چنین نگاشتی استفاده می‌کند.

[۱۲] Maximum a Posteriori
[۱۳] Topic Modeling
[۱۴] Baysian Network
[۱۵] Structure Learning
[۱۶] Convex combination of Semantic Embeddings

ابتدا یک شبکه عصبی پیچشی[۱۷] برای دسته‌بندی نمونه‌های دسته‌های دیده‌شده آموزش داده می‌شود. این مسئله،یک مسئله یادگیری دسته‌بند عادی است و شبکه‌ها در اکثر موارد از قبل به صورت پیش‌آموزش دیده شده وجود دارند. تابع فعال‌سازی[۱۸] لایه‌ی آخر این شبکه به این صورت تعریف می‌شود:

$$softmax(z)_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \quad j = ۱, \ldots, n_s. \tag{۸-۲}$$

تابع بالا به ازای هر j، امتیاز تعلق نمونه به دسته‌ی $j-$م را نشان می‌دهد. در هنگامی که با مسئله دسته‌بندی عادی روبرو هستیم، روی j بیشینه گرفته می‌شود و دسته‌ای که بیشترین امتیاز را گرفته به عنوان پیش‌بینی خروجی داده می‌شود. در روش ConSE برای مسئله یادگیری صفرضرب، هنگامی که یک نمونه از دسته‌های آزمون را به شبکه می‌دهیم، خروجی بدست آمده از رابطه (۲-۸) می‌تواند به عنوان میزان شباهت آن نمونه به هر یک دسته‌های آموزش در نظر گرفته شود. فرض کنید که برای هر نمونه $\hat{y}(x,n)$، $n-$مین عنصر بزرگ $softmax(x)$ را نشان دهد، یعنی $n-$مین برچسب محتمل برای x از میان دسته‌های آموزش. حالا برای پیش‌بینی برچسب x از میان دسته‌های آموزش از این رابطه استفاده می‌کنیم:

$$\phi(x) = \frac{۱}{Z} \sum_{n=۱}^{T} P(\hat{y}(x,n)|x) c_{\hat{y}(x,n)}, \tag{۹-۲}$$

که T یک پارامتر[۱۹] مدل و $Z = \sum_{n=۱}^{T} P(\hat{y}(x,n)|x)$ ضریب نرمال‌سازی است. در این حالت نمونه‌ی x با تابع $\phi(\cdot)$ به فضای توصیف‌ها نگاشته شده است. به عبارت دقیق‌تر به صورت جمع وزن‌دار توصیف T دسته‌ی شبیه‌تر نمایش داده شده است که وزن‌های این جمع میزان شباهت هستند. روش COSTA[۲۰] [۲۳] نیز از دیدگاه مشابهی استفاده می‌کند. در این روش همانند رابطه (۲-۹)، پارامترهای دسته‌بند برای دسته‌های دیده نشده به صورت جمع وزن‌دار پارامترهای دسته‌بندهای دسته‌های دیده شده بیان می‌گردد. در این پژوهش برای بدست آوردن وزن‌های مربوط به شباهت میان دسته‌ها توابع مختلفی از تعداد رخ‌داد همزمان برچسب‌ها پیشنهاد شده است.

۲-۶ نگاشت‌های دوخطی

حالت دیگری از چارچوب کلی معرفی شده در ابتدای فصل این است که نگاشت به فضای مشترک یک نگاشت دوخطی باشد. یعنی به این صورت که W نگاشتی خطی است که $x^T W$ تصویر x را به فضای توصیف‌ها نگاشته و Wc توصیف

[۱۷] Convolutional
[۱۸] Activation Function
[۱۹] Parameter
[۲۰] Co-Occurrence Statistics

c را به فضای تصاویر می‌نگارد.در نهایت تابع مطابقت میان یک توصیف و تصویر به صورت زیر تعریف می‌شود:

$$F(x,c) = \phi(x)^T W \theta(y) \tag{۱۰-۲}$$

در این حالت، این که فضای مشترک در حقیقت کدام یک از فضاهای تصاویر یا توصیفات هستند، جواب روشنی ندارد. نقطه‌ی قوت این روش‌ها در امکان پیچیده‌تر کردن تابع هزینه است. چرا که در حالتی که نگاشت خطی است مسائل بهینه‌سازی پیچیده‌تری نسبت به حالت غیرخطی قابل حل خواهند بود.

۲-۶-۱ یادگیری با تابع هزینه بیشینه حاشیه

یک انتخاب متداول برای تابع هزینه، توابع رتبه‌بند[۲۱] هستند. با توجه به این که عموما بعد از یادگیری این نگاشت، دسته‌ای که نزدیک‌ترین توصیف را (با معیاری مثل فاصله یا ضرب داخلی) دارد، به عنوان پیش‌بینی تولید می‌شود، چنین تابع هزینه‌ای یک انتخاب طبیعی است. چرا که مسئله‌ی نزدیکترین همسایه در اصل یک مسئله رتبه‌بندی است و استفاده از یک تابع هزینه‌ی رتبه‌بند برای یادگیری نگاشت بهتر از مجموع مربعات است [۲۴].

در [۲۵] تابع هزینه رتبه‌بند WSABIE [۲۶] که برای حاشیه‌نویسی تصاویر پیشنهاد شده، به مسئله یادگیری صفرضرب انطباق داده شده‌است. تابع هزینه WSABIE به این صورت تعریف شده است:

$$L(x_s, Y_s; W, \theta) = \frac{1}{N_s} \sum_{n=1}^{N_s} \lambda_{r_\Delta(x_n, y_n)} \sum_{y \in \mathcal{Y}} \max(0, l(x_n, y_n, y)), \tag{۱۱-۲}$$

$$l(x_n, y_n, y) = \mathbb{1}(y \neq y_n) + \phi(x_n)^T W \theta(y) - \phi(x_n)^T W \theta(y_n), \tag{۱۲-۲}$$

که در آن $r_\Delta(x_n, y_n) = \sum_{y \in \mathcal{Y}} \mathbb{I}(l(x_n, y_n, y) > 0)$ و λ_k یک تابع نزولی از k است. این تابع، پیش‌بینی اشتباه صفت‌ها 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. در انطباق برای یادگیری صفرضرب، بهینه‌سازی تنها روی نگاشت W انجام شده و تابع θ دانسته فرض می‌شود: $\theta(y) = c_y$.

ایده‌ی بالا در [۲۷] ادامه داده شده و نگاشت شباهت ساخت‌یافته SJE[۲۲] نامیده شده است. در این حالت تابع

[۲۱] Ranking Function

[۲۲] Structured Joint Embedding

مطابقت بین توصیف‌ها و تصاویر از رابطه (۲-۱۰) تعریف می‌شود. تابع هزینه ساده‌تر از حالت قبل به صورت

$$\frac{1}{N_s}\sum_{n=1}^{N_s}\max_{y\in\mathcal{Y}}(0, l(x_n, y_n, y)), \tag{۲-۱۳}$$

در نظر گرفته شده که l همانند رابطه (۲-۱۲) است. هم‌چنین برای استفاده از چند توصیف به صورت هم‌زمان، تعریف تابع مطابقت به صورت زیر تعمیم داده می‌شود:

$$F(x, y; \{W\}_{1\ldots K}) = \sum_k \alpha_k \theta(x)^T W_k \phi_k(y), \tag{۲-۱۴}$$

$$s.t. \sum_k \alpha_k = 1,$$

که $\phi_k(y)$ توصیف‌های مختلف از دسته‌ی y را نشان می‌دهد و $W_1, \ldots W_K$ نگاشت‌های میان هر یک از این توصیف‌ها و فضای تصاویر را. وزن‌های α_k که میزان اهمیت یا اطمینان هر یک از توصیف‌ها را نشان می‌دهد، با اعتبارسنجی تعیین می‌شوند. روش SJE با انواع اطلاعات جانبی سازگار است. اطلاعات جانبی که آزمایشات با آن‌ها انجام شده است شامل بردار صفت‌های دودویی یا پیوسته تعیین شده توسط انسان و نمایش برداری متون دائره‌المعارفی با روش‌های word2vec [۲۸] و GloVe [۲۹] است. هم‌چنین نویسندگان این پژوهش یک نسخه با نظارت از word2vec ارائه می‌دهند که در جریان آموزش آن از موضوع هر متن هم استفاده می‌شود.

روش SJE در [۳۰] برای برخی نگاشت‌های غیرخطی نیز تعمیم داده شده است. در این روش که LatEm[۲۳] نام دارد تابع هزینه مانند حالت قبل (رابطه (۲-۱۳)) تعریف شده است با این تفاوت که تابع مطابقت میان توصیف و تصویر به‌جای رابطه دوخطی (۲-۱۰) از این رابطه تبعیت می‌کند:

$$F(x, y) = \max_{1\leqslant i\leqslant L} \phi(x)^T W_i \theta(y). \tag{۲-۱۵}$$

در این حالت تابع مطابقت به صورت ترکیب نگاشت‌های دوخطی $W_1, \ldots W_M$ بیان شده است و یک تابع غیرخطی ولی تکه‌تکه خطی[۲۴] برای تصمیم‌گیری مورد استفاده قرار می‌گیرد.

یک تعمیم دیگر از SJE در [۳۱] ارائه شده است که در آن فرض وجود اطلاعات نظارتی قوی‌تر در نظر گرفته شده است. در این حالت فرض می‌شود که در تصاویر قسمت‌های مختلفی که توصیفی از آن‌ها موجود است، مشخص شده‌اند. البته تناظر میان قسمت‌های توصیف و تصویر موجود نیست، مثلا در مجموعه دادگان مربوط به پرنده‌ها، قسمت‌های

[۲۳] Latent Embedding Model

[۲۴] Piece-wise Linear

مختلف بدن پرنده مانند نوک و پا در همه تصاویر جدا شده است اما این اطلاعات که هر کدام از این‌ها به چه قسمتی از توصیف آن دسته مربوط می‌شوند، در دسترس نیست. با این فرض تابع مطابقت F تعریف شده در رابطه (۲-۱۰) به گونه‌ای تعمیم داده می‌شود که مطابقت قسمت‌های مختلف متن و تصویر را بسنجد:

$$F(x,y) = \frac{1}{|g_x||g_y|} \sum_{i \in g_x} \sum_{j \in g_y} \max(0, v_i^T s_j), \tag{۲-۱۶}$$

که در آن g_x مجموعه قسمت‌های مختلف تصویر x و g_y مجموعه قسمت‌های توصیف ارائه شده‌ی دسته‌ی y است. s_j و v_i که به ترتیب بازنمایی یک قسمت از متن و تصویر هستند به صورت زیر تعریف می‌شوند:

$$s_j = f\left(\sum_m W_m^{\text{language}} l_m + b^{\text{language}}\right)$$
$$v_i = W^{\text{visual}}[CNN_\zeta(I_v)] + b^{\text{visual}}. \tag{۲-۱۷}$$

نماد l_m انواع مختلف توصیف را نشان می‌دهند که در این پژوهش شامل بردار صفت، نمایش word2vec و کیسه‌ی کلمات[۲۵] متون توصیف کننده است. W_m^{language} ماتریس‌هایی هستند که هر کدام از m توصیف زبانی را به فضای مشترک می‌نگارند و b^{language} جمله‌ی بایاس نگاشت از توصیف‌های متنی است. به صورت مشابه، برای تصاویر ابتدا استخراج ویژگی به وسیله‌ی شبکه عصبی پیچشی CNN_ζ با پارامترهای ζ انجام می‌شود؛ سپس این ویژگی‌ها با نگاشت خطی W^{visual} و جمله‌ی بایاس b^{visual} به فضای مشترک نگاشته می‌شوند. در نهایت یادگیری این پارامترها به صورت توام با یکدیگر با تابع هزینه‌ی بیشینه حاشیه روی تابع مطابقت F انجام می‌شود.

در [۲۴] نیز که برای اولین بار توصیف تنها نام برچسب دسته‌ها در نظر گرفته شده، از نگاشت دوخطی استفاده شده است. در این روش نام برچسب‌ها با استفاده از مدل نهان‌سازی کلمات word2vec به بردارهایی نگاشته می‌شوند. ابعاد فضای نهان‌سازی کلمات یک پارامتر است که در این مقاله با اعتبار سنجی تعیین شده است. استخراج ویژگی از تصاویر با استفاده از شبکه عصبی پیچشی [۳۲] که روی دسته‌های دیده شده آموزش داده شده، انجام می‌شود. در نهایت یک تابع یشترین حاشیه[۲۶] برای یادگیری نگاشت دوخطی پیشنهاد می‌شود:

$$L((x_n, y_n); W) = \sum_{y \neq y_n} \max(0, \xi - x_n W c_{y_n} + x_n W c_y). \tag{۲-۱۸}$$

[۲۵] Bag of Words
[۲۶] Max Margin

که در آن ξ حاشیه دسته‌بندی است. دسته‌بندی نمونه‌های جدید با نگاشتن x به فضای برچسب‌ها و استفاده از دسته‌بند نزدیکترین همسایه صورت می‌گیرد.

۲-۶-۲ روش‌های مبتنی بر خطای مجموع مربعات

یک نحوه‌ی استفاده دیگر از نگاشت‌های دوخطی، دسته‌بندی مستقیم با این نگاشت است. در مقاله [۱۸] چنین رویکردی پیش گرفته شده و از مسئله‌ی بهینه‌سازی زیر استفاده شده است.

$$\underset{W\in\mathbb{R}^{d\times a}}{\text{minimize}} \left\|X_s^T W C_s - Y_s\right\|_{Fro} + \Omega(W), \tag{۱۹-۲}$$

که در آن Ω یک جمله منظم‌سازی است. در این حالت اگر تبدیل را از فضای تصاویر به فضای صفت‌ها نگاه کنیم، نگاشت W باید تصاویر را به زیرفضایی عمود به تمامی بردار صفت‌های مربوط به برچسب‌های نادرست بنگارد. عملکرد خوب این روش، با وجود استفاده از تابع هزینه ساده مجموع مربعات خطا که در یادگیری ماشین تابع هزینه‌ی مناسبی برای دسته‌بندی به شمار نمی‌آید، به جمله منظم سازی آن نسبت داده می‌شود. جمله منظم‌سازی Ω به این صورت تعریف می‌شود:

$$\Omega(W) = \lambda \left\|WC_s\right\|_{Fro} + \gamma \left\|X_s^T W\right\|_{Fro} + \lambda\gamma \left\|W\right\|_{Fro}, \tag{۲۰-۲}$$

این جمله منظم‌سازی با دیدگاه نگاشت دوخطی طبیعی است. چرا که ماتریس WC_S را می‌توان یک دسته‌بند خطی روی فضای تصاویر در نظر گرفت و از طرفی ماتریس $X_s^T W$ یک دسته‌بند روی بردارهای صفت است در نتیجه طبیعی است که پارامترهای این دو دسته‌بند با نرم فروبنیوس آن‌ها کنترل شود تا از بیش‌برازش[۲۷] جلوگیری شود. استفاده از توابع نرم دوم برای خطا و منظم‌سازی در این روش باعث شده است که مسئله بهینه‌سازی جواب به صورت فرم بسته داشته باشد و زمان اجرا نسبت به سایر روش‌ها بسیار کمتر باشد.

این روش در [۳۳] برای توصیفات متنی توسعه داده شده است. با توجه به ابعاد بالای داده‌های متنی و همچنین نویز زیادی که در آن‌ها در مقایسه با بردارهای صفت وجود دارد، ماتریس تبدیل W به دو ماتریس تجزیه می‌شود:

$$W = V_x^T V_c. \tag{۲۱-۲}$$

[۲۷] Over Fitting

با این تجزیه از افزایش شدید تعداد پارامترها در اثر افزایش بعد بردار توصیف‌ها جلوگیری می‌شود (دقت کنید که بعد W در رابطه (۲-۱۹) برابر $d \times a$ است). علاوه بر این V_c می‌تواند برای استخراج ویژگی‌های مفید و حذف نویز از C_s به کار گرفته شود و V_x مانند W در حالت اصلی عمل کند؛ یعنی پارامترهای یک دسته‌بند را از روی توصیف‌ها تولید کند. در نهایت تابع هزینه برای این روش به صورت زیر تعریف می‌شود:

$$\min_{V_x,V_c} \left\|X_s^T V_x^T V_c C_s - Y_s\right\|_{Fro} + \lambda_1 \left\|V_x^T V_c C\right\|_{Fro} + \lambda_2 \left\|V_c^T\right\|_{2,1}, \qquad (۲۲-۲)$$

که $\left\|M^T\right\|_{2,1} = \sum_i \left\|M_{(i)}\right\|_2$ و این نوع منظم‌سازی، ستون‌های ماتریس V_c را به سمت تنک بودن سوق خواهد داد. در واقع اگر λ_2 بزرگ انتخاب شود، V_c نقش یک ماتریس انتخاب ویژگی[۲۸] را خواهد داشت. جمله‌های منظم سازی دیگر در (۲-۲۰) به دلیل تاثیر اندکشان در آزمایشات عملی حذف شده‌اند.

۲-۷ نگاشت به فضای تصاویر

در برخی از روش‌ها فضای مشترک فضای ویژگی‌های تصویر است و نگاشتی از توصیف‌ها به این فضا یاد گرفته می‌شود و مطابقت تصویر و توصیف در این فضا قابل سنجیدن می‌شود. از آن‌جا که در این روش‌ها، استخراج ویژگی از تصاویر با توابع از پیش معین صورت می‌گیرد این روش‌ها را با عنوان نگاشت به فضای تصاویر بررسی می‌کنیم.

یک تعمیم از SJE در [۳۴] ارائه شده است. در این روش برای تصاویر مجموعه متون بزرگتری نسبت به دادگان قبلی [۱۰] جمع‌آوری و استفاده شده است. این ازدیاد داده‌ها امکان آموزش مدل‌های پیچیده‌تر و پیشرفته‌تر را برای یادگیری نگاشت توصیف دسته‌ها به فضای مشترک، فراهم می‌کند. در نتیجه فاصله میان عمل‌کرد یادگیری صفرضرب هنگام استفاده از توصیف‌های متنی و توصیف‌های به صورت بردار صفت را کمتر کرده است. در این حالت فرض می‌شود که داده‌های آموزش به صورت $\{(v_n, t_n, y_n), n = 1, ..., N\}$ است که متشکل است از $v \in \mathcal{V}$ که ویژگی‌های تصویری هستند، $t \in \mathcal{T}$ توصیفات متنی و $y \in \mathcal{Y}$ برچسب‌ها. دقت کنید که در توصیف این روش بر خلاف سایر روش‌ها از نمادگذاری معرفی شده در این بخش استفاده نکرده‌ایم. نمادهای استفاده شده منطبق بر نمادهای مقاله اصلی می‌باشند. دلیل این موضوع این است که ویژگی‌های تصویری v_n با تصاویر x_n متفاوت است. در نمادگذاری ما هر x در رابطه یک‌به‌یک با یک تصویر آموزش یا آزمون است در حالی‌که در مجموعه آموزش معرفی‌شده در بالا هر تصویر با چند مجموعه ویژگی بصری v در مجموعه آموزش حضور دارد و هر کدام از این ویژگی‌های بصری v_n، یک متن مربوط به خود دارد که با t_n

[۲۸] Feature Selection

نشان داده شده است. هم‌چنین فرض کنید که $\mathcal{T}(y)$ و $\mathcal{V}(y)$ به ترتیب مجموعه تمامی متون و ویژگی‌های بصری مربوط به کلاس y را نشان می‌دهند. در این حالت هدف یادگیری تابع مطابقت $F : \mathcal{V} \times \mathcal{T} \to \mathbb{R}$ میان تصاویر و توصیف‌هاست. که به صورت

$$F(v,t) = \theta(v)^T \phi(t), \tag{۲۳-۲}$$

در نظر گرفته شده است. با داشتن چنین تابعی، مشابه سایر روش‌ها پیش‌بینی برچسب برای تصاویر یا حتی متون جدید با معادلات زیر صورت می‌پذیرد:

$$f_v(v) = \arg\max_{y \in \mathcal{Y}} (\mathbb{E}_{t \sim \mathcal{T}(y)}[F(v,t)]), \tag{۲۴-۲}$$

$$f_t(t) = \arg\max_{y \in \mathcal{Y}} (\mathbb{E}_{v \sim \mathcal{V}(y)}[F(v,t)]). \tag{۲۵-۲}$$

یادگیری تابع F با تابع هزینه‌ی زیر صورت می‌گیرد:

$$\frac{1}{N} \sum_{n=1}^{N} \ell_v(v_n, t_n, y_n) + \ell_t(v_n, t_n, y_n), \tag{۲۶-۲}$$

که توابع ℓ_t و ℓ_v این گونه تعریف شده اند:

$$\ell_v(v_n, t_n, y_n) = \max_{y \in \mathcal{Y}}(0, \Delta(y_n, y) + \mathbb{E}_{t \sim \mathcal{T}(y)}[F(v_n, t) - F(v_n, t_n)]),$$

$$\ell_t(v_n, t_n, y_n) = \max_{y \in \mathcal{Y}}(0, \Delta(y_n, y) + \mathbb{E}_{v \sim \mathcal{V}(y)}[F(v, t_n) - F(v_n, t_n)]).$$

تفاوت تابع هزینه (۲-۲۶) با رابطه (۲-۱۳) در اضافه شدن جمله‌ی دوم است. در رابطه (۲-۱۳) این مسئله که هر تصویر طوری نگاشته شود که به توصیف درست نزدیک‌تر از بقیه توصیف‌ها باشد در نظر گرفته می‌شد. در رابطه بالا علاوه به این مسئله، نگاشت‌ها باید طوری باشد که هر توصیف به ویژگی بصری خود نزدیک‌تر باشد تا سایر ویژگی‌های بصری. نگاشت θ مانند سایر روش‌ها یک شبکه عصبی ژرف پیچشی است که از قبل با داده‌های ImageNet آموزش داده شده است. برای هر تصویر قسمت‌های بصری مختلف با بریدن قسمت‌های متفاوت از تصویر حاصل می‌شود. نگاشت ϕ برای متون با سه شبکه عصبی مختلف پیچشی، بازگردنده و پیچشی بازگردنده (CNN-RNN) مدل شده است. استفاده از این شبکه‌ها برای نگاشت متن نخستین بار در این روش رخ داده‌است. جمع‌آوری مجموعه دادگان متنی بزرگتر، آموزش چنین شبکه‌هایی را ممکن کرده است.

در [۱۰] که برای نخستین بار توصیف‌ها از نوع متنی مورد بررسی قرار گرفته شده است، راه‌حل پیشنهادی یادگیری نگاشتی از این توصیفات به فضای تصاویر است. حاصل این نگاشت یک دسته‌بند خطی در فضای تصاویر در نظر گرفته می‌شود. اگر این نگاشت را طبق نمادگذاری معرفی شده با ϕ نشان دهیم دسته بندی با استفاده از رابطه زیر انجام خواهد شد:

$$y^* = \arg\max_y \phi(c^y)^T x. \qquad (۲۷-۲)$$

برای یادگیری $\phi(c)$ از ترکیب دو تخمین‌گر استفاده می‌شود:

۱. رگرسیون احتمالی: توزیع P_{reg} طوری یادگرفته می‌شود که برای یک توصیف c و نگاشت w در فضای تصاویر احتمال $P_{reg}(w|c)$ را مدل می‌کند.

۲. تابع مطابقت: نگاشت دوخطی D که تطابق میان دامنه تصاویر و توصیف‌ها مدل می‌کند به عبارت دیگر $c^T Dx$ زمانی که x به دسته‌ای که c توصیف می‌کند تعلق دارد بزرگتر از مقدار آستانه‌ای است و در غیر این صورت کوچک‌تر از آن. می‌توان مشاهده کرد که در این حالت با استفاده از رابطه (۲-۲۷)، $c^T w$ یک دسته‌بند خطی برای دسته‌ای که c توصیف می‌کند، خواهد بود.

پارامترهای P_{reg} و D با استفاده از نمونه‌های آموزش بدست می‌آیند. در نهایت تابع پیشنهادی برای نگاشت ϕ برای دسته‌های آزمون به صورت زیر تعریف می‌شود:

$$\phi(c) = \arg\min_{w,\zeta_i} w^T w - \alpha c^T Dw - \beta\, ln(P_{reg}(w|c)) + \gamma \sum \zeta_i, \qquad (۲۸-۲)$$

$$s.t. : -(w^t x_i) \geqslant \zeta_i, \quad \zeta_i \geqslant ۰,\ i = ۱, \dots N_s,$$

$$c^T Dc \geqslant l,$$

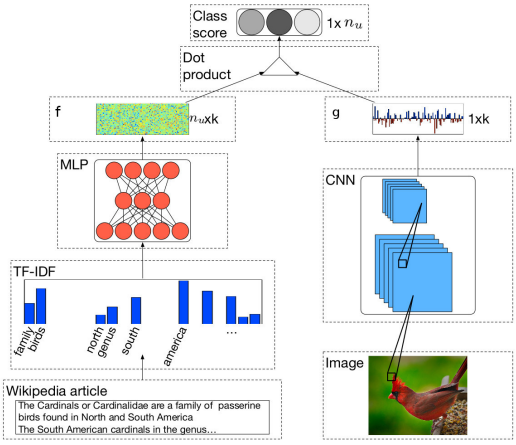
که α, β, γ, l فراپارامترهای مدل هستند. جمله اول در این تابع هزینه، برای منظم‌سازی دسته‌بند خطی w است. جمله دوم مشابهت w با $c^T D$ را الزام می‌کند و جمله سوم، مقدار راستی‌نمایی[۲۹] یک رگرسیون احتمالی برای تخمین w از x است. محدودیت $-(w^T x_i) \geqslant \zeta_i$ بر اساس فرض عدم تعلق نمونه‌های آزمون به کلاس‌های دیده‌شده تعریف شده است

[۲۹] Likelihood

و اجبار می‌کند که تمامی نمونه‌های دیده‌شده باید در طرف منفی دسته‌بند خطی w قرار گیرند. نویسندگان این پژوهش، روش خود را با استفاده از تکنیک هسته [۳۰] برای دسته‌بندهای غیرخطی نیز توسعه داده‌اند [۳۵].

۲-۸ نگاشت به یک فضای میانی

در برخی روش‌ها هر دوی نگاشت‌های ϕ و θ، معرفی شده در ابتدای فصل با توجه به داده‌ها یاد گرفته می‌شوند و در نتیجه فضای مشترک مورد استفاده نه فضای تصاویر و نه فضای توصیف‌هاست؛ بلکه فضای ثالثی است. این فضای میانی در برخی از روش‌ها یک فضای با بعد کمتر است و تعبیر معنایی برای آن موجود نیست. در برخی روش‌های دیگر، فضای میانی را با بعد n_s یعنی تعداد دسته‌های دیده شده در نظر گرفته‌اند و تعبیر معنایی برای آن ارائه شده است. این فضای میانی بر اساس توصیف دسته‌ها و نمونه‌های دیده نشده بر اساس شباهت آن‌ها با دسته‌های دیده شده استوار است.



شکل ۲-۲: شبکه مورد استفاده برای یادگیری توام نگاشت تصاویر و توصیف‌ها که یک شبکه عصبی ژرف با دو ورودی است. ورودی اول از نوع تصویر است و ابتدا با یک شبکه پیچشی سپس با چند لایه چگال به فضایی $k-$بعدی می‌رود. ورودی دوم که یک مقاله از ویکی‌پدیای انگلیسی است پس از تبدیل به نمایش برداری به صورت tf-idf با چندلایه با اتصالات چگال پردازش شده و به فضایی $k-$بعدی می‌رود. در نهایت امتیاز تعلق تصویر به دسته‌ی متن با ضرب داخلی این دو نگاشت تعیین می‌شود [۱].

در [۱] از شبکه‌های عصبی ژرف برای یادگیری توام نگاشت‌های ϕ و θ استفاده شده است. نمای کلی شبکه مورد استفاده در این روش در تصویر ۲-۲ نشان داده شده است. توصیف‌های متنی و ویژگی‌های بصری دو ورودی جداگانه به چنین شبکه‌ای هستند که ابتدا به صورت جداگانه با یک یا چند لایه‌ی با اتصالات کامل به یک فضای مشترک نگاشته

[۳۰] kernel trick

شده و سپس بر اساس شباهت نمایش آن‌ها در این فضای میانی دسته‌بندی می‌شوند. تفاوت این روش با سایر روش‌هایی که مرور شد یادگیری توامان نگاشت‌های ϕ و θ است که با استفاده از شبکه‌های عصبی ممکن شده است. معیار یادگیری این دو نگاشت تنها خطای دسته‌بندی نهایی است. این روش را می‌توان به صورت ساخت دسته‌بند از روی توصیفات نیز تعبیر کرد؛ با این تفاوت که در این حالت یک تبدیل نیز روی فضای تصاویر اعمال شده و سپس دسته‌بند خطی یادگرفته شده از متون در این فضا به نگاشت تصاویر اعمال می‌شود. در این حالت دسته‌بند خطی w^y یک تابع غیرخطی از توصیف کلاس y است: $w^y = f(c^y)$ که f شبکه عصبی مخصوص متن است (نیمه‌ی چپ تصویر ۲-۲). استخراج ویژگی غیرخطی از تصاویر نیز با یک شبکه عصبی که تابع آن را g می‌نامیم، انجام شده است (نیمه‌ی راست تصویر ۲-۲). در نهایت دسته‌بندی با تابع زیر انجام می‌شود:

$$y^* = \arg\max_y w^{yT} g(x). \qquad (۲۹-۲)$$

این روش فراتر از دسته‌بند خطی به حالت فوق نیز با معرفی دسته‌بند پیچشی توسعه پیدا می‌کند. در شبکه‌های عصبی پیچشی، اطلاعات مکانی در لایه‌های با اتصال چگال که بعد از لایه‌های پیچشی قرار می‌گیرند، از بین می‌رود. هم‌چنین تعداد وزن‌ها در این لایه‌ها بسیار بیشتر از لایه‌های پیچشی زیرین است. در نتیجه بنظر می‌رسد استفاده مستقیم از خروجی لایه‌ی پیچشی و اضافه کردن یک لایه پیچشی دیگر که یادگیری فیلتر آن بر اساس متن انجام می‌شود، می‌تواند راه‌حل مناسب‌تری از یادگرفتن یک یا چند لایه‌ی چگال باشد.

فرض کنید b خروجی یک لایه‌ی پیچشی با M نقشه از ویژگی‌های تصویر باشد: $b \in \mathbb{R}^{M \times l \times h}$ که h و l ارتفاع و عرض نقشه ویژگی‌ها هستند. دسته‌بند روی b به صورت یک لایه‌ی پیچشی فورمول‌بندی می‌شود. ابتدا یک کاهش ابعاد غیرخطی روی هر یک از نقشه‌های ویژگی صورت می‌گیرد که آن را با g' نشان می‌دهیم: $g' : \mathbb{R}^{M \times l \times h} \mapsto \mathbb{R}^{K' \times l \times h}$ که $K' << M$. در ادامه از نماد a' برای نقشه ویژگی کاهش بعد یافته استفاده می‌کنیم $a' = g'(a)$. از یک توصیف مثل c^y یک فیلتر پیچش[۳۱] $w^y = f'(c^y)$ ایجاد می‌شود که اگر اندازه فیلتر را با m نشان دهیم: $w'_c \in \mathbb{R}^{K' \times m \times m}$. همانند حالت قبل، f' با یک شبکه عصبی چند لایه مشخص می‌شود. در نهایت دسته‌بند پیچشی به صورت زیر تعریف می‌شود:

$$\text{score}(x, y) = o\left(\sum_{i=1}^{K'} w_i^{y'} \overset{*}{,} a'_i\right), \qquad (۳۰-۲)$$

که $\text{score}(x, y)$ امتیاز تعلق x به دسته‌ی y است؛ $o(\cdot)$ یک تابع ادغام[۳۲] به صورت $o : \mathbb{R}^{l \times h} \mapsto \mathbb{R}$ و $\overset{*}{,}$ نشانگر عمل

[۳۱] Convolution

[۳۲] pooling

پیچش[۳۳] است. در این حالت فیلترهای یادگرفته شده به علت این که به محل تصویر وابسته هستند می‌توانند با دقت بهتری تطابق توصیف‌های متنی و تصویر را نشان دهند.

در نهایت در این پژوهش استفاده همزمان از دسته‌بندهای خطی و پیچشی پیشنهاد می‌شود که با استفاده از آزمایشات عملی نشان داده شده عمل‌کرد بهتری خواهد داشت. برای استفاده همزمان از این دو دسته‌بند امتیاز تطابق از جمع این دو بدست می‌آید:

$$\text{score}(x,y) = w^{yT} g(x) + o\Big(\sum_{i=1}^{K'} w_i^{y'} \check{*}\, g'(a)_i\Big), \tag{۳۱-۲}$$

در این حالت پارامترهای مربوط به g, g', f, f' به صورت همزمان یادگرفته می‌شوند. یادگیری در شبکه بر اساس خطای تنها خروجی شبکه که نشان می‌دهد آیا این متن و توصیف هم‌دسته هستند یا نه، صورت می‌گیرد. در این پژوهش دو تابع هزینه برای خطا در نظر گرفته شده ۱) آنتروپی تقاطعی [۳۴] ۲) تابع هزینه لولا[۳۵]. بررسی عمل‌کرد این دو نوع تابع هزینه نشان می‌دهد که بر اساس معیار ارزیابی نهایی هر کدام می‌تواند عمل‌کرد بهتری نسبت به دیگری داشته باشد. اگر معیار ارزیابی دقت دسته‌بندی در k انتخاب اول[۳۶] باشد تابع هزینه لولا بهتر عمل می‌کند و اگر معیار مساحت زیر نمودار دقت و فراخوان[۳۷] باشد، آنتروپی متقاطع عمل‌کرد بهتری دارد.

در [۱۲] روشی برای ساخت بردارهای صفت برای تصاویر، برای دسته‌بندی بهتر آن‌ها، در حالت عادی دسته‌بندی تصاویر، ارائه شده است. این روش برای هر دسته یک بردار صفت و برای هر یک از صفت‌ها یک دسته‌بند یاد می‌گیرد. این روش برای یادگیری صفرضرب هم تعمیم داده شده است. این روش با سایر روش‌ها در نوع توصیفی که برای دسته‌ها استفاده می‌کند کاملا متفاوت است. در این روش بردار صفت برای دسته‌ها جزء خروجی‌های روش است نه ورودی‌های آن. در این‌جا الگوریتم هیچ توصیفی از دسته‌های دیده شده دریافت نمی‌کند و دسته‌های دیده نشده بر اساس شباهتشان با دسته‌های دیده شده توصیف می‌شوند و در نهایت الگوریتم برای همه دسته‌ها بردار صفت تولید می‌کند. فرض کنید در کل n دسته موجود باشد و قصد داشته باشیم بردار صفت‌های l بعدی تولید کنیم (l یک پارامتر است). ماتریس این ویژگی‌ها را با $A \in \mathbb{R}^{n\times l}$ نشان می‌دهیم. هدف در این جا بدست آوردن A و هم‌چنین دسته‌بند $f = [f_1 \dots f_l]^T$ برای

[۳۳] Convolution
[۳۴] Cross Entropy
[۳۵] hinge loss
[۳۶] top-k accuracy
[۳۷] Precision Recall Area Under the Curve

صفت‌هاست. در نهایت یک نمونه با استفاده از رابطه زیر قابل دسته‌بندی خواهد بود:

$$y^* = \arg\min_i \|A_{(i)} - f(x)^T\|. \quad (۳۲-۲)$$

نویسندگان این پژوهش عنوان می‌کنند که بردار صفت یادگرفته شده برای خوب بودن باید دو خاصیت را داشته باشد:

- ایجاد تمایز: بردار صفت هر دسته باید با دسته دیگر، به اندازه کافی متفاوت باشد.به عبارت دیگر سطرهای ماتریس A از هم فاصله داشته باشند.
- قابل یادگیری بودن: صفت‌ها باید با خطای کم از روی تصاویر قابل پیش‌بینی باشند. یک روش برای ایجاد چنین حالتی این است که صفت‌ها باید میان دسته‌های مشابه یکدیگر، شبیه باشد.

اثبات می‌شود خطای دسته‌بندی کرانی بر اساس دو عامل بالا، یعنی حداقل فاصله سطرهای A و حداکثر خطای دسته‌بند f خواهد داشت. برای یادگیری A طوری که دو خاصیت فوق را داشته باشد تابع هزینه

$$\max_A \sum_{i,j} \|A_{(i)} - A_{(j)}\|_۲^۲ - \lambda \sum_{i,j} S_{ij} \|A_{(i)} - A_{(j)}\|_۲^۲ \quad (۳۳-۲)$$

پیشنهاد شده است. $S \in \mathbb{R}^{n \times n}$ ماتریسی است که عناصر آن شباهت میان دسته‌ها را نشان می‌دهد. جمله اول، جمع فاصله سطرهای A از هم است و برای ایجاد خاصیت اول یعنی ایجاد تمایز در نظر گرفته شده است. جمله دوم تحمیل می‌کند که دسته‌های مشابه یکدیگر بایست صفت‌های بصری مشابه داشته باشند تا بتوان این صفت‌ها را از تصویر پیش‌بینی کرد. در مسئله دسته‌بندی عادی، S از روی داده‌های برچسب‌دار و فاصله تصاویر هر دسته از دسته‌ی دیگر تعیین می‌شود. برای مسئله یادگیری صفرضرب، مقادیر S برای دسته‌های دیده نشده به عنوان ورودی دریافت می‌شود و با کمک f که از داده‌های آموزش یادگرفته شده دسته‌بندی آن‌ها با رابطه (۲-۳۲) انجام می‌شود.

۲-۸-۱ نگاشت به فضای دسته‌های دیده شده

با توجه به این که یادگیری تابع تعیین شباهت هر نمونه با دسته‌های آموزش تنها به نمونه‌های آموزش نیاز دارد می‌تواند به طور کامل در زمان آموزش انجام شود. بر این اساس اگر دسته‌های دیده نشده به خوبی بر اساس شباهتشان با دسته‌های دیده شده قابل توصیف باشند، می‌توان یک معیار مطابقت میان آن‌ها و نمونه‌های آزمون بدست آورد (مثلا بر اساس ضرب داخلی یا فاصله اقلیدسی در این فضا). در زمینه‌ی یادگیری صفرضرب چند روش بر این اساس ارائه شده است. بعضی از

این روش‌ها توصیف دسته‌های آزمون بر اساس دسته‌های آموزش را به عنوان ورودی دریافت می‌کنند و برخی دیگر توانایی بدست آوردن این نمایش را بر اساس توصیف‌های جانبی دارند.

در روشی که در [۳۶] ارائه شده است ابتدا هر دسته به صورت نسبتی از دسته‌های دیده شده یا به عبارتی هیستوگرامی از آن‌ها نشان داده می‌شود. سپس بر اساس این نمایش از دسته‌ها و تنها با استفاده از نمونه‌های آموزش، نگاشت از فضای تصاویر به فضای هیستوگرام دسته‌های دیده شده یاد گرفته می‌شود. نمایش توصیف c با استفاده از رابطه زیر بدست می‌آید:

$$\theta(\mathbf{c}) = \underset{\boldsymbol{\alpha}\in\Delta^{|\mathcal{S}|}}{\arg\min} \left\{ \frac{\gamma}{۲}\|\boldsymbol{\alpha}\|^۲ + \frac{۱}{۲}\|\mathbf{c} - \sum_{y\in\mathcal{S}} \mathbf{c}_y \alpha_y\|^۲ \right\}, \tag{۳۴-۲}$$

که در آن $\Delta^{|\mathcal{S}|}$ یک سادک[۳۸] به ابعاد تعداد دسته‌های دیده شده را نشان می‌دهد. جمله منظم سازی $\frac{\gamma}{۲}\|\boldsymbol{\alpha}\|^۲$ در عبارت بالا، مانع از بدست آمدن این نمایش بدیهی می‌شود که برای دسته‌های دیده شده، تنها عنصر متناظر با همان دسته در $\boldsymbol{\alpha}$ یک شود و سایر درایه‌ها صفر. γ یک فراپامتر در این مدل است که باید با اعتبارسنجی تعیین شود. نگاشت از تصاویر به هیستوگرام‌ها یا به عبارتی تعیین شباهت هر نمونه با دسته‌های دیده شده در این روش به این صورت انجام می‌شود که برای هر یک از دسته‌های دیده شده یک نگاشت اختصاصی برای تعیین شباهت به آن وجود دارد. این نگاشت بر اساس تابع واحد خطی اصلاح‌کننده ReLU[۳۹] یا نگاشت اشتراک INT تعریف می‌شود که سپس با یک تبدیل خطی مشترک w به امتیاز شباهت تبدیل می‌شود. اگر نگاشت مربوط به دسته‌ی y را با $\psi_y(\cdot)$ نشان دهیم، داریم:

$$\text{INT:} \quad \phi_y(\mathbf{x}) = \min(\mathbf{x}, \mathbf{v}_y), \tag{۳۵-۲}$$

$$\text{ReLU:} \quad \phi_y(\mathbf{x}) = \max(۰, \mathbf{x} - \mathbf{v}_y), \tag{۳۶-۲}$$

که v_y نگاشت اختصاصی شباهت با دسته‌ی y است. در آزمایشات عملی نشان داده شده است که نگاشت‌های ReLU و INT عملکرد نسبتا مشابهی دارند. در نهایت امتیاز شباهت با دسته‌ی y با عملگر خطی w تعیین می‌شود و خواهیم داشت:

$$\phi(x) = \big(w^T\psi_۱(x), w^T\psi_۲(x), \ldots, w^T\psi_{n_s}(x)\big). \tag{۳۷-۲}$$

دسته‌بندی نمونه‌های آزمون با ضرب داخلی در فضای هیستوگرام‌ها تعیین می‌شود:

$$y^* = \underset{y\in\mathcal{Y}}{\arg\max} \langle \phi(x), \theta(c^y) \rangle. \tag{۳۸-۲}$$

[۳۸] Simplex

[۳۹] Rectified Linear Unit

یادگیری w و v با استفاده از مسئله بهینه‌سازی زیر تعیین صورت می‌گیرد:

$$\min_{\mathcal{V},\mathbf{w},\boldsymbol{\xi},\boldsymbol{\epsilon}} \frac{۱}{۲}\|\mathbf{w}\|^۲ + \frac{\lambda_۱}{۲}\sum_{\mathbf{v}\in\mathcal{V}}\|\mathbf{v}\|^۲ + \lambda_۲\sum_{y,s}\epsilon_{ys} + \lambda_۳\sum_{i,y}\xi_{iy}, \tag{۳۹-۲}$$

$$\text{s.t. } \forall i \in \{۱,\cdots,N\}, \forall y \in \mathcal{S}, \forall s \in \mathcal{S},$$

$$\sum_{i=۱}^{N}\frac{\mathbb{I}_{\{y_i=y\}}}{N_y}\Big[f(\mathbf{x}_i,y) - f(\mathbf{x}_i,s)\Big] \geqslant \Delta(y,s) - \epsilon_{ys}, \tag{۴۰-۲}$$

$$f(\mathbf{x}_i,y_i) - f(\mathbf{x}_i,y) \geqslant \Delta(y_i,y) - \xi_{iy}, \tag{۴۱-۲}$$

$$\epsilon_{ys} \geqslant ۰, \xi_{iy} \geqslant ۰, \forall \mathbf{v} \in \mathcal{V}, \mathbf{v} \geqslant ۰,$$

که در آن $\Delta(\cdot,\cdot)$ یک تابع هزینه‌ی خطای ساختار مند میان دسته‌ی پیش‌بینی شده و دسته‌ی صحیح را نشان می‌دهد $\lambda_۱ \geqslant ۰$، $\lambda_۲ \geqslant ۰$ and $\lambda_۳ \geqslant ۰$ فراپارمترهای مربوط به منظم‌سازی هستند و $\boldsymbol{\epsilon} = \{\epsilon_{ys}\}$ and $\boldsymbol{\xi} = \{\xi_{iy}\}$ متغیرهای مربوطه به محدودیت‌های نرم در بهینه‌سازی‌اند. در این روش تابع هزینه‌ی خطای ساختارمند به صورت $\Delta(y,s) = ۱ - \mathbf{c}_y^T\mathbf{c}_s$ تعریف شده است.

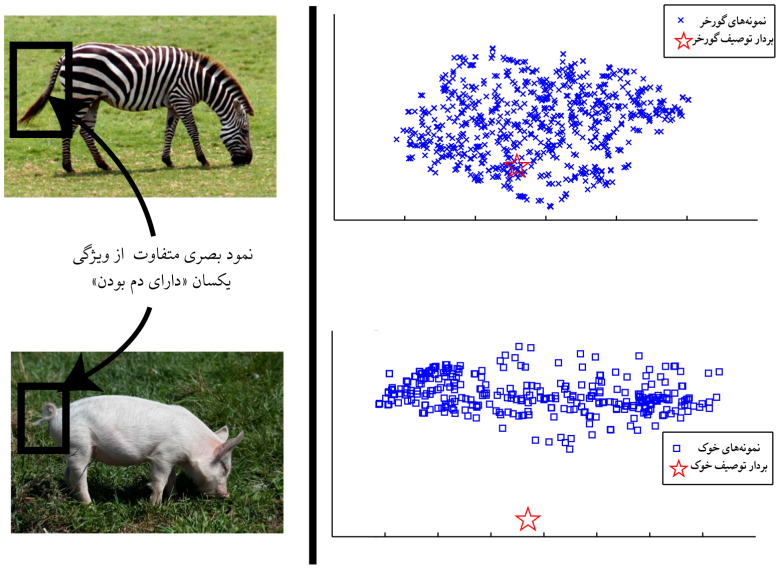
صورت‌بندی بالا یک صورت‌بندی دسته‌بندی با بیشینه حاشیه است با این تفاوت که علاوه بر محدودیت بیشینه حاشیه (رابطه(۲-۴۱)) یک محدودیت برای دسته‌بندی صحیح به صورت میانگین هم در رابطه (۲-۴۰) اضافه شده است. این محدودیت جدید می‌تواند باعث شود که دادها به گونه‌ای نگاشته شود که نه تنها دسته‌بندی صحیح صورت گیرد بلکه یک توزیع با مرکز $\theta(c^y)$ ایجاد کنند و برای نگاشت‌یافته‌ی مراکز دسته‌ها نیز یک حاشیه در نظر گرفته شود. این حالت باعث ایجاد خوشه‌هایی جدا از هم می‌شود که مراکزشان توصیف‌هاست و در نتیجه برای مسئله یادگیری از صفر مناسب‌تر است.

نویسندگان این پژوهش روش خود را در [۳۷] با یادگیری توامان نگاشت توصیف‌ها و تصاویر توسعه داده‌اند. علاوه بر یادگیری توامان پارامترهای نگاشت‌ها، برای داده‌های آزمون، نمایش طوری به دست می‌آید که علاوه بر هم‌خوانی با پارامترهای بدست آمده برای نگاشت، از داده‌های دسته‌های دیده شده نیز دور باشند. این یک شرط شهودی برای بهتر شدن نگاشت است چرا که فرض بر این است که دسته‌های آموزش و آزمون اشتراکی ندارند و در نتیجه برای مثال نمایش تصاویر آزمون نباید در نزدیکی توصیف دسته‌های آموزش باشد.

۲-۹ روش‌های نیمه‌نظارتی

در این بخش به بررسی روش‌های نیمه‌نظارتی می‌پردازیم. این روش‌ها از نظر نوع نگاشت‌های مورد استفاده در یکی از دسته‌های قبلی قابل بیان بودند ولی با توجه به این که روش پیشنهادی ما نیز نیمه‌نظارتی است، برای پر رنگ‌تر شدن نحوه‌های استفاده از داده‌های آزمون در جریان آموزش این دسته را به طور جداگانه مورد بررسی قرار می‌دهیم.

در [۳۸] برای نخستین بار مشکل جابجایی دامنه[۴۰] معرفی شد. این مشکل که در شکل ۲-۳ قابل مشاهده است به متفاوت بودن خواص صفت‌ها برای دسته‌های مختلف اشاره می‌کند. برای مثال صفت راه‌راه بودن برای دو حیوان گورخر و ببر از نظر بصری خواص متفاوتی دارد و یادگیری یک دسته‌بند برای تشخیص راه‌راه بودن با استفاده از تصاویر گورخر در تشخیص وجود و یا عدم وجود این صفت در تصویر ببر ضعیف خواهد بود.



شکل ۲-۳: مشکل جابجایی دامنه بین دو دسته‌ی دیده شده (گورخر) و دیده نشده (خوک) نمایش داده شده است. صفت یکسان «دارای دم بودن» در این دو دسته دارای دو نمود بصری متفاوت است (سمت چپ) و نگاشت یادگرفته شده برای بردن این صفت به فضای مشترک برای دسته‌ی دیده نشده عملکرد ضعیف‌تری نسبت به دسته‌ی دیده شده به نمایش می‌گذارد (سمت راست) [۳۸].

در [۳۸] برای حل این مشکل دو تکنیک به کار گرفته شده است. ابتدا یافتن نمایش مشترک برای سه دامنه‌ی تصاویر، بردار صفت و بردار نام دسته‌ها به صورت توامان با استفاده از CCA[۴۱][۳۹] و سپس برچسب‌گذاری داده‌های بدون

[۴۰] Domain shift problem

[۴۱] Canonical Correlation Analysis

برچسب در این فضای مشترک با استفاده از یک تکنیک انتشار برچسب[۴۲] بیزی.

در [۴۰] مسئله به صورت یک دسته‌بندی روی دسته‌های دیده شده و نسبت دادن برچسب به داده‌های دسته‌های دیده نشده مدل شده‌است. در این روش یک دسته‌بند خطی روی تصاویر یادگرفته می‌شود که این دسته‌بند ترکیبی از پارامترهای مدل و توصیف‌هاست. به صورت دقیق‌تر چارچوب یادگیری برابر خواهد بود با:

$$\min_{Y,U,W,\boldsymbol{\xi}} \quad \frac{\beta}{۲}\left\|W\right\|_{Fro}^{۲}+\frac{\beta}{۲}\left\|U\right\|_{Fro}^{۲}+\mathbf{۱}^{T}\boldsymbol{\xi}, \tag{۴۲-۲}$$

$$s.t. \quad diag\big((Y-\mathbf{۱}\mathbf{۱}_{k}^{T})\big)UWX^{T})\geqslant(\mathbf{۱}-Y\mathbf{۱}_{k})-\boldsymbol{\xi},\ \forall k\in\mathcal{Y}, \tag{۴۳-۲}$$

$$Y\in\{۰,۱\}^{(N_s+N_u)\times(n_s+n_u)}, \quad BY=Y_s^T,, \tag{۴۴-۲}$$

$$Y\mathbf{۱}=\mathbf{۱}, \quad l\mathbf{۱}\leqslant Y^T\mathbf{۱}\leqslant h\mathbf{۱}, \tag{۴۵-۲}$$

که در این صورت‌بندی فوق، U را می‌توان توصیف‌های موجود برای هر دسته در نظر گرفت، Y برچسب‌ها را نشان می‌دهد و B یک ماتریس انتخاب‌گر است که قسمتی از Y را که مربوط به نمونه‌های آموزش است انتخاب می‌کند. β و l و h فراپارامترهای مدل هستند که β وزن جمله منظم‌سازی را تعیین می‌کند و l وh حداقل و حداکثر نمونه‌هایی که باید هر دسته دریافت کند را تعیین می‌کنند. یک خاصیت جالب این صورت‌بندی این است که اگر دوگان مسئله بهینه‌سازی فوق را بنویسیم، U تنها به شکل UU^T ظاهر می‌شود، یعنی تنها اطلاعاتی که از دسته‌ها نیاز است میزان شباهتشان به یکدیگر است که ممکن است از روی کواریانس توصیف‌ها محاسبه شود، اما در نبود توصیف به صورت مستقیم هم قابل بیان است. در این چارچوب اگر U را ثابت در نظر بگیریم، W یک دسته‌بندی SVM روی دسته‌های دیده شده انجام می‌دهد و برچسب نمونه‌های مربوط به دسته‌های دیده نشده هم به گونه‌ای پیدا می‌شود که علاوه بر ارضای شرایط تابع هدف مربوطه حداقل شود. ضعف این چارچوب در عدم استفاده از اطلاعات موجود در موقعیت مکانی داده‌های آزمون در دسته‌بندی انجام شده روی آن‌هاست و هم‌چنین مسئله بهینه‌سازی تعریف شده برای داده‌های واقعی یک مسئله سخت است که به منابع زمانی و محاسباتی زیادی نیاز دارد. برای حل مشکل اول، نویسندگان این پژوهش نوع دیگری از چارچوب فوق ارائه می‌کنند که با اضافه کردن یک جمله هموار سازی اطلاعات نزدیکی مکانی نمونه‌ها را وارد می‌کند.

$$\min_{Y,U,W} \quad \sum_{i=۱}^{N_s+N_u}\ell(X_{(i)}^{T}W,Y_iU)+\frac{\alpha}{۲}\left\|W\right\|_{Fro}^{۲}+\frac{\beta}{۲}\left\|U-U_{\cdot}\right\|_{Fro}^{۲}+\frac{\rho}{۲}tr(Y_uLY_u^T), \tag{۴۶-۲}$$

$$s.t. \quad (۴۴\text{-}۲),\ (۴۵\text{-}۲)$$

[۴۲] Label Propagation

که در آن α و ρ فراپامترهای جملات منظم‌سازی هستند و U. ماتریس توصیف دسته‌هاست. L ماتریس لاپلاسین مربوط به ماتریس مشابهت میان نمونه‌هاست که در اینجا عکس فاصله اقلیدسی نمونه‌ها به عنوان شباهت در نظر گرفته شده است. به عبارتی اگر A ماتریس متقارنی باشد که عکس فاصله دوبدوی نمونه‌های آزمون را از یکدیگر نشان می‌دهد، خواهیم داشت $L = diag(A\mathbf{1}) - A$. صورت‌بندی معادله (۲-۴۶) با صورت بندی انجام شده در (۲-۴۲) چند تفاوت دارد. اضافه شدن جمله لاپلاسین برای استفاده بهتر از اطلاعات موجود در نمونه‌های آزمون یکی از آن‌هاست. علاوه بر این، در این روش یادگیری نمایش برای برچسب‌ها همواره صورت می‌گیرد. این در حالی‌ست که در صورت‌بندی قبلی U عموما برابر با توصیف‌های موجود در صورت مسئله در نظر گرفته می‌شد. در اینجا U. چنین مقداری را اختیار می‌کند و U اجازه دارد تغییر کند تا نمایش بهتری یاد گرفته شود. این دو روش، علاوه بر نیمه‌نظارتی بودن، تفاوت مهم دیگری با سایر روش‌های ارائه شده برای یادگیری صفرضرب دارند: در این دو روش برچسب‌های داده‌های آزمون در جریان بهینه‌سازی حدس زده می‌شوند و از روش‌هایی مثل نزدیک‌ترین همسایه یا انتشار برچسب به عنوان یک مرحله جداگانه برای تعیین برچسب داده‌ها استفاده نمی‌شود. ضعف این روش‌ها سنگین بودن مسئله بهینه‌سازی تعریف شده است که به همین علت امکان استفاده از نمایش ابعاد بالا برای تصاویر که از شبکه‌های ژرف به دست می‌آید، از بین می‌رود.

در [۴۱] مسئله یادگیری صفرضرب به صورت یک مسئله تطبیق دامنه[۴۳] مدل می‌شود. مسئله دسته‌بندی به صورت صفرضرب ذاتا یک مسئله تطبیق دامنه نیست. در مسئله تطبیق دامنه، یک پیش‌بینی یکسان روی داده‌هایی از دو دامنه متفاوت انجام می‌شود؛ حال آن‌که در مسئله یادگیری صفرضرب علاوه بر تفاوت دامنه در نمونه‌ها، پیش‌بینی‌ها نیز برد متفاوتی دارند و در دسته‌های یکسانی نمی‌گنجد. اگر مسئله یادگیری صفرضرب را به شیوه یافتن توصیف از روی تصاویر، یا به عبارتی پیش‌بینی صفت نگاه کنیم، این مسئله یک مسئله استاندارد تطبیق دامنه بدون نظارت است؛ چرا که یک مجموعه صفت یکسان برای داده‌هایی از دو دامنه متفاوت پیش‌بینی می‌شوند. در این روش، از یادگیری لغت‌نامه[۴۴] برای پیش‌بینی صفت استفاده می‌شود و با معرفی دو جمله منظم‌سازی، مسئله تطبیق دامنه و مشکل جابجایی دامنه در نظر گرفته می‌شوند. برای هر یک از دامنه‌ها یک لغت‌نامه یادگرفته می‌شود که این شامل نمایش هر یک از صفت‌ها در فضای تصاویر است. سپس هر تصویر با توجه به اینکه چه میزان از هر صفت در آن وجود دارد، به صورت ترکیب این پایه‌ها بیان می‌شود. برای دامنه دسته‌های دیده شده، با توجه به این که صفت‌ها از پیش دانسته شده است، مسئله در حقیقت یافتن یک نگاشت

[۴۳] Domain Adaptation

[۴۴] Dictionary Learning

خطی است، نه یادگیری یک لغت‌نامه:

$$D_s = \arg\min_{D_s} \|X_s - D_s Z_s\|_{Fro}^2 + \gamma \|D_s\|_{Fro}^2, \quad s.t. \ \left\|D_{(i)}\right\|_2^2 \leqslant 1, \tag{۴۷-۲}$$

که γ یک پارامتر و D_s نگاشت خطی مورد نظر یا به عبارتی پایه‌های لغت‌نامه است. برای دامنه آزمون، صفت‌های تصاویر دانسته نیستند در نتیجه یک مسئله یادگیری لغت‌نامه داریم که باید صفت‌ها همراه با پایه‌های لغت‌نامه D_u یادگرفته شوند:

$$\begin{aligned} \{D_u, Z_u\} = \min_{D_u, Z_u} & \|X_u - D_u Z_u\|_{Fro}^2 + \lambda_1 \|D_u - D_s\|_{Fro}^2 \\ & + \lambda_2 \sum_{i,j} w_{ij} \left\|Z_{u(i)} - S_{u(j)}\right\|_2^2 + \lambda_3 \|Z_u\|_1 \\ s.t. \quad & \left\|D_{(i)}\right\|_2^2 \leqslant 1 \end{aligned} \tag{۴۸-۲}$$

که در آن λ_1 و λ_2 و λ_3 پارامتر مدل هستند. w_{ij} امتیاز شباهت نمونه‌ی $X_u(i)$ به دسته‌ی j از دسته‌های دیده نشده است که با روش IAP بدست آمده است. در تابع هزینه‌ی فوق، جمله‌ی اول و آخر، جملات معمول مربوط به یادگیری لغت‌نامه‌ی تنک هستند. جمله‌ی دوم برای تطبیق دامنه اضافه شده است و شبیه بودن پایه‌های لغت‌نامه را میان دو دامنه اعمال می‌کند. به عبارت دیگر نمایش بصری هر یک صفت‌های دو دامنه باید نزدیک به یکدیگر باشد. جمله سوم برای حل مشکل جابجایی دامنه اضافه شده است. این جمله اجبار می‌کند که صفت‌های پیش‌بینی شده برای هر یک تصاویر به امضای دسته‌های آزمون مشابهت داشته باشد. در این روش بعد از پیش‌بینی صفت‌های Z_u برای تصاویر آزمون، از انتشار برچسب برای تعیین دسته‌ها استفاده می‌شود. مزیت این روش سادگی مسئله بهینه‌سازی تعریف شده نسبت به دیگر روش‌های نیمه‌نظارتی است. در انجام بهینه‌سازی تناوبی روی D_u و Z_u، مسئله اول جواب بسته دارد و مسئله دوم یک رگرسیون لاسو[۴۵] است که بسته‌های نرم‌افزاری زیادی برای آن وجود دارد. از طرفی متفاوت در نظر گرفتن D_u و D_s موجه به نظر نمی‌رسد. درست است که خواص بصری هر یک از صفت‌ها برای هر دسته متفاوت است (مثل راه‌راه بودن دسته‌های ببر و گورخر) ولی این تفاوت به دسته‌های دیده شده یا دیده‌نشده مرتبط نیست و بین دو دسته‌ی دیده شده یا دو دسته‌ی دیده نشده نیز وجود دارد.

در [۴۲] روش نیمه‌نظارتی کلمه‌محور SS-Voc[۴۶] ارائه می‌شود که بجای استفاده از نمونه‌های بدون برچسب از توصیف‌هایی (که اینجا کلمه هستند) که نمونه‌ای از آن‌ها موجود نیست استفاده می‌کند. این روش با استفاده از چنین کلماتی سعی در رفع کردن چهار نقص در روش‌های دیگر را دارد. این چهار مورد عبارتند از: ۱) فرض جدا بودن دسته‌های

[۴۵] LASSO Regression

[۴۶] Semi-Supervised VOCabulary informed learning

آموزش و آزمون واقعی نیست و ممکن است در زمان آزمون نمونه‌هایی از دسته‌های دیده شده هم وجود داشته باشد. ۲) مجموعه دسته‌های دیده نشده عموما کم‌تعداد است، در حالیکه در مسائل واقعی تعداد دسته‌های دیده نشده می‌تواند بسیار زیاد باشد. ۳) تعداد زیادی نمونه از دسته‌های دیده شده برای آموزش لازم است. ۴) دانش غنی موجود در رابطه معنایی کلمات (نام دسته‌ها) مورد استفاده قرار نمی‌گیرد. در این روش نگاشتی از تصاویر به فضای معنایی نمایش کلمات یادگرفته می‌شود که به صورت همزمان باید دارای سه خاصیت زیر باشد:

۱. هر تصویر برچسب‌دار نزدیک به نمایش معنایی برچسب خود نگاشته شود.

۲. نمایش هر تصویر در فضای کلمات به نمایش برچسب درست خود نزدیکتر باشد تا به سایر برچسب‌های موجود

۳. نمایش هر تصویر در فضای کلمات به نمایش برچسب درست نزدیکتر باشد تا به سایر کلمات لغت‌نامه.

معیار سومی که برشمرده شد تفاوت اصلی این روش با سایر روش‌هایی مثل [۲۴] است که از تابع هزینه‌ی رتبه‌بند استفاده می‌کنند. در نظر گرفتن فاصله با کلماتی که در مجموعه آموزش و آزمون وجود ندارند باعث می‌شود که این روش توانایی دسته‌بندی مجموعه باز[۴۷] را هم داشته باشد، یعنی حالتی که دسته‌های آزمون از پیش تعیین شده نیستند.

برای تامین خاصیت اول، از تابع هزینه‌ی بیشینه حاشیه استفاده می‌شود:

$$(|\xi|_\epsilon)_j = \max\left\{۰, |W_{\star j}^T \mathbf{x}_i - (\mathbf{c}_{z_i})_j| - \epsilon\right\}, \tag{۴۹-۲}$$

$$\mathcal{L}_\epsilon(\mathbf{x}_i, \mathbf{u}_{z_i}) = ۱^T \mid \xi \mid_\epsilon^۲, \tag{۵۰-۲}$$

که $|\xi|_\epsilon \in \mathbb{R}^a$ و $()_j$، $j-$مین عنصر بردار را نشان می‌دهد. این جمله مشابه تابع هزینه رگرسیون بردار پشتیبان[۴۸] است که با استفاده از جمله‌ی درجه ۲ هموار شده است.

برای تامین موارد دوم و سوم برای نگاشت از جمله زیر استفاده می‌شود:

$$\mathcal{M}(\mathbf{x}_i, \mathbf{c}_{y_i}) = \frac{۱}{۲}\sum_v \left[G + \frac{۱}{۲}D(\mathbf{x}_i, \mathbf{c}_{y_i}) - \frac{۱}{۲}D(\mathbf{x}_i, \mathbf{c}_v)\right]_+^۲, \tag{۵۱-۲}$$

که در آن v نمایش یک کلمه در فضای معنایی است، G متغیر مربوط به حاشیه است و $[\cdot]_+^۲$ نشان‌دهنده‌ی تابع هزینه‌ی لولای هموار شده [۴۹] است. برای این که بهینه‌سازی امکان‌پذیر باشد v بجای کل کلمات لغت‌نامه تنها چند مقدار نزدیک

[۴۷] Open Set

[۴۸] Support Vector Regression

[۴۹] quadratically smoothed hinge loss

به نمایش برچسب صحیح یعنی c_{y_i} را اختیار می‌کند. تابع هزینه‌ی پیشنهادی برای یادگرفتن نگاشتی با خواص فوق به این صورت تعریف شده است:

$$W = \arg\min_{W} \lambda \|W\|_{Fro}^{۲} + \sum_{n=۱}^{N_u} \alpha \mathcal{L}_\epsilon(\mathbf{x_i}, \mathbf{c_{y_i}}) + (۱ - \alpha)\mathcal{M}(\mathbf{x_i}, \mathbf{c_{y_i}}). \qquad (۵۲\text{-}۲)$$

در نهایت در این روش با جایگزین کردن c با cV در تابع هزینه‌ی فوق، نگاشت V روی توصیف‌ها نیز یاد گرفته می‌شود تا نمایش کلمات که با استفاده از مجموعه متن بدون برچسب بدست آمده، با توجه به برچسب‌های موجود در مسئله تنظیم دقیق شود.

۲-۱۰ جمع‌بندی

در پایان این فصل به یک مقایسه کلی از روش‌های پیشین و مزایا و معایب آن‌ها می‌پردازیم که در جدول ۲-۱ آمده است.

جدول ۲-۱: مقایسه مهم‌ترین روش‌های ارائه شده برای یادگیری از صفر

نام روش	سال ارائه	نوع توصیف	مزایا و معایب
DAP [۱۳]	۲۰۰۹	بردار صفت	+ارائه یک چارچوب نظام‌مند + امکان تعویض برخی قسمت‌ها مانند نوع دسته‌بند مورد استفاده ـ مدل نکردن ارتباط میان صفت‌ها ـ در نظر نگرفتن خطای دسته‌بندی در آموزش
طراحی صفت برای دسته‌ها [۱۲]	۲۰۱۳	شباهت دسته‌ها با هم	+ عدم نیاز به توصیف صریح دسته‌ها + ارائه یک کران نظری برای خطای دسته‌بندی + امکان استفاده در یادگیری با نظارت یا صفرضرب ـ عدم امکان استفاده از توصیف‌های دقیق‌تر و بسنده کردن به شباهت میان دسته‌ها

جدول ۲-۱: مقایسه مهم‌ترین روش‌های ارائه شده برای یادگیری از صفر

نام روش	سال ارائه	نوع توصیف	مزایا و معایب
دسته‌بند نوشتاری [۱۰]	۲۰۱۳	متن	+ معرفی مسئله استفاده از توصیف متنی و جمع‌آوری مجموعه دادگان لازم + استفاده از روش‌های تطبیق دامنه + امکان یادگیری دسته‌بند برای هر کلاس دیده نشده‌ی جدید ـ سادگی مدل تحلیل متن ـ محدود بودن به نگاشت‌های خطی
DeViSE [۲۴]	۲۰۱۳	نام دسته‌ها	+ عدم نیاز به تهیه توصیف توسط انسان + بهره‌گیری از پیش‌آموزش روی داده‌های فراوان ـ عدم دسته‌بندی دقیق برای دسته‌های نزدیک به هم
نگاشت القایی چند منظری[۵۰] [۳۸]	۲۰۱۴	بردار صفت و نام دسته‌ها	+ معرفی مشکل جابجایی دامنه در یادگیری صفرضرب و ارائه یک راه‌حل برای آن + ارائه یک روش انتشار برچسب برای دسته‌بندی در مقابل نزدیک‌ترین همسایه + استفاده از چند توصیف به صورت همزمان ـ نیاز به داده‌های آزمون در زمان آموزش
یادگیری صفرضرب با صفت‌های غیرقطعی [۴۳]	۲۰۱۴	بردار صفت	+ در نظر گرفتن عدم قطعیت پیش‌بینی صفت در داده‌های آزمون + تعمیم به مسئله یادگیری تک‌ضرب ـ در نظر نگرفتن روابط بین صفت‌ها
COSTA [۲۳]	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 صفت‌های غیر دودویی

[۵۰] Transductive Mult-View Embedding

جدول ۲-۱: مقایسه مهم‌ترین روش‌های ارائه شده برای یادگیری از صفر

نام روش	سال ارائه	نوع توصیف	مزایا و معایب
ConSE [۱۱]	۲۰۱۴	نام دسته‌ها	+ عدم نیاز به تهیه توصیف توسط انسان + بهره‌گیری از پیش‌آموزش با داده‌های بدون برچسب فراوان + عدم وجود فاز آموزش مخصوص به مسئله + امکان تشخیص برای هر دسته‌ی جدید ـ عدم دسته‌بندی دقیق برای دسته‌های نزدیک به هم
ESZSL [۱۸]	۲۰۱۵	بردار صفت	+ درنظرگرفتن خطای دسته‌بند در آموزش + دارای جواب بسته و پیاده‌سازی یک خطی + سرعت آموزش و آزمون بالا ـ محدود بودن رابطه به روابط خطی ـ عمل‌کرد ضعیف برای ویژگی‌های تصویر با ابعاد بالا
SSE [۳۶]	۲۰۱۵	بردار صفت	+ امکان طبیعی استفاده از صفت‌ها 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
SJE [۲۷]	۲۰۱۵	بردار صفت یا نام دسته‌ها	+ ارائه یک چارچوب کلی برای نگاشت به یک فضای مشترک + ارائه یک روش برای نگاشت نام دسته‌ها + امکان طبیعی استفاده از صفت‌ها با مقدار حقیقی ـ محدود بودن به نگاشت‌های دوخطی
یادگیری از صفر نیمه‌نظارتی با یادگیری نمایش برچسب‌ها [۴۴]	۲۰۱۵	بردار صفت 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

جدول ۲-۱: مقایسه مهم‌ترین روش‌های ارائه شده برای یادگیری از صفر

نام روش	سال ارائه	نوع توصیف	مزایا و معایب
یادگیری صفرضرب با دسته‌بند حداکثر حاشیه [۴۰]	۲۰۱۵	بردار صفت	+ پیش‌بینی مستقیم برچسب‌های نهایی + صورت‌بندی نیمه‌نظارتی ـ مسئله بهینه‌سازی سنگین ـ عدم استفاده از ویژگی‌های فضای تصاویر آزمون
تطبیق دامنه بدون نظارت برای یادگیری صفرضرب [۴۱]	۲۰۱۵	بردار صفت یا نام دسته‌ها	+ صورت‌بندی مسئله به صورت یک مسئله تطبیق دامنه بدون نظارت + استفاده از اطلاعات بدون‌نظارت موجود در داده‌های آزمون + مسئله بهینه‌سازی سبک ـ نیاز به یک پیش‌بینی اولیه از یک روش دیگر به عنوان ورودی
پیش‌بینی دسته‌بند از متن توصیفی [۱]	۲۰۱۵	متن	+ معرفی دسته‌بند پیچشی + صورت‌بندی مسئله با شبکه‌های عصبی ـ استخراج ویژگی‌های نه چندان خوب از متن ـ تعداد پارامترهای زیاد مدل
تشخیص هم‌دسته بودن توصیف و تصویر [۳۷]	۲۰۱۶	بردار صفت	+ امکان طبیعی استفاده از انواع صفت‌های پیوسته + پارامترهای مستقل از تعداد دسته‌ها ـ استنتاج سنگین که تخمین زده شده‌است
SS-VOC [۴۲]	۲۰۱۶	نام دسته‌ها	+ در نظرنگرفتن فرض محدود کننده جدا بودن دسته‌های آزمون و آموزش + استفاده از کلمات لغت‌نامه برای نیمه‌نظارتی کردن روش + کارکرد روش در مسائل یادگیری عادی، صفرضرب و مجموعه باز + توانایی اجرا زمانی که دسته‌های آزمون بسیار زیاد هستند ـ عدم امکان استفاده از اطلاعات نظارتی قوی‌تر مثل بردار صفت‌ها

جدول ۲-۱: مقایسه مهم‌ترین روش‌های ارائه شده برای یادگیری از صفر

نام روش	سال ارائه	نوع توصیف	مزایا و معایب
یادگیری ژرف بازنمایی توصیف‌های متنی [۳۴]	۲۰۱۶	متن	+ جمع‌آوری مجموعه دادگان متنی بزرگ + استفاده از شبکه‌های عصبی بازگشتی[۵۱] برای تحلیل متن + ارائه یک فورمول‌بندی جامع بر اساس شبکه‌های عصبی با قابلیت یادگیری توامان تمام قسمت‌ها ـ عدم ارائه راه‌کار برای انتخاب معماری مدل متنی
یادگیری صفرضرب از متون آنلاین با حذف نویز [۳۳]	۲۰۱۶	متن	+ الگوریتم یادگیری آسان + تشخیص ابعاد مهم نمایش متنی و کلمات مهم برای هر دسته ـ استفاده از مدل محدود خطی برای تحلیل متن
یادگیری صفرضرب با چند راهنما [۳۱]	۲۰۱۶	توصیف‌های گوناگون	+ استفاده از سطح دقیق‌تری برای تناظر میان تصویر و توصیف + امکان استفاده از توصیف‌های متنی که بدون نظارت بدست می‌آیند + امکان استفاده همزمان از توصیف‌های مختلف ـ نیاز به اطلاعات نظارتی بیشتر در تصاویر برای تعیین قسمت‌های مختلف ـ مسئله بهینه‌سازی با محدودیت‌های زیاد و سنگین
LatEm [۳۰]	۲۰۱۶	توصیف‌های گوناگون	+ عدم محدودیت به نگاشت‌های خطی و در نظر گرفتن نگاشت‌های غیرخطی به صورت تکه‌تکه دوخطی + امکان استفاده همزمان از توصیف‌های مختلف

۵۱

فصل ۳

روش پیشنهادی

در این فصل به بیان روش‌های پیشنهادی در این پژوهش برای مسئله یادگیری صفرضرب می‌پردازیم. روش‌های مطرح شده در این فصل از دو رویکرد متفاوت برای حل مسئله یادگیری صفرضرب استفاده می‌کنند. یک رویکرد یافتن نگاشت از فضای تصاویر به فضای توصیف دسته‌هاست که این نگاشت با استفاده از شبکه‌های ژرف مدل شده است. رویکرد دوم با انجام یک خوشه‌بندی در فضای ویژگی‌های ژرف استخراج شده از تصاویر و با یادگرفتن نگاشتی از فضای توصیف دسته‌ها به فضای ویژگی‌های ژرف تصاویر همراه است.

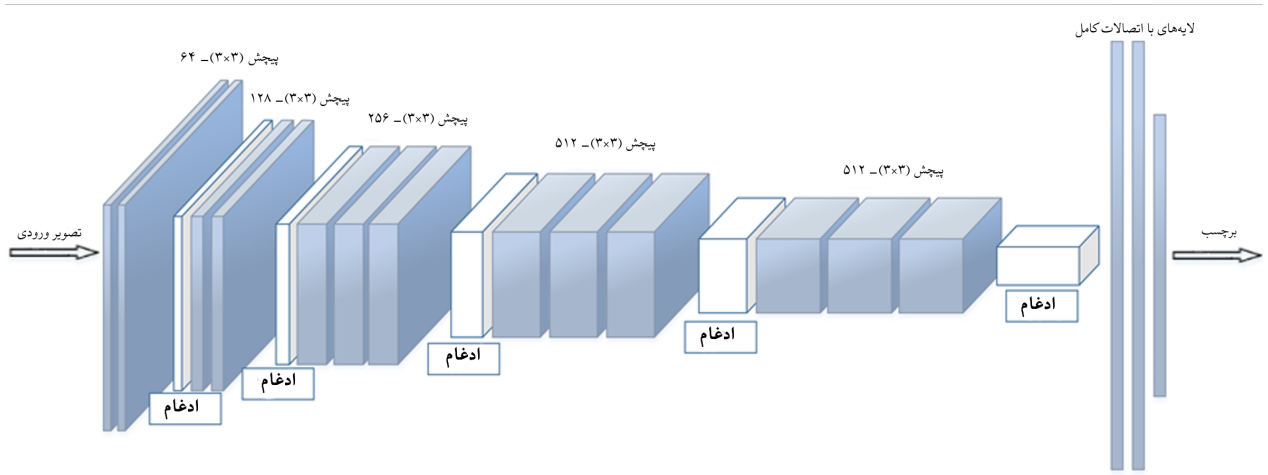
در ابتدای این بخش به مسئله استخراج ویژگی از تصاویر با استفاده از شبکه‌های ژرف می‌پردازیم، فضای تشکیل شده از ویژگی‌های تصاویر هنگام استفاده از این شبکه‌ها، دارای خاصیت جدایی پذیری دسته‌های مختلف از هم و تشکیل خوشه‌هایی از نمونه‌های هر دسته است؛ فرض وجود چنین خاصیت‌هایی در فضای ویژگی‌های تصاویر، اساس روش‌های ارائه شده در این فصل است. در بخش ۳-۲ یک شبکه‌ی عصبی چندوظیفه‌ای برای پیش‌بینی ویژگی از تصاویر معرفی می‌کنیم که با در نظر گرفتن نمونه‌های آزمون در زمان آموزش می‌تواند مشکل جابجایی دامنه را کاهش دهد. در بخش ۳-۵ یک تابع مطابقت نوین برای مسئله دسته‌بندی صفرضرب معرفی می‌کنیم که استفاده از اطلاعات غیرنظارتی موجود در ساختار نمونه‌های دسته‌های دیده نشده را ممکن می‌سازد. این تابع مطابقت از یک خوشه‌بندی روی نمونه‌های آزمون بهره می‌برد که با توجه به استخراج ویژگی‌ها با استفاده از شبکه‌های عصبی ژرف و جداسازی مناسب در فضای این ویژگی‌ها، از دقت مناسبی برخوردار است. این تابع مطابقت به نمونه‌هایی که در یک خوشه قرار دارند برچسب یکسانی نسبت می‌دهد. با توجه به استفاده از خوشه‌بندی در این تابع مطابقت، یک روش خوشه‌بندی نیمه‌نظارتی که منطبق بر فرضیات مسئله

یادگیری صفرضرب است ارائه می‌گردد و سپس یک روش دسته‌بندی با استفاده از تابع مطابقت و خوشه‌بندی ارائه شده و یادگیری نگاشتی خطی از توصیف دسته‌ها به فضای تصاویر، تدوین می‌گردد. هرچند که عملکرد این روش ارائه شده برتر از روش‌های پیشگام موجود است ولی محدودیت‌هایی نیز دارد که ناشی از جدا بودن مرحله خوشه‌بندی و نگاشت به فضای مشترک است؛ برای رفع این محدودیت‌ها روش دیگری معرفی می‌شود که خوشه‌بندی و یادگیری نگاشت در آن به صورت توام انجام می‌شود. این یادگیری توام باعث بهبود دقت دسته‌بندی نسبت به روش پیشنهادی قبلی می‌شود.

نمادگذاری مورد استفاده در این فصل سازگار با نمادگذاری معرفی شده در بخش ۲ است که در جدول ۳-۱ برای مراجعه سریع خلاصه شده است.

جدول ۳-۱: معرفی نمادهای مورد استفاده

نماد	شرح
$\mathcal{S}(\mathcal{U})$	مجموعه دسته‌های دیده‌شده (دیده‌نشده)
$n_s(n_u)$	تعداد دسته‌های دیده‌شده (دیده‌نشده)
$N_s(N_u)$	تعداد نمونه‌های آموزش (آزمون)
$X_s(X_u)$	ماتریس نمونه‌های آموزش (آزمون)
$Y_s(Y_u)$	برچسب‌های نمونه‌های آموزش (آزمون)
$C_s(C_u)$	ماتریس توصیف‌های دسته‌های دیده‌شده (دیده‌نشده)
$\mathbf{x_i} \in \mathbb{R}^d$	بردار ویژگی‌های تصویر $i-$م
$\mathbf{c_y} \in \mathbb{R}^a$	بردار توصیف دسته‌ی y
$X_{(i)}$	سطر $i-$م ماتریس X
$\|X\|_{Fro}$	نرم فروبنیوس ماتریس X
$diag(\mathbf{x})$	یک ماتریس قطری که بردار $\mathbf{x}$ روی قطر اصلی آن قرار داده شده
$\mathbf{1}$	یک بردار که تمام عناصر آن برابر یک است
$\mathbf{1}_k$	یک بردار که درایه‌ی $k-$م آن یک و سایر عناصرش صفر است



شکل ۳-۱: ساختار شبکه vgg که در آن لایه‌های سفید مراحل ادغام که اینجا انتخاب بیشینه در پنجره‌های ۲ × ۲ است را نشان می‌دهند. لایه‌های پیچشی با مکعب‌های آبی مشخص شده‌اند که عرض آن‌ها متناسب با تعداد کانال‌های موجود در آن لایه است [۴۶].

۳-۱ استخراج ویژگی با شبکه‌های عصبی ژرف

در سال‌های اخیر استفاده از شبکه‌های عصبی پیچشی ژرف کاراترین روش برای استخراج ویژگی از تصاویر بوده است [۴۵]. این روش که در آن نحوه‌ی استخراج ویژگی با استفاده از تعداد زیادی داده‌ی برچسب‌دار *یاد گرفته می‌شود*، جایگزین روش‌های قبلی مانند SIFT و HOG شده است که در آن‌ها، نحوه‌ی استخراج ویژگی توسط یک خبره تعیین شده و همواره ثابت است. در این شبکه‌ها در هر لایه عموما از چندین صافی استفاده می‌شود. تعداد کم پارامترهای فیلتر و استقلال آن از اندازه تصویر ورودی، باعث شده تعداد پارامترهای موجود در یک لایه‌ی پیچشی بسیار کمتر از یک لایه با اتصالات کامل[۱] باشد و در نتیجه امکان افزایش عمق شبکه بیشتر باشد. معماری مورد استفاده در روش‌های این فصل برای استخراج ویژگی، مبتنی بر معماری ۱۹ لایه شبکه vgg [۱۴] است (شکل ۳-۱). در این شبکه از ۱۶ لایه‌ی پیچشی استفاده شده است. ساختار هر لایه به این صورت است که تعدادی کانال از ویژگی‌ها (در لایه‌ی اول خود تصویر) به عنوان ورودی وارد لایه می‌شوند و با استفاده از تعدادی صافی با اندازه ۳ × ۳ به ویژگی‌های خروجی تبدیل می‌شوند. تعداد کانال‌های ورودی در لایه‌ی اول سه کانال رنگی RGB است و در لایه‌های بعدی تعداد صافی‌ها به گونه‌ای تعیین شده که تعداد کانال‌های ویژگی‌ها برابر: ۶۴ در لایه‌ی اول و دوم، ۱۲۸ در لایه سوم و چهارم، ۲۵۶ در لایه پنجم تا هشتم و ۵۱۲

[۱] fully connected layer

در لایه نهم تا شانزدهم است. تابع فعال‌سازی مورد استفاده در لایه‌های پیچشی تابع ReLU[۲] است که ضابطه آن به این صورت است:

$$ReLU(\mathbf{x}) = max(۰, \mathbf{x}). \quad (۱-۳)$$

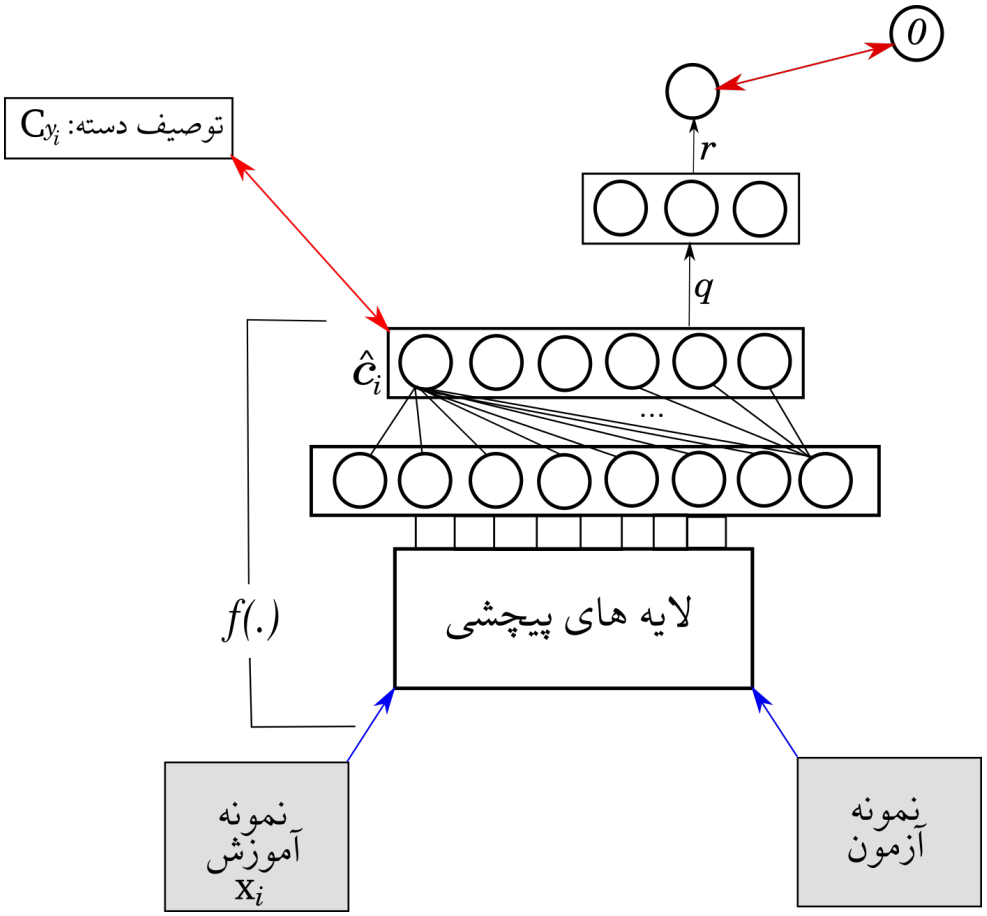
برای کاهش اندازه ماتریس ویژگی‌ها، میان برخی لایه‌های پیچشی از یک تابع ادغام[۳] استفاده می‌شود. تابع ادغام مورد استفاده در این شبکه تابع ادغام بیشینه است یعنی در ماتریس ویژگی حاصل یک پنجره‌ی ۲ × ۲ حرکت داده می‌شود و تنها بزرگترین مقدار میان چهار مقداری پنجره بر آن‌ها منطبق شده به خروجی منتقل می‌شود. بعد از ۱۶ لایه پیچشی سه لایه با اتصالات کامل وجود دارد. ما برای استخراج ویژگی از خروجی لایه‌ی هفدهم یعنی نخستین لایه با اتصالات کامل استفاده می‌کنیم و دو لایه‌ی نهایی کنار گذاشته می‌شوند. ورودی این لایه به این صورت به دست می‌آید که تمام ماتریس‌های ویژگی لایه‌ی شانزدهم به صورت بردارهای یک بعدی در آمده و در کنار هم قرار می‌گیرند، سپس به صورت یک برادر ۲۵۰۸۸ – بعدی وارد لایه‌ی هفدهم شده و در این لایه با استفاده از یک نگاشت خطی و تابع فعال‌سازی ReLU به بردارهای ویژگی ۴۰۹۶ – بعدی تبدیل می‌شود. در شبکه اصلی این خروجی این لایه به یک لایه‌ی مشابه خود و در نهایت با یک لایه با اتصالات کامل که خروجی آن به اندازه تعداد دسته‌هاست با تابع فعال‌سازی softmax به پیش‌بینی برچسب تبدیل می‌شود.

۳-۲ یک شبکه‌عصبی چندوظیفه‌ای

یادگیری نگاشت‌ها با استفاده از داده‌های دسته‌های دیده‌شده، همان‌طور که در بخش ۲-۹ اشاره شد، دچار مشکل جابجایی دامنه است و روی داده‌های دسته‌های دیده‌نشده به خوبی قابل تعمیم نیست. یک راه حل برای مقابله با این مشکل این است که در حین یادگیری نگاشت اجبار شود که حاصل نگاشت یک نمونه‌ی آزمون به نوعی نزدیک به نگاشت توصیف یکی از دسته‌های آزمون باشد. همان طور که در بخش ۲-۹ بیان شد، چنین راه‌حلی در [۴۱] استفاده شده است. معیار نزدیکی نگاشت‌ها در آن روش یک امتیاز پیشین از شباهت هر نمونه‌ی آزمون با دسته‌های دیده نشده است که توسط یک روش دیگر استخراج شده می‌شود. یعنی ابتدا یک روش دسته‌بندی احتمالی که در آن پژوهش روش IAP [۱۳] برای این کار انتخاب شده بود، به صورت مستقل روی مجموعه دادگان اجرا شده و احتمال‌هایی که برای انتساب هر نمونه به

[۲] Rectified Linear Unit

[۳] Pooling



شکل ۳-۲: ساختار شبکه چند وظیفه‌ای پیشنهادی. فلش‌های آبی رنگ ورودی‌های شبکه را نشان می‌دهند و فلش‌های قرمز رنگ مقایسه خروجی شبکه با خروجی مورد انتظار را. خطوط سیاه‌رنگ اتصالات شبکه را نشان می‌دهند. زیر شبکه‌ی برگرفته شده از شبکه vgg و یک لایه‌ی با اتصالات چگال اضافه شده بین دو دو ورودی مشترک هستند. لایه‌های r و q مخصوص نمونه‌های آزمون هستند. خروجی لایه‌ی r همواره با مقدار صفر مقایسه می‌شود.

دسته‌های آزمون از آن روش بدست می‌آید بعنوان وزن‌های شباهت در نظر گرفته می‌شود و فاصله هر توصیف پیش‌بینی شده برای هر نمونه با توصیف دسته‌های آزمون متناسب با این وزن‌های شباهت جریمه می‌شود. ما در این بخش یک روش مبتنی بر شبکه‌های عصبی ژرف معرفی می‌کنیم که در آن نگاشتی غیرخطی و چندلایه از تصاویر به بردارهای صفت یادگرفته می‌شود. معیار یادگیری این نگاشت، پیش‌بینی صحیح صفت برای نمونه‌های آموزش (که بردار صفت صحیح برای آن‌ها مشخص است) و هم‌چنین نزدیک بودن حاصل نگاشت هر نمونه‌ی آزمون به توصیف یکی از دسته‌های دیده نشده است. برای مدل کردن این نگاشت، از یک شبکه‌ی عصبی استفاده شده است. اگر نگاشت مدل شده با شبکه عصبی را با f نشان دهیم، آنگاه $\hat{\mathbf{c}}_i = f(\mathbf{x}_i)$ نشان‌دهنده‌ی بردار توصیف پیش‌بینی شده برای نمونه‌ی $i-$م است و تابع هزینه‌ی مورد

استفاده برای آموزش شبکه به صورت زیر تعریف می‌شود:

$$\min_{f} \frac{۱}{N_s} \sum_{i=۱}^{N_s} loss(\hat{\mathbf{c}}_i), \mathbf{c}_{y_i}) + \frac{\gamma}{N_u} \sum_{i=N_s}^{N_s+N_u} \Big(\min_{j=n_s,\ldots,n_s+n_u} \|\hat{\mathbf{c}}_\mathbf{i} - \mathbf{c}_\mathbf{j}\|_۲^۲ \Big), \qquad (۳\text{-}۲)$$

که γ یک پارامتر است. جمله‌ی اول، جمله‌ی مربوط به خطای پیش‌بینی صفت‌هاست و تفاوت میان صفات پیش‌بینی شده توسط شبکه و صفات صحیح را برای نمونه‌های آموزش جریمه می‌کند. جمله‌ی دوم برای رفع مشکل جابجایی دامنه طراحی شده است و تحمیل می‌کند که حاصل نگاشت یک نمونه‌ی آزمون حتما نزدیک توصیف یکی از دسته‌های دیده‌نشده باشد، این دسته‌ی دیده نشده، دسته‌ای در نظر گرفته شده است که توصیف آن با نگاشت کمترین فاصله را دارد. این قسمت از رابطه فوق را می‌توان به صورت شهودی این گونه توضیح داد که در غیاب جمله‌ی دوم برای هر نمونه یک بردار توصیف پیش‌بینی می‌شد و سپس نزدیک‌ترین بردار توصیف از میان توصیف دسته‌های آزمون به عنوان توصیف صحیح در نظر گرفته شده و برچسب بر اساس آن پیش‌بینی می‌شد. حال جمله‌ی دوم رابطه (۳-۲) جریمه‌ای به میزان فاصله‌ی توصیف پیش‌بینی شده برای هر نمونه با بردار توصیف همان دسته‌ای که به آن نزدیک‌تر است، در نظر می‌گیرد. حال اگر این فرض صحیح باشد که حاصل نگاشت در اکثر موارد به توصیف صحیح نزدیکتر است، یا به عبارتی در اکثر مواقع استفاده از دسته‌بند نزدیکترین همسایه روی نگاشتی که تنها با جمله‌ی اول آموزش دیده، دقتی بیش از ٪۵۰ داشته باشد، وجود چنین جمله‌ای باعث می‌شود که مواردی که قبلا درست تشخیص داده می‌شدند حالا با دقت بیشتر (فاصله کمتر از بردار توصیف دسته‌ی مورد نظر) باز هم درست پیش‌بینی شوند. با توجه به افزایش دقت نگاشت روی این نمونه‌ها، انتظار می‌رود برای برخی نمونه‌هایی که در حالت قبل پیش‌بینی نادرست به آن‌ها تعلق می‌گرفت نیز با این نگاشت بهبود یافته، پیش‌بینی صحیح برای آن‌ها صورت بگیرد.

تابع $loss(\cdot,\cdot)$ در معادله (۳-۲) در مجموعه دادگانی که صفات دودویی هستند تابع آنتروپی متقاطع[۴] در نظر گرفته شده است یعنی:

$$loss(y,z) = z\log(۱-y) + (۱-z)\log(y). \qquad (۳\text{-}۳)$$

برای مجموعه دادگانی که مقادیر بردارهای توصیف در آن‌ها مقادیر دلخواه حقیقی است تابع هزینه مربع اختلاف در نظر گرفته شده است:

$$loss(y,z) = \|y-z\|_۲^۲. \qquad (۴\text{-}۳)$$

[۴] Cross Entropy

۳-۲-۱ بهینه‌سازی

تابع کمینه به کار برده شده در جمله‌دوم معادله (۳-۲) در برخی نقاط مشتق‌پذیر نیست، اما با توجه به اینکه اندازه‌ی این نقاط صفر است تابع تقریبا همه‌جا مشتق‌پذیر است و آموزش شبکه با استفاده از پس‌انتشار[۵] مقدار گرادیان ممکن خواهد بود. به صورت دقیق‌تر، بهینه‌سازی رابطه (۳-۲) عملیات محاسبه‌ی مقدار کمینه را داخل شبکه تعبیه می‌کنیم (شکل ۳-۲)؛ به این صورت که لایه‌های جدید q وr برای نمونه‌های دیده نشده اضافه می‌شود که:

$$(q(\mathbf{v}))_j = \|f(\mathbf{v}) - \mathbf{c_j}\|_۲^۲, \tag{۳-۵}$$

$$r(\mathbf{z}) = \min_{j=۱\ldots n_u} (\mathbf{z})_j. \tag{۳-۶}$$

در رابطه (۳-۵)، لایه‌ی q یک بردار توصیف پیش‌بینی شده را به عنوان ورودی دریافت کرده است و خروجی آن برداری است که تعداد ابعادش برابر تعداد دسته‌های دیده‌نشده است و مقدار هر بعد آن برابر فاصله‌ی بردار v با بردار توصیف (امضای) یک دسته‌ی دیده‌نشده‌است. سپس خروجی این لایه به لایه‌ی r وارد می‌شود و در این لایه کوچکترین مقدار این بردار انتخاب می‌شود. نتیجتاً ترکیب این دولایه کمینه‌ی فاصله‌ی v با امضاهای دسته‌های دیده‌نشده را تولید خواهد کرد که برابر جمله‌ی دوم در رابطه (۳-۲) خواهد بود.

در هنگام آموزش با پس‌انتشار، مشتق تابع هزینه‌ی l نسبت به هر ورودی مثل z در لایه‌ی r با ضابطه‌ی زیر محاسبه می‌شود:

$$\frac{\partial l}{\partial z} = \sum_j \mathbb{1}[(z)_j = \min(z)] \frac{\partial l}{(z)_j}. \tag{۳-۷}$$

پس از آموزش شبکه، در فاز آزمون لایه‌های q و r حذف شده و بردار توصیف برای تصاویر آزمون با استفاده از شبکه پیش‌بینی می‌شود، در نهایت دسته‌بندی با استفاده از دسته‌بند نزدیک‌ترین همسایه روی نمونه‌های آزمون انجام خواهد شد.

مراحل آموزش شبکه در الگوریتم ۱ آورده شده است.

۳-۲-۲ معماری شبکه

ما از قسمتی از شبکه‌ی ۱۹ لایه‌ی vgg [۱۴] که شامل ۱۶ لایه‌ی پیچشی ابتدا و لایه‌اول با اتصالات چگال است به عنوان یک زیر شبکه در ورودی شبکه خود استفاده می‌کنیم. همان‌طور که در بخش ۳-۱ شرح داده شد، با این زیر شبکه تصاویر

[۵] Back Propagation

الگوریتم ۱ الگوریتم آموزش و آزمون شبکه عصبی پیشنهادی

۱ **ورودی:** تصاویر و توصیف‌های آموزش و آزمون و برچسب‌های نمونه‌های آموزش.

۲ **خروجی:** برچسب‌های پیش‌بینی شده برای نمونه‌های آزمون.

۳ پیش آموزش شبکه تنها با نمونه‌های آموزش و مقایسه خروجی با توصیف صحیح.

۴ آموزش کامل شبکه با داده‌های آموزش و آزمون.

۵ حذف لایه‌های r و q.

۶ خروجی شبکه را به ازای X_u در P_u بریز.

۷ دسته‌بند نزدیک‌ترین همسایه NN را با بردارهای توصیف دسته‌های آزمون بساز

۸ عناصر P_u را با استفاده از NN دسته‌بندی کن.

۹ حاصل مرحله قبل را به عنوان پیش‌بینی نهایی برگردان.

ورودی به بردارهای ۴۰۹۶−بعدی نگاشته می‌شوند. سپس یک لایه‌ی با اتصالات چگال قرار دارد که این حاصل را به بردارهای توصیف دسته‌ها می‌نگارد. برای نمونه‌های آموزش، خروجی این لایه با بردار توصیف صحیح مقایسه می‌شود. برای نمونه‌های آزمون خروجی این لایه به لایه‌های q وr متصل می‌شود و مقدار خروجی r با مقدار مطلوبش که صفر است مقایسه خواهد شد.

تابع فعال‌سازی در همه‌ی لایه‌ها تابع ReLU است؛ با این استثنا که برای مجموعه دادگانی که بردار توصیف دودویی دارند، در لایه‌ی آخر از تابع سیگموید با ضابطه

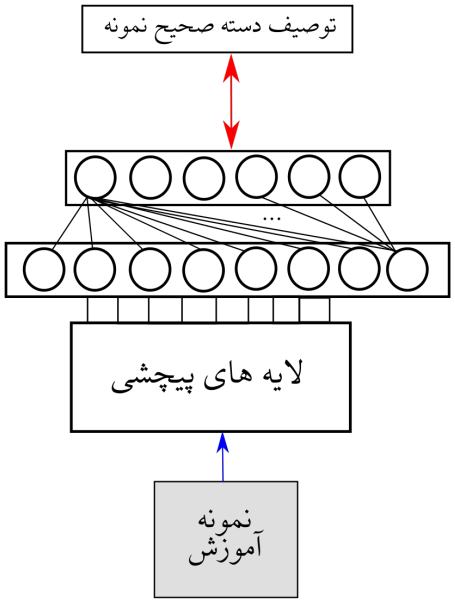
$$\sigma(x) = \frac{۱}{۱ + e^{-x}}, \tag{۸-۳}$$

بعنوان تابع فعال‌سازی استفاده شده است تا مقادیر در بازه‌ی $[۰, ۱]$ نگاشته شوند.

۳-۲-۳ یک مدل پایه برای مقایسه

برای روشن شدن تاثیر استفاده از اطلاعات بدون نظارت نمونه‌های آزمون در یادگیری بهتر نگاشت، قصد داریم در فصل آتی مدل ارائه شده را با یک مدل ساده برای پیش‌بینی صفت مقایسه کنیم که در این‌جا این مدل پایه را معرفی می‌کنیم. در این مدل ساده تنها از لایه‌های با اتصالات کامل بعد از استخراج ویژگی با لایه‌های پیچشی، برای پیش‌بینی صفت استفاده

می‌کنیم. ساختار این مدل در تصویر ۳-۳ نمایش داده شده است. در این شبکه از یک یا چند لایه با اتصالات کامل بعد از لایه‌های پیچشی استفاده می‌شود. مشابه حالت قبل تابع فعال‌سازی برای مجموعه دادگانی که مقادیر توصیف دسته‌هایشان دودویی است تابع سیگموید، و برای مجموعه دادگانی که مقادیر بردارهای توصیف در آن‌ها مقادیر دلخواه حقیقی است تابع ReLU استفاده شده است. ابعاد لایه‌های با اتصالات کامل پایانی الزاما برابر تعداد ابعاد بردارهای توصیف است و برای سایر لایه با اتصالات کامل نیز همین تعداد ابعاد انتخاب شده است. مقایسه نتایج دقت دسته‌بندی بین مدل قبلی و این مدل در بخش ۴-۴ نشان‌دهنده‌ی تاثیر مثبت استفاده از اطلاعات بدون نظارت موجود در نمونه‌های آزمون است که باعث بهبود حداقل ۱۰ درصدی دقت دسته‌بندی می‌شود.



شکل ۳-۳: ساختار شبکه پایه. فلش آبی رنگ ورودی‌های شبکه را نشان می‌دهند و فلش‌های قرمز رنگ مقایسه خروجی شبکه با خروجی مورد انتظار را.

۳-۳ نگاشت به هیستوگرام دسته‌های دیده‌شده با شبکه عصبی

با توجه به عمل‌کرد خوب نمایش تصاویر و توصیف دسته‌های آزمون به صورت هیستوگرام دسته‌های دیده شده، فضای میانی برای نمایش تصاویر و توصیف‌ها در مسئله یادگیری صفر ضرب در اخیرترین روش‌های یادگیری صفرضرب [۳۶]، در این بخش روشی برای استفاده از این فضای میانی معرفی می‌کنیم. این روش می‌تواند نتایج بهتری نسبت به حالتی که از

فضای توصیف‌ها به عنوان فضای مشترک استفاده شده و پیش‌بینی صفت از تصاویر صورت می‌گیرد، کسب نماید. روش پیشنهادی برای نگاشت تصویر به یک هیستوگرام از دسته‌های دیده‌شده، مبتنی بر دسته‌بندی با شبکه‌های عصبی است. پراستفاده‌ترین روش دسته‌بندی چند دسته‌ای با شبکه‌های عصبی، بهره‌گیری از یک لایه با تابع فعال‌سازی softmax با اندازه تعداد دسته‌ها، به عنوان لایه‌ی آخر شبکه است. ضابطه این تابع را که در معادله (۲-۸) ذکر شد در این‌جا برای پیگیری بهتر بحث تکرار می‌کنیم. اگر مقادیر لایه‌ی آخر شبکه را با z نمایش دهیم، با اعمال این تابع فعال‌سازی روی این لایه، عنصر $j-$م آن به این صورت تغییر می‌کنید.

$$softmax(z)_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \quad j = ۱, \ldots, n_s.$$

با دقت در ضابطه این تابع مشاهده می‌شود که این تابع نسبت هر عنصر را به جمع سایر عناصر حساب می‌کند که به تعبیری برابر است میزان وزنی که عنصر $j-$م نسبت به کل وزن‌های موجود در لایه کسب کرده است. برای پررنگ‌تر شدن تفاوت، به جای محاسبه‌ی این نسبت میان خود عناصر از یک تابع نمایی برحسب آن‌ها استفاده شده‌است. اندازه این لایه در شبکه‌های عصبی برابر تعداد دسته‌هایی که علاقمند به دسته‌بندی در آن‌ها هستیم در نظر گرفته می‌شود و هر گره[۶] از آن متناظر با یکی از دسته‌های دیده نشده است. در خروجی این لایه، اگر $z)_j$ بیشینه به میزان کافی با سایر مقادیر z تفاوت داشته باشد، مقدار تابع به ازای $z)_j$ بیشینه نزدیک به یک خواهد بود و برای سایر عناصر z نزدیک به صفر یعنی با استفاده از این تابع فعال‌سازی، خروجی این لایه می‌تواند نمایش یکی‌یک برچسب را تولید کند. به همین علت در هنگام آموزش شبکه از تابع هزینه‌ی آنتروپی متقاطع میان z و نمایش یکی‌یک برچسب صحیح استفاده می‌شود.

از طرفی به علت عمل میانگین‌گیری، مقادیر این تابع روی یک سادک قرار می‌گیرند یعنی به عبارت دقیق‌تر داریم:

$$\forall j, \quad softmax(z_j) \geqslant ۰, \tag{۳-۹}$$

$$\sum_j softmax(z_j) = ۱. \tag{۳-۱۰}$$

در نتیجه می‌توان از خروجی این لایه به عنوان برداری از احتمال تعلق نمونه‌ی ورودی به هر دسته یا به عبارت دیگر هیستوگرام دسته‌ها تعبیر کرد. ما از این خاصیت برای نگاشت تصاویر به هیستوگرام دسته‌های دیده‌شده در یادگیری صفرضرب استفاده می‌کنیم. در روش پیشنهادی یک شبکه عصبی عمیق که برای دسته‌بندی در دسته‌های دیده شده می‌سازیم و با استفاده از نمونه‌های دسته‌های دیده‌شده که همگی دارای برچسب هستند آن را آموزش می‌دهیم. در نتیجه این شبکه برای هر تصویر

[۶] node

ورودی (اعم از تصاویر دسته‌های دیده‌شده یا دیده‌نشده) یک بردار از امتیاز شباهت آن به هر دسته‌ی دیده‌شده یا به عبارتی هیستوگرامی از دسته‌های دیده شده تولید می‌کند.

همان‌طور که گفته شد تابع فعال‌سازی softmax طوری طراحی شده که تفاوت میان مقادیر هر گره را بزرگنمایی کرده و خروجی آن نزدیک به کدگذاری یکی‌یک بردار برچسب باشد، این مسئله می‌تواند باعث از بین رفتن اطلاعات شباهت نمونه به دسته‌هایی شود که در رتبه‌های بعد از دسته‌ای که امتیاز بیشینه را کسب کرده قرار دارند [۴۷]. برای حل این معضل یعنی افزایش کیفیت هیستوگرام بدست آمده و دور کردن خروجی از کدگذاری یکی‌یک، از یک نسخه تغییر یافته از تابع softmax استفاده می‌کنیم:

$$softmax_T((z)_j) = \frac{exp((z)_j/T)}{\sum_i exp((z)_i/T)}. \qquad (۱۱-۳)$$

با ازدیاد پارامتر T در رابطه (۳-۱۱) باعث تفاوت کمتر مقدار خروجی تابع به ازای $(z)_j$ بیشینه با سایر مقادیر شده و خروجی هموارتری نسبت ضابطه معمول که در که در آن $T = ۱$ است می‌شود. ما در زمان آموزش شبکه، به علت این که خروجی با کدگذاری یکی‌یک برچسب صحیح مقایسه می‌شود، از مقدار $T = ۱$ استفاده می‌کنیم. اما برای بدست آوردن نمایش تصاویر آزمون در فضای هیستوگرام دسته‌های دیده‌شده از مقدار $T > ۱$ بهره می‌گیریم تا خروجی شبکه میزان شباهت به دسته‌های مختلف را به صورت هموارتر نشان دهد. هیستوگرام حاصل از تصویر $\mathbf{x}$ با این روش را با نماد ψx نمایش می‌دهیم. نگاشت ψ که با یک شبکه‌ی عصبی عمیق مدل شده، از سه قسمت تشکیل شده است: ۱) ۱۶ لایه پیچشی شرح داده شده در بخش ؟؟، ۲) سه لایه با اتصالات کامل که وزن‌های آن‌ها با آموزش روی نمونه‌های دسته‌های دیده شده به دست می‌آید و ۳) تابع فعال‌سازی نهایی از رابطه ۳-۱۱.

برای تکمیل روش پیشنهادی برای دسته‌بندی صفرضرب باید نگاشتی برای بردن بردارهای توصیف دسته‌های دیده نشده به این فضا، یعنی فضای هیستوگرام دسته‌های دیده‌شده ارائه کنیم. برای این کار از مجموع عکس فاصله‌ی اقلیدسی و فاصله بلوکی[۷] بردارهای توصیف دسته‌ها با یکدیگر استفاده می‌کنیم، به عبارتی برای بردار توصیف c متعلق به یک دسته‌ی دیده‌نشده داریم:

$$\theta_j(\mathbf{c}) = \frac{۱}{\|\mathbf{c} - \mathbf{c_j}\|_۲ + \|\mathbf{c} - \mathbf{c_j}\|_۱}, \quad j = ۱, \ldots, n_s. \qquad (۱۲-۳)$$

دسته‌بندی در این فضا با استفاده از دسته‌بند نزدیک‌ترین همسایه صورت می‌گیرد، به عبارت دیگر اگر تابع اختصاص

[۷] Manhattan Distance

برچسب را با $\ell(\cdot)$ نشان دهیم:

$$\ell(\mathbf{x}) = \underset{i=n_s,\ldots,n_s+n_u}{\arg\min} \|\psi(\mathbf{x}]) - \theta(\mathbf{c}_i)\|_2^2 . \qquad (۱۳-۳)$$

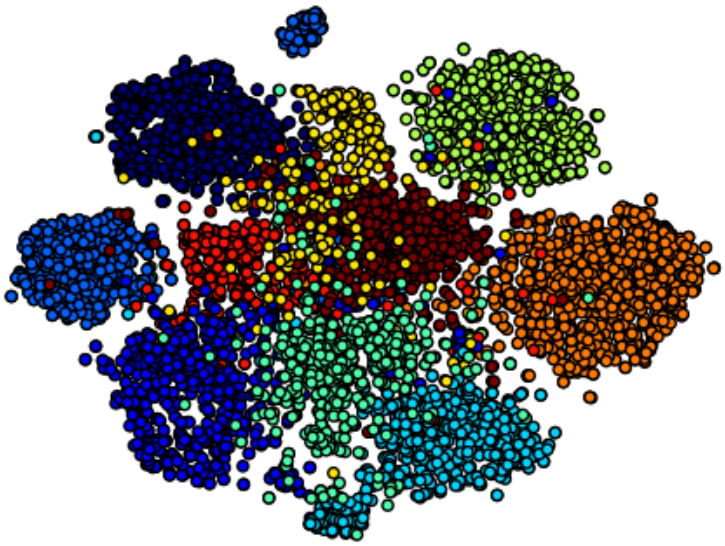
در نهایت با استفاده از تابع مطابقتی که در بخش ۳-۴ معرفی می‌شود، می‌توان نتایج حاصل از دسته‌بند نزدیک‌ترین همسایه را بهبود داد.

۳-۴ تابع مطابقت مبتنی بر خوشه‌بندی

در اکثر روش‌های پیشین که در فصل ۲ مرور شد، تابع مطابقت میان تصاویر و توصیف‌ها برای اختصاص برچسب به داده‌های آزمون بر اساس فاصله کمینه یا ضرب داخلی بیشینه در یک فضای مشترک محاسبه می‌شد. استثناهای این موضوع، استفاده از روش انتشار برچسب در [۳۸] و [۴۱] و هم‌چنین پیش‌بینی مستقیم برچسب‌ها در [۴۰] و [۴۴] هستند.

در این بخش ما یک تابع مطابقت جدید بر اساس یک خوشه‌بندی روی داده‌های دسته‌های دیده نشده، تعریف می‌کنیم. اگر فضای نمایش تصاویر دارای این خاصیت باشد که دسته‌های مختلف به صورت خوشه‌های مجزا باشند، استفاده از خوشه‌بندی برای انتساب برچسب از نظر شهودی توجیه‌پذیر است. با توجه به نمایش غنی بوجود آمده برای تصاویر توسط شبکه‌های ژرف این فرض در بسیاری از موارد برقرار است. برای نمونه، نمایش t-SNE نمونه‌های آزمون مجموعه داده‌های AwA در تصویر ۳-۴ نشان داده شده است و برقراری فرض قابل خوشه‌بندی بودن در آن قابل مشاهده است. این ادعا با استفاده از آزمایش در بخش ۴-۸ اثبات خواهد شد. روش‌های پیشنهادی ما در این فصل بر اساس این ساختار و استفاده از وجود چنین خاصیتی در فضای تصاویر است.

یک راه استفاده از چنین خاصیتی در فضای تصاویر، معرفی یک تابع مطابقت است که علاوه بر شباهت نگاشت‌یافته‌ی نمونه‌ها و توصیف‌ها به سایر نمونه‌های موجود در همسایگی هر نمونه نیز وابسته باشد. بدین منظور ما یک تابع مطابقت جدید پیشنهاد می‌دهیم که در آن برچسب تعلق گرفته به هر نمونه به نمونه‌هایی که با آن‌ها در یک خوشه قرار گرفته است وابسته است. برای این منظور ابتدا باید یک خوشه‌بندی روی نمونه‌ها انجام شود سپس با استفاده از یک معیار (که یک نمونه از آن را در بخش ۳-۶ معرفی می‌کنیم) میزان شباهت خوشه به توصیف تعیین می‌شود. این در مقابل حالتی است که تابع مطابقت، میزان شباهت هر نمونه را به طور جداگانه با توصیف دسته‌ها محاسبه می‌کرد. در این حالت هر خوشه



شکل ۳-۴: نمایش دوبعدی بوسیله t-SNE برای ده دسته‌ی آزمون از مجموعه دادگان AwA با ده رنگ متفاوت نشان داده شده است. درستی فرض قابل خوشه‌بندی در تصویر مشخص است، یعنی ویژگی‌های استخراج شده با استفاده از شبکه‌های ژرف توانایی ایجاد تمایز بالا میان دسته‌ها را دارا هستند و نمونه‌های هر دسته نیز نزدیک به یکدیگر هستند.

باید یک برچسب دریافت کند و برچسب اختصاص یافته به هر خوشه، توسط تمام اعضای آن به ارث برده می‌شود. این تابع مطابقت تا کنون در روش‌های موجود برای یادگیری صفرضرب استفاده نشده است. نسخه‌های متفاوتی از این تابع مطابقت بر حسب چگونگی تعیین برچسب هر خوشه قابل ارائه است که ما در اینجا دو مورد از آن‌ها را بیان می‌کنیم. یک نحوه برای انتساب خوشه‌ها به دسته‌های دیده نشده استفاده از رای اکثریت است، در این حالت بایست ابتدا یک پیش‌بینی برای همه نمونه‌های آزمون صورت بگیرد (برای مثال با استفاده از روش معرفی شده در بخش ۳-۲)، فرض کنید که این برچسب‌گذاری را با z_n برای $N_s < n \leqslant N_s + N_u$ نشان دهیم. همچنین یک خوشه‌بندی روی داده‌ها انجام شده که آن را با r_n برای $N_s < n \leqslant N_s + N_u$ نشان می‌دهیم. حال $\ell(k)$ که برچسب خوشه‌ی $k-$م است از رابطه زیر تعیین خواهد شد:

$$\ell(k) = \underset{n_s < i \leqslant n_s + n_u}{\arg\max} \Big[\sum_{m=N_s+1}^{N_s+N_u} \mathbb{1}(r_n = k) \times \mathbb{1}(z_n = i) \Big]. \qquad (۱۴-۳)$$

در این حالت، این تابع مطابقت قابل اضافه شدن به روش‌های دیگر نیز هست. به این صورت که پیش‌بینی‌های انجام شده در آن روش را در نظر گرفته و با استفاده از آن‌ها در هر خوشه رای‌گیری انجام دهیم تا برچسبی که کل خوشه دریافت می‌کند

تعیین شود. آزمایش‌ها نشان می‌دهند که اضافه شدن این تابع مطابقت عمل‌کرد شبکه عصبی چندوظیفه‌ای پیشنهادی را بهبود می‌دهد.

یک نسخه‌ی دیگر از این تابع مطابقت که در روش ارائه شده در بخش ۳-۶ مورد استفاده قرار می‌گیرد مربوط به حالتی است که نگاشتی از فضای توصیف دسته‌ها به فضای تصاویر وجود داشته باشد. فرض کنید که چنین نگاشتی یادگرفته شده و با θ نشان داده شود. هم‌چنین نگاشت $\phi(x)$ نگاشت تبدیل تصاویر به ویژگی‌های ژرف است. مانند حالت قبل یک خوشه‌بندی r_n روی نمونه‌های آزمون صورت گرفته و $\boldsymbol{\mu_k}$ مرکز خوشه $k-$م را نشان می‌دهد. در نتیجه داریم:

$$r_n = \arg\min_k \|\phi(\mathbf{x_n}) - \boldsymbol{\mu_k}\|_۲^۲ . \qquad (۱۵\text{-}۳)$$

حالا میزان مطابقت نمونه‌ی $\mathbf{x_n}$ و توصیف $\mathbf{c}$ با استفاده از رابطه زیر تعریف می‌شود:

$$compatibility(\mathbf{x}, \mathbf{c}) = - \|\boldsymbol{\mu}_{\boldsymbol{r_n}} - \theta(\mathbf{c})\|_۲ . \qquad (۱۶\text{-}۳)$$

تعبیر رابطه فوق این است که میزان مطابقت نمونه x با دسته‌ی آزمون y، بر اساس میزان نزدیکی مرکز خوشه‌ای که x به آن تعلق دارد با تصویر توصیف دسته‌ی y در فضای ویژگی‌های تصاویر تعریف می‌شود.

۳-۵ یک خوشه‌بندی نیمه‌نظارتی

عمل‌کرد تابع مطابقت معرفی شده در بخش قبل وابسته به دقت خوشه‌بندی انجام شده روی داده‌هاست. در واقع دقت خوشه‌بندی انجام شده، حد بالای دقت نهایی روش خواهد بود؛ چرا که در تابع مطابقت معرفی شده، تمام اعضای یک خوشه برچسب یکسانی را دریافت می‌کنند در نتیجه اگر اعضای درون یک خوشه هم‌دسته نباشند حداکثر اعضای متعلق به یکی از دسته‌ها برچسب صحیح دریافت می‌کنند و پیش‌بینی برای سایر اعضای خوشه که متعلق به دسته‌های دیگر هستند نادرست خواهد بود. این حد بالا در حالتی رخ می‌دهد که هر خوشه برچسبی را دریافت کند که برچسب صحیح اکثر اعضای آن است. با توجه به این موضوع وجود یک خوشه‌بندی دقیق برای استفاده از این تابع مطابقت ضروری است. البته در آزمایش‌های انجام شده، با به کارگیری تابع مطابقت پیشنهادی و استفاده از الگوریتم خوشه‌بندی k-means [۴۸] نیز می‌توان به عمل‌کرد پیشگام دست پیدا کند. اما این الگوریتم در خوشه‌بندی نمونه‌های آزمون استفاده‌ای از برچسب‌هایی که برای نمونه‌های آموزش وجود دارد، نخواهد کرد و این اطلاعات می‌توان باعث بهبود عمل‌کرد خوشه‌بندی شود. از طرفی الگوریتم‌های نیمه‌نظارتی موجود برای خوشه‌بندی نیز بر مسئله یادگیری صفرضرب تطابق ندارند. در حالت معمول یادگیری

نیمه‌نظارتی [۲]، مسئله به این صورت تعریف می‌شود که داده‌های برچسب‌دار و بدون برچسب همگی به یک مجموعه دسته‌ی یکسان تعلق دارند و داده‌های بدون برچسب نیز در نهایت برچسب یکسانی با داده‌های برچسب‌دار دریافت می‌کنند. این در حالی‌ست که در مسئله یادگیری صفرضرب، نمونه‌های بدون برچسب در دسته‌های مجزا از نمونه‌های برچسب‌دار قرار می‌گیرند. با توجه به این موضوع، یک روش خوشه‌بندی نیمه‌نظارتی پیشنهاد می‌کنیم که با فرض‌های مسئله یادگیری صفرضرب منطبق باشد. در این روش خوشه‌بندی همانند k-means عمل می‌شود با این تفاوت که اگر شماره خوشه نمونه‌های دسته‌های دیده شده برابر با برچسب صحیح آن‌ها نباشد، جریمه‌ای در نظر گرفته می‌شود. تابع هزینه این روش به این صورت تعریف شده است:

$$\min_{R,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k} \sum_{n,k} r_{nk} \left\|\mathbf{x_n} - \boldsymbol{\mu_k}\right\|_2^2 + \beta \sum_{n=1}^{N_s} \mathbb{1}(\mathbf{r_n} \neq \mathbf{y_n}). \tag{۱۷-۳}$$

در این معادله $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu_k}$ مراکز خوشه‌ها و R ماتریس اختصاص داده‌ها به خوشه‌هاست؛ جمله اول همان جمله موجود در تابع هزینه‌ی k-means است. علاوه بر این، در جمله‌ی دوم برای هر نمونه‌ی برچسب‌دار، اگر به خوشه‌ای تعلق بگیرد که شماره آن با برچسبش متفاوت باشد، جریمه β در نظر گرفته می‌شود. در نتیجه این روش، n_s خوشه ابتدایی را به سمت این سوق می‌دهند که همان n_s دسته‌ی دیده شده باشند. β یک پارامتر مدل است که اهمیت این جمله اضافه شده را تعیین می‌کند.

۳-۵-۱ بهینه‌سازی

کمینه‌کردن تابع هزینه معرفی شده در رابطه (۳-۱۷)، با توجه به این که R یک افراز[۸] روی نمونه‌هاست، مانند بهینه‌سازی تابع هزینه‌ی k-means یک مسئله‌ی اِن‌پی‌ـسخت است [۴۹]. در نتیجه ما از یک تقریب مشابه الگوریتم خوشه‌بندی k-means استفاده می‌کنیم که یک بهینه محلی برای این تابع را پیدا می‌کند. به این منظور، یک روند تناوبی[۹] میان بهینه کردن بر اساس R و μ_kها به کار گرفته می‌شود. برای بروز رسانی μ_k روی اعضای خوشه k میانگین گرفته می‌شود:

$$\boldsymbol{\mu_k} = \frac{\sum_{n=1}^{N_s+N_u} \mathbb{1}(r_{nk}=1)\mathbf{x_n}}{\sum_{n=1}^{N_s+N_u} \mathbb{1}(r_{nk}=1)}. \tag{۱۸-۳}$$

[۸] Partitioning

[۹] Alternative

برای بروز رسانی R هر نمونه که متعلق به دسته‌های دیده‌نشده است و برچسب صحیحی برای آن موجود نیست، به خوشه‌ای اختصاص می‌یابد که کمترین فاصله را با مرکز آن دارد:

$$R_{(n)} = \mathbf{1}_{\arg\min_k \|x_n - \mu_k\|_2^2}, \quad n = N_s + 1, \ldots, N_s + N_u \tag{۱۹-۳}$$

اما برای نمونه‌های دسته‌های دیده شده که برچسب صحیحی برای آن‌ها موجود است علاوه بر فاصله تا مرکز خوشه مقدار جمله دوم رابطه (۳-۱۷) نیز در تخصیص خوشه موثر است. در این حالت برای تخصیص نمونه به خوشه‌ای با شماره‌ای متفاوت با برچسب صحیح جریمه‌ای به مقدار β در نظر گرفته می‌شود.

$$R_{(n)} = \mathbf{1}_{\arg\min_k \|x_n - \mu_k\|_2^2 + \beta \mathbb{1}(y_n \neq 1_k)}, \quad n = 1, \ldots, N_s \tag{۲۰-۳}$$

با توجه به این که در قوانین بروزرسانی در روابط (۳-۱۸) تا (۳-۲۰) مقدار پیشنهاد شده برای هر پارامتر با فرض ثابت بودن پارامترها، مقدار بهینه است این روند به یک بهینه‌ی محلی همگرا خواهد شد.

برای مقداردهی اولیه به μ_k برای خوشه‌های مربوط به دسته‌های دیده شده، میانگین عناصر آن‌ها را قرار می‌دهیم:

$$\boldsymbol{\mu_k^{\cdot}} = \frac{\sum_{n=1}^{N_s} \mathbb{1}(Y_{s(n)} = \mathbf{1_k}) \cdot \mathbf{x_n}}{\sum_{n=1}^{N_s} \mathbb{1}(Y_{s(n)} = \mathbf{1_k})}, \quad 1 \leqslant k \leqslant n_s \tag{۲۱-۳}$$

که $\boldsymbol{\mu_k^{\cdot}}$ برای نشان دادن مقدار در لحظه‌ی صفر یا همان مقدار اولیه برای شروع الگوریتم بهینه‌سازی بکار رفته است. برای سایر خوشه‌ها، یعنی خوشه‌های مربوط به دسته‌های دیده نشده از الگوریتم k-means++ [۵۰] با $k' = k - n_s$ یعنی تعداد خوشه‌هایی که به جز دسته‌های دیده‌شده وجود دارد، استفاده می‌کنیم.

۳-۶ روش یادگیری صفرضرب خوشه‌بندی و یادگیری نگاشت مجزا

در این بخش روشی معرفی می‌شود که همراه با خوشه‌بندی بخش قبل یک چارچوب برای دسته‌بندی در مسئله یادگیری صفرضرب را تشکیل می‌دهند. برای نسبت دادن برچسب به خوشه‌ها، به دنبال یافتن نمایشی از امضای هر دسته در فضای تصاویر به عنوان نماینده آن دسته در فضای تصاویر هستیم. از نظر شهودی مطلوب است که این نماینده‌ها بر مرکز خوشه‌هایی که در فضای تصاویر تشکیل می‌شود منطبق باشند. برای محقق شدن این خاصیت، نگاشت را به صورتی یاد می‌گیریم که حاصل نگاشت توصیف دسته‌های آموزش منطبق بر میانگین نمونه‌های این دسته‌ها باشد:

$$D = \arg\min_D \ \|X_s - DZ_s\|_{Fro}^2 + \alpha \|D\|_{Fro}^2 . \tag{۲۲-۳}$$

در این معادله، ستون‌های $Z_s \in \mathbb{R}^{a \times N_s}$ امضای دسته‌های نمونه‌های X_s هستند و α یک پارامتر است که با اعتبارسنجی تعیین خواهد شد. مسئله تعریف شده برای یافتن نگاشت D، امضای کلاس را طوری می‌نگارد که نزدیک به مرکز نمونه‌های آن دسته باشد و این در حالت ایده‌آل همان مرکز خوشه‌ها خواهد بود. یعنی انتظار می‌رود حاصل نگاشت امضای هر دسته با استفاده از D در مرکز نمونه‌های آن دسته قرار بگیرد، از طرفی در یک خوشه‌بندی ایده‌آل خوشه‌بندی سازگار با برچسب‌های صحیح داده‌هاست در نتیجه میانگین اعضای یک خوشه در حقیقت میانگین اعضای یکی از دسته‌های آزمون خواهد بود. حالا تنها گام باقی‌مانده برای تکمیل روش این است که به گونه‌ای تشخیص داده شود که هر کدام از خوشه‌ها با کدام یک از دسته‌های دیده‌نشده در تناظر است برای این کار از دسته‌بند نزدیک‌ترین همسایه استفاده می‌کنیم به این صورت که مراکز خوشه‌ها و حاصل نگاشت امضای دسته‌ها در فضای تصاویر را در نظر گرفته و هر خوشه را به دسته‌ای انتساب می‌دهیم که نمایش امضای آن دسته در این فضا به مرکز خوشه نزدیک‌تر است.

یافتن نگاشت D بر اساس کمینه‌کردن رابطه (۳-۲۲) به وسیله‌ی یک رابطه فرم بسته قابل انجام است. به این منظور از رابطه‌ی (۳-۲۲) برحسب عناصر D مشتق می‌گیریم و برابر صفر قرار می‌دهیم:

$$\frac{\partial}{\partial D}\left\|X_s - DZ_s\right\|_{Fro}^{۲} + \alpha\left\|D\right\|_{Fro}^{۲} = \frac{\partial}{\partial D}tr((X_s - DZ_s)^T(X_s - DZ_s)) + \alpha\frac{\partial}{\partial D}tr(D^TD)$$
$$= ۲(DZ_s - X_s)Z_s^T + ۲\alpha D = ۰$$
$$\Rightarrow DZ_sZ_s^T - X_sZ_s^T + \alpha D = ۰ \Rightarrow D(Z_sZ_s^T + \alpha I) = X_sZ_s^T$$

و در نتیجه خواهیم داشت:

$$D = X_sZ_s^T(Z_sZ_s^T + \alpha I)^{-۱}. \qquad (۳-۲۳)$$

برای تخصیص برچسب به هر خوشه از این رابطه استفاده می‌کنیم:

$$\ell(\boldsymbol{\mu_k}) = \underset{u=۱,\ldots,n_u}{\arg\min}\ \left\|\boldsymbol{\mu_k} - DC_u\right\|_{Fro}^{۲} \qquad (۳-۲۴)$$

و تمامی عناصر خوشه‌ی kم برچسب $\ell(\boldsymbol{\mu_k})$ را دریافت می‌کنند. با توجه به انجام مستقل مراحل خوشه‌بندی و یادگیری نگاشت این روش را *یادگیری نگاشت و خوشه‌بندی مجزا می‌نامیم* که با توجه به نوع خوشه‌بندی مورد استفاده (خوشه‌بندی نیمه‌نظارتی پیشنهادی یا الگوریتم k-means) ممکن است پسوند نیمه‌نظارتی نیز به آن اضافه شود.

در این روش سه پارامتر وجود دارد، یک پارامتر α در معادله (۳-۲۲) است و دو پارامتر دیگر که مربوط به خوشه‌بندی نیمه‌نظارتی هستند، یعنی k و β در معادله (۳-۱۷). در آزمایش‌ها عملی دریافتیم که روش به مقدار پارامتر α حساس

الگوریتم ۲ الگوریتم یادگیری صفرضرب خوشه‌بندی و یادگیری نگاشت مجزای نیمه‌نظارتی

۱ **ورودی:** تصاویر و توصیف‌های آموزش و آزمون و برچسب‌های نمونه‌های آموزش X_s, X_u, Y_s, Z_s, C_u

۲ **خروجی:** برچسب‌های پیش‌بینی شده برای نمونه‌های آزمون:Y_u

۳ $\boldsymbol{\mu_k}$ را برای $k = ۱, \ldots, n_s$، با رابطه (۳-۲۱) مقداردهی کن.

۴ $\boldsymbol{\mu_k}$ را برای $k = n_s + ۱, \ldots, n_s + n_u$، با استفاده از k-means++ مقداردهی کن.

۵ تا همگرایی به یک بهینه‌ی محلی، موارد زیر را تکرار کن

۶ $\arg\min_i \|x_n - \mu_i\|_۲^۲ \rightarrow a_n$ برای $n = N_s + ۱, \ldots, N_s + N_u$

۷ $\arg\min_i \|x_n - \mu_i\|_۲^۲ + \beta \mathbb{1}(y_n \neq ۱_i) \rightarrow a_n$ برای $n = ۱, \ldots, N_s$

۸ $\sum_n \mathbf{x_n} \mathbb{1}(a_n = k) / \sum_n (\mathbb{1}(a_n = k) \rightarrow \mu_{\mathbf{k}}$

۹ $X_s Y_s^T (Y_s Y_s^T + \alpha I)^{-۱} \rightarrow D$ برای $k \in \{۱, ۲, \ldots, n_s + n + u\}$

۱۰ $\arg\min_j \|\mu_{\mathbf{k}} - (DS_u)_{(j)}\|_۲ \rightarrow l[k]$ برای $k \in \{۱, ۲, \ldots, n_s + n + u\}$

۱۱ $\mathbf{۱}_{l[a_n]} \rightarrow (\mathbf{Y_u})_{(\mathbf{n})}$ برای $n \in \{N_s + ۱ \ldots N_s + N_u\}$

۱۲ Y_u را برگردان

است در نتیجه مقدار آن توسط یک روند اعتبارسنجی تعیین خواهد شد، نحوه‌ی اعتبارسنجی به صورت دقیق در بخش ۴-۲ بیان خواهد شد. در مقابل، مدل به پارامترهای k و β حساس نبود، در نتیجه برای ساده و سریع‌تر شدن روند آموزش مقدار آن‌ها را ثابت در نظر گرفته‌ایم. برای k مقدار $k = n_s + ۲n_u$ در نظر گرفته شده است چرا که عموما افزایش تعداد خوشه‌ها نسبت به دسته‌ها می‌تواند دسته‌هایی که الزاما به صورت یک خوشه نیستند را هم مدل کند. با ارائه نتایج عملی تاثیر این دو پارامتر در فصل ۴-۷ نشان داده می‌شود که این انتخاب‌ها، انتخاب‌های تاثیرگذاری نبوده و عملکرد روش به مقدار این دوپارامتر حساس نیست. در آزمایش‌ها عملی که در فصل ۴ گزارش می‌شود، مشاهده می‌شود که این روش عملکرد پیشگام در دقت دسته‌بندی صفرضرب را روی سه مجموعه دادگان از چهار مجموعه بهبود می‌بخشد.

روند کامل این روش پیشنهادی در الگوریتم ۲ بیان شده است.

۳-۷ خوشه‌بندی و نگاشت توام

روش ارائه شده در فصل قبل، هر چند که به دقت دسته‌بندی بالاتری از روش‌های پیشین دست پیدا می‌کند اما دقت دسته‌بندی در آن توسط دقت خوشه‌بندی صورت گرفته محدود شده است. هم‌چنین انجام جداگانه عمل خوشه‌بندی و یادگیری نگاشت از فضای توصیف‌ها به فضای تصاویر امکان استفاده از کامل از اطلاعات برای یادگیری توام و سازگاری بین این دو یادگیری را از بین می‌برد. این درحالی است که با توجه به وجود داده‌های برچسب‌دار از دسته‌های دیده شده، یادگیری توام این دو قسمت یعنی خوشه‌بندی و نگاشت از فضای توصیف‌ها به فضای تصاویر می‌تواند باعث شود که اختصاص نمونه‌های آزمون به خوشه‌ها به گونه‌ای انجام شود که همزمان هر دو معیار شبیه بودن به سایر نمونه‌های درون خوشه (که تنها در مرحله خوشه‌بندی روش قبلی در نظر گرفته می‌شد) و معیار نزدیکی نمونه‌های یک خوشه به حاصل نگاشت توصیف دسته‌ی آن‌ها (که تنها در مرحله یادگیری نگاشت دیده می‌شد) در نظر گرفته شوند. برای دست‌یابی به چنین هدفی یک مسئله بهینه‌سازی معرفی می‌کنیم که خوشه‌بندی و نگاشت توصیف دسته‌ها به فضای تصاویر در آن به صورت توام انجام شود:

$$\min_{R,D} \|X_s - DZ_s\|_{Fro}^2 + \lambda \left\|X_u - DC_u R^T\right\|_{Fro}^2 + \eta \|D\|_{Fro}^2 , \qquad (۲۵-۳)$$

$$s.t. \quad R \in \{0,1\}^{N_u \times n_u}.$$

در این معادله η و λ فراپارامترهای مدل هستند. جمله اول و سوم در رابطه بالا مشابه رابطه (۳-۲۲) هستند و تاثیر آن‌ها همانند حالت قبل این است که نگاشت D بتواند امضای دسته‌های دیده نشده را به مرکز تصاویر هر دسته بنگارد. جمله دوم که در این معادله اضافه شده، ذاتا یک جمله خوشه‌بندی است. اگر جمله دوم در عبارت بالا را از فرم ماتریسی خارج کرده و بر حسب عناصر R بیان کنیم این مسئله واضح‌تر خواهد شد:

$$\sum_{n=N_s+1}^{N_s+N_u} \sum_{k=1}^{n_u} r_{nk} \|\mathbf{x_n} - D\mathbf{c_k}\|_2^2 , \qquad (۲۶-۳)$$

که مشابه تابع هزینه‌ی k-means است، با این تفاوت که مراکز خوشه‌ها کاملا آزاد نیستند بلکه مراکز خوشه‌ها باید تصویر امضای دسته‌های دیده نشده باشد که توسط نگاشت D به فضای تصاویر نگاشته شده است. در این حالت برچسب‌های پیش‌بینی شده برای نمونه‌ها همان انتساب‌های آن‌ها به خوشه‌هاست که در طول جریان آموزش توامان با نگاشت D یادگرفته می‌شود. در نتیجه مشکل بیان شده برای روش قبل، در این روش وجود ندارد. جمله خوشه‌بندی را در این مسئله بهینه‌سازی می‌توان به این صورت نیز تعبیر کرد که این جمله یادگیری نگاشت D را به صورتی بهبود می‌دهد که مشکل جابجایی

الگوریتم ۳ الگوریتم یادگیری نگاشت و خوشه‌بندی به صورت توام

۱ **ورودی:** تصاویر و توصیف‌های آموزش و آزمون و برچسب‌های نمونه‌های آموزش X_s, X_u, Y_s, Z_s, C_u

۲ **خروجی:** برچسب‌های پیش‌بینی شده برای نمونه‌های آزمون:R

۳ R را با خروجی الگوریتم ۲ مقدار دهی کن.

۴ تا هنگامی که مقدار R تغییر می‌کند، تکرار کن:

۵ D را با رابطه (۳-۲۷) بروزرسانی کن.

۶ عناصر R را با استفاده از رابطه (۳-۲۸) بروزرسانی کن.

۷ R را برگردان

دامنه در آن وجود نداشته باشد. در حالت عادی برای یادگیری نگاشت D توسط رابطه (۳-۲۲) تنها از نمونه‌های آموزش استفاده می‌شد، در نتیجه مشکل جابجایی دامنه برای داده‌های آزمون بوجود می‌آمد، چرا که این داده‌ها در تعیین نگاشت D بی‌تاثیر بوده‌اند. اما جمله اضافه شده در روش فوق الزام می‌کند که امضای هر دسته‌ی دیده نشده نزدیک به تعدادی از داده‌های آزمون (که توسط R مشخص می‌شوند) نگاشته شود. این مسئله می‌تواند مانع از مشکل جابجایی دامنه شود. این موضوع در بخش ۴-۹ بیشتر بررسی خواهد شد.

۳-۷-۱ بهینه‌سازی

مسئله بهینه‌سازی رابطه (۳-۲۵) بر حسب هر دو متغیر R و D محدب[۱۰] نیست، در نتیجه برای یافتن یک بهینه محلی از یک روند تناوبی میان بهینه‌کردن بر حسب R و D استفاده می‌کنیم. با فرض ثابت بودن R بهینه‌سازی بر اساس D دارای جواب به فرم بسته است، برای بدست آوردن این جواب نسبت به عناصر D از رابطه (۳-۲۵) مشتق می‌گیریم:

$$\frac{\partial}{\partial D} \|X_s - DZ_s\|_{Fro}^2 + \lambda \left\|X_u - DC_u R^T\right\|_{Fro}^2 + \eta \|D\|_{Fro}^2$$
$$= 2(DZ_s - X_s)Z_s^T + \lambda(DC_u R^T - X_u)RC_u^T + \eta D = 0$$
$$\Rightarrow D(Z_s Z_s^T + C_u R^T R C_u^T + \eta I) - X_s Z_s^T + X_u R C_u^T = 0$$

[۱۰] Convex

در نتیجه خواهیم داشت:

$$D = (X_s Z_s^T + \beta X_u R C_u^T)(Z_s Z_s^T + \beta C_u R^T R C_u^T + \eta I)^{-1}, \quad (۲۷-۳)$$

و مقدار بهینه برای R، زمانی که D ثابت باشد، با نسبت دادن هر نمونه به نزدیک‌ترین مرکز خوشه به دست می‌آید:

$$r_{ij} = \mathbb{1}[j = \arg\min_k \|X_{u(i)} - D S_{u(k)}\|_2]. \quad (۲۸-۳)$$

در این روند بین بروز رسانی D و R تناوب انجام می‌شود تا جایی که R ثابت بماند یعنی تغییری در برچسب‌های پیش‌بینی شده برای هیچ‌کدام از نمونه‌ها رخ ندهد. در آزمایش‌های انجام شده این همگرایی همواره در کمتر از ۲۰ بار بروز رسانی به دست می‌آید.

مراحل این روش در الگوریتم ۳ آمده است. در مورد گام ۳ از این الگوریتم این توضیح لازم است که از میان R و D تنها یکی نیاز به مقداردهی اولیه دارد؛ چرا که روابط بروز رسانی هر کدام تنها به مقدار پارامتر دیگر بستگی دارد و از مقدار پیشین خود مستقل است. در نتیجه در روند بهینه‌سازی تناوبی هرکدام از R و D که ابتدا بروز رسانی شوند، در بروز رسانی آن‌ها تنها به مقدار اولیه پارامتر دیگر نیاز است و خود آن نیاز به مقداردهی اولیه ندارند. ما در اینجا R را مقداردهی اولیه کرده و روند بهینه‌سازی را با بروزرسانی D آغاز می‌کنیم. این انتخاب نسبت به حالت مقابلش یعنی مقداردهی اولیه D با رابطه (۳-۲۳) در گام سوم الگوریتم و تعویض گام‌های ۵ و ۶ برتری دارد. چرا که در مقداردهی اولیه استفاده شده برای R از اطلاعات موجود در تمام داده‌ها از جمله نمونه‌های آزمون نیز استفاده شده است حال آن‌که مقداردهی D با رابطه‌ی (۳-۲۳) تنها به نمونه‌های آموزش وابسته بوده و از اطلاعات بدون نظارت موجود در نمونه‌های آزمون بهره‌ای نمی‌برد. برای نشان دادن صحت این ادعا نتیجه دقت دسته‌بندی در هردوی این حالات سنجیده شده و نتایج آن در بخش ۴-۸ گزارش شده است.

۳-۸ جمع‌بندی

در این بخش ابتدا نحوه‌ی استخراج ویژگی با شبکه‌های عصبی پیچشی ژرف شرح داده شد. سپس یک شبکه عصبی برای انجام پیش‌بینی صفت در مسئله یادگیری صفرضرب ارائه شد. پس از آن یک تابع مطابقت جدید برای مسئله یادگیری صفرضرب ارائه شد. برای بهره‌گیری مناسب از این تابع مطابقت یک خوشه‌بندی دقیق روی نمونه‌های آزمون مورد نیاز بود. به این خاطر، سپس یک الگوریتم خوشه‌بندی نیمه‌نظارتی که با فرض‌های مسئله‌ی یادگیری صفرضرب هم‌خوانی

داشته باشد ارائه گردید. با فراهم آمدن این مقدمات یک روش برای دسته‌بندی صفرضرب با استفاده از تابع مطابقت و خوشه‌بندی پیشنهادی و یک نگاشت خطی از فضای توصیف دسته‌ها به فضای تصاویر ارائه شد. بعد از آن یک روش که یادگیری نگاشت و خوشه‌بندی در آن به صورت توام انجام شود ارائه شد و در مورد نحوه‌ی بهینه‌سازی توابع پیشنهادی در این روش‌ها بحث شد.

فصل ۴

نتایج عملی

در این فصل، روش پیشنهادی را روی چند مجموعه دادگان آزمایش کرده و نتایج آن را با سایر روش‌های ارائه شده برای یادگیری صفرضرب مقایسه می‌کنیم. در این فصل ابتدا مجموعه دادگان مورد استفاده در آزمایشات معرفی می‌شوند. سپس کارایی روش‌های ارائه شده در بخش‌های ۳-۲ تا ۳-۷ با آزمایش روی این مجموعه دادگان مورد بررسی قرار می‌گیرد و تاثیر هر قسمت‌های مختلف هر یک از روش‌های پیشنهادی و پارامترهای موجود در آن‌ها سنجیده می‌شود.

۴-۱ مجموعه دادگان مورد استفاده

برای آزمایشات عملی ما از چهار مجموعه داده‌ی مرسوم برای سنجش عمل‌کرد روش‌های یادگیری صفرضرب استفاده می‌کنیم.

Animal with Attributes (AwA) [۱۳]: این مجموعه داده شامل تصاویری از ۵۰ گونه از پستانداران است. هر دسته توسط یک بردار صفت ۸۵–بعدی توصیف می‌شود. در این مجموعه داده توصیف‌های دسته‌ها هم به صورت مقادیر دودویی به معنای وجود یا عدم وجود آن صفت وجود دارند و هم توسط اعداد حقیقی با توجه به میزان وجود آن صفت در هر دسته در دسترس هستند. در آزمایش‌های انجام شده از مقادیر پیوسته برای توصیف دسته‌ها استفاده شده است، چرا که در روش‌های پیشین نشان داده شده که این مقادیر توانای ایجاد تمایز بیشتری دارند [۲۷]. هم‌چنین از تقسیم‌بندی آموزش و آزمون انجام شده در خود مجموعه داده استفاده می‌کنیم که در آن ۴۰ دسته به عنوان دسته‌های دیده

شده و ۱۰ دسته به عنوان دسته‌های دیده نشده در نظر گرفته شده‌اند.

aPascal/aYahoo (aPY)[۸]: مجموعه تصاویر VOC 2008 [۵۱] که شامل ۲۰ دسته است بعنوان دسته‌های دیده شده در نظر گرفته شده است و تصاویر aYahoo که شامل ۱۲ دسته هستند به عنوان دسته‌های دیده نشده. برای این دو مجموعه داده، بردار صفت‌های ۶۴−بعدی دودویی برای هر تصویر موجود است. برای بدست آوردن توصیف هر دسته که در مسئله یادگیری صفرضرب مورد نیاز است، همانند روش‌های پیشین، روی بردار صفت‌های تصاویر هر دسته میان گرفته شده است [۱۳].

SUN Attribute [۵۲]: مجموعه تصاویر SUN شامل ۷۱۷ دسته می‌باشد و در این مجموعه برای هر یک از تصاویر یک بردار صفت ۱۰۲−بعدی موجود است که برای تبدیل آن به توصیف‌های در سطح دسته‌ها، روی بردار صفت‌های تصاویر هر دسته میانگین گرفته شده است. ما تقسیم‌بندی آموزش/آزمون انجام گرفته در [۴۳] استفاده می‌کنیم که در آن ۱۰ دسته به عنوان دسته‌های دیده نشده در نظر گرفته شده‌اند.

Caltech UCSD Birds-2011 (CUB) [۵۳]: این مجموعه داده شامل تصاویری از ۲۰۰ گونه از پرندگان است. هر تصویر با ۳۱۲ صفت دودویی توصیف می‌شود و توصیف در نظر گرفته شده برای هر دسته میانگین توصیف نمونه‌های آن دسته است. تقسیم‌بندی مورد استفاده برای دسته‌های آموزش و آزمون، دسته‌بندی مورد استفاده در [۵۴] است که توسط کارهای بعدی نیز مورد استفاده قرار گرفته است [۳۶، ۲۷، ۳۴].

در تمام مجموعه داده‌ها، برای تصاویر از ویژگی‌های بدست آمده با شبکه‌های ژرف استفاده می‌کنیم چرا که توانایی ایجاد تمایز این ویژگی‌ها نسبت به ویژگی‌های کم‌عمق سنتی مانند SIFT و HOG بیشتر است. ویژگی‌های مورد استفاده از اولین لایه با اتصالات چگال از شبکه ۱۹ لایه‌ی VGG [۱۴] بدست آمده است. پیش آموزش شبکه روی زیرمجموعه‌ای از مجموعه دادگان ImageNet [۵۵] مربوط به چالش سال ۲۰۱۲ دسته‌بندی تصاویر در مقیاس بالا [۱] [۵۶] انجام شده است. این تصاویر شامل ۱۵۰۰۰۰ تصویر از ۱۰۰۰ دسته هستند. این ویژگی‌ها به صورت عمومی توسط نویسندگان [۳۶] در اختیار قرار گرفته است.

مشخصات مجموعه دادگان مورد استفاده به صورت خلاصه در جدول ۴-۱ آمده است.

[۱] ImageNet Large Scale Visual Recognition Challenge (ILSVRC12)

جدول ۴-۱: مشخصات مجموعه دادگان مورد استفاده در آزمایشات عملی

مجموعه داده	ابعاد توصیف	ابعاد تصاویر	دسته‌های آموزش	دسته‌های آزمون	نمونه‌های آموزش	نمونه‌های آزمون
AwA	۸۵	۴۰۹۶	۴۰	۱۰	۲۴۲۹۵	۶۱۸۰
aPY	۶۴	۴۰۹۶	۲۰	۱۲	۱۲۶۹۵	۲۶۴۴
CUB-۲۰۱۱	۳۱۲	۴۰۹۶	۱۵۰	۵۰	۸۸۵۵	۲۹۳۳
SUNA	۱۰۲	۴۰۹۶	۷۰۷	۱۰	۱۴۱۴۰	۲۰۰

۴-۲ نحوه‌ی اعتبارسنجی

برای تعیین پارامترهای مورد استفاده در روش‌های ارائه شده، از یک الگوریتم اعتبار سنجی مرسوم در روش‌های یادگیری صفرضرب استفاده می‌شود. پارامترهای موجود در روش‌ها عبارتند از:

- پارامتر β در رابطه (۳-۲). این پارامتر که در شبکه عصبی چندوظیفه‌ای پیشنهاد شده به کار رفته و نشان‌دهنده میزان تاثیر نمونه‌های آزمون در تابع هزینه است.
- مقدار α در رابطه (۳-۲۲) که وزن جمله‌ی منظم‌سازی را در یادگیری نگاشت از فضای توصیف دسته‌ها به فضای تصاویر تعیین می‌کند.
- مقادیر λ و η در رابطه (۳-۲۵) که به ترتیب میزان اهمیت جمله مربوط به نمونه‌های آزمون و وزن جمله‌ی منظم‌سازی را در یادگیری نگاشت از فضای توصیف دسته‌ها به فضای تصاویر تعیین می‌کنند.

در این شیوه‌ی اعتبار سنجی تعدادی از دسته‌های آموزش به عنوان دسته‌های اعتبارسنجی در نظر گرفته شده و اعتبار سنجی به این صورت انجام می‌شود که آموزش روی سایر دسته‌ها صورت گرفته و روی دسته‌های اعتبارسنجی که دیده نشده فرض شده‌اند، سنجیده می‌شود. بدیهی است که مجموعه دسته‌های آزمون اصلی در این روند به هیچ صورتی مورد استفاده قرار نمی‌گیرند. وقتی مقادیر پارامترها تعیین شد، روش روی کل دسته‌های دیده‌شده آموزش می‌بیند. ما تعداد دسته‌های اعتبارسنجی را برای هر مجموعه به گونه‌ای انتخاب کردیم که نسبت تعداد دسته‌های اعتبارسنجی به سایر دسته‌های آموزش برابر نسبت تعداد دسته‌های آزمون به کل دسته‌های آموزش باشد. برای اعتبار سنجی الگوریتم به ازای هر مقدار پارامتر ۱۰ بار با انتخاب تصادفی دسته‌های اعتبارسنجی از دسته‌های آزمون اجرا شده و عمل‌کرد روی این ۱۰ حالت میانگین گرفته

شده است.

۴-۳ معیار سنجش روش‌ها

معیار مورد استفاده برای این مقایسه که پرکاربردترین معیار در این زمینه است، دقت دسته‌بندی چنددسته‌ای[۲] است که به این صورت تعریف می‌شود. فرض کنید برچسب‌های صحیح نمونه‌های آزمون را با $l_۱, l_۲, \ldots, l_m$ و برچسب‌های پیش‌بینی شده برای آن‌ها را با $p_۱, p_۲, \ldots, p_m$ نشان دهیم که $l_i, p_i \in \mathbb{N}$. این معیار تعداد پیش‌بینی‌های درست را نسبت به تعداد کل پیش‌بینی‌های انجام شده نشان می‌دهد. اگر برای نمایش آن از نماد MCA استفاده کنیم، داریم:

$$MCA = \frac{\sum_{i=۱}^{m} \mathbb{1}(l_i = p_i)}{m}. \qquad (۱-۴)$$

۴-۴ پیش‌بینی صفت با شبکه عصبی چند وظیفه‌ای

در این بخش، شبکه‌ی عصبی معرفی شده در بخش ۳-۲ با سایر روش‌های پیش‌بینی صفت مقایسه می‌کنیم. ساختار شبکه مورد استفاده به این صورت است که ابتدا تصویر برای استخراج ویژگی به ۱۷ لایه با وزن‌های منجمد که در جریان آموزش قرار نمی‌گیرند وارد می‌شود. این ۱۷ لایه از شبکه ۱۹ لایه‌ی vgg که در بخش ۳-۱ شرح داده شد، گرفته شده‌اند. وزن‌های این لایه‌ها با پیش آموزش روی یک زیرمجموعه از مجموعه دادگان ImageNet مربوط به ILSVRC12 بدست آمده است. بعد از این ۱۷ لایه یک یا دو لایه با اتصالات کامل به کار گرفته شده است. اندازه خروجی لایه‌ی آخر همواره باید برابر با ابعاد توصیف‌ها باشد. بنابراین در هنگام استفاده از تنها یک لایه، اندازه این لایه برابر $(a \rightarrow ۴۰۹۶)$ خواهد بود. هنگام استفاده از دو لایه اندازه خروجی لایه میانی را نیز برابر با تعداد ابعاد توصیف‌ها در نظر گرفته‌ایم، در نتیجه در این حالت ابتدا یک لایه با ابعاد $(a \rightarrow ۴۰۹۶)$ سپس یک لایه با اتصالات کامل دیگر به ابعاد $(a \rightarrow a)$ به کار گرفته شده است. برای جلوگیری از بیش‌برازش، میان این دو لایه با اتصالات کامل از یک لایه‌ی حذف تصادفی[۳] [۵۷] با احتمال ۰/۵ نیز استفاده شده است. نتایج مربوط به حالت اول و دوم در جدول ۴-۲ به ترتیب با عناوین یک لایه و دو لایه مشخص شده‌اند.

مشاهده می‌شود که حالت یک‌لایه نتایج بهتری نسبت به شبکه دولایه کسب کرده است. برای این تحلیل این موضوع

[۲] Mulit-Class Accuracy

[۳] dropout

باید توجه کرد که تنها یک بهینه برای نگاشت یک لایه وجود دارد ولی نگاشت دو لایه دارای بهینه‌های محلی متعدد است. از طرفی با توجه به ۱۶ لایه‌ی پیچشی مورد استفاده پیش از این لایه‌های با اتصالات کامل یک فضای ویژگی غنی را بوجود می‌آورد که پیش‌بینی با تنها یک لایه هم امکان‌پذیر است و نگاشت بهینه در این حالت بدون مشکل پیدا می‌شود این در حالی است که برای حالت دولایه با وجود بهینه‌های محلی متعدد یافتن نگاشتی که عملکرد مشابه حالت یک لایه داشته باشد با تعداد محدود نمونه‌های آموزش امکان‌پذیر نیست.

تابع فعال‌سازی برای مجموعه دادگان AwA و CUB-2011 که مقادیر بردارهای صفات در آن‌ها حقیقی است، تابع ReLU در نظر گرفته شده است. برای مجموعه دادگان SUN و aPY مقادیر بردارهای صفات برای نمونه‌های آن‌ها دودویی بوده و در نتیجه مقادیر بردارهای صفات برای دسته‌ها که میانگین این بردارها برای نمونه‌هاست در بازه $[۰, ۱]$ قرار می‌گیرد. در نتیجه از تابع فعال‌سازی سیگموید استفاده شده تا مقادیر در این فاصله قرار بگیرند.

اندازه دسته‌ها[۴] در جریان آموزش برابر ۱۲۸ در نظر گرفته شده است. پیش از آموزش شبکه به صورت کامل، از یک روند پیش‌آموزش استفاده کرده‌ایم که در آن تنها نمونه‌های آموزش به شبکه وارد شده و خروجی با توصیف صحیح آن‌ها مقایسه می‌شود (نیمه‌ی چپ تصویر ۳-۲). تعداد تکرارها در جریان پیش آموزش ۱۵ و در آموزش کلی شبکه ۳۰ در نظر گرفته شده است چرا که روند همگرایی در همین تعداد تکرار اتفاق می‌افتد و افزایش تکرارها تاثیری در بهبود نتایج ندارد. جهت آموزش شبکه برای مجموعه دادگان AwA و CUB-2011 از الگوریتم بهینه‌سازی adam [۵۸] استفاده شده است. برای مجموعه دادگان SUN و aPY الگوریتم adadelta [۵۹] مورد استفاده قرار گرفته است.

در این بخش هم‌چنین برای روشن‌تر شدن تاثیر استفاده از نمونه‌های بدون برچسب آزمون و اطلاعات بدون نظارت موجود در ساختار ویژگی‌های ژرف استخراج شده از تصاویر، نتایج مربوط به مدل پایه‌ی شرح داده شده در بخش ۳-۲-۳ نیز گزارش شده است. ساختار و تنظیمات مورد استفاده برای شبکه حالت پایه کاملا مشابه شبکه چندوظیفه‌ای در نظر گرفته شده است. یعنی تعداد لایه‌های و اندازه هرلایه و هم‌چنین تابع فعال‌سازی مورد استفاده برای مجموعه دادگان مختلف و هم چنین اندازه دسته مانند حالت قبل است. تعداد تکرارها در جریان آموزش برای شبکه ساده ۸۰ تکرار در نظر گرفته شده است. نتایج مربوط به این شبکه در جدول ۴-۲ با عنوان *شبکه پایه* آمده است.

پیاده‌سازی این شبکه با استفاده از ابزارهای متن باز Theano [۶۰] و Keras [۶۱] صورت گرفته است و برای اجرای آن‌ها از پردازنده گرافیکی Nvidia Titan Black با ۶ گیگابایت حافظه گرافیکی استفاده شده است. زمان اجرای

[۴] Batch Size

جدول ۴-۲: مقایسه دقت دسته‌بندی چنددسته‌ای روش پیشنهادی با سایر روش‌ها. جدول شامل دقت دسته‌بندی چنددسته‌ای به صورت (میانگین± انحراف معیار) است. نتایج سایر روش‌ها از مقالاتی که روش در آن‌ها ارائه شده نقل شده و آزمایش‌ها توسط ما تکرار نشده است. خانه‌هایی که از جدول با – مشخص شده‌اند به معنای عدم ارائه نتایج روش برای مجموعه‌دادگان مربوطه در مقاله اصلی است.

روش	AwA	CUB-۲۰۱۱	aPY	SUNA
Jayaraman and Grauman [۴۳]	۴۳/۰۱ ± ۰/۰۷	–	۲۶/۰۲ ± ۰/۰۵	۵۶/۱۸ ± ۰/۲۷
Lampert et al (DAP) [۱۳]	۴۱/۴	–	۱۹/۱	۲۲/۲ ± ۱/۶
Lampert et al (IAP) [۱۳]	۴۲/۲	–	۱۶/۹	۱۸/۰ ± ۱/۵
Akata et al [۲۵]	۳۷/۴	۱۸/۰	–	–
شبکه پایه (بخش ۳-۲-۳) ـ یک لایه	۵۶/۷۸ ± ۱/۲۹	۳۲/۶۰ ± ۰/۸۲	۲۴/۵۷ ± ۱/۳۶	۵۸/۳۳ ± ۱/۵۲
شبکه پایه (بخش ۳-۲-۳) دو لایه	۵۲/۱۴ ± ۰/۳۱	۳۱/۶۵ ± ۰/۴۱	۲۲/۵۶ ± ۱/۲۹	۶۲/۰۰ ± ۲/۶۴
شبکه چندوظیفه‌ای (بخش ۳-۲) ـ یک لایه	**۷۴/۵۲ ± ۱/۹۳**	**۳۳/۹۱ ± ۰/۲۱**	**۳۳/۱۰ ± ۱/۳۶**	**۶۶/۱۳ ± ۰/۵۰**
شبکه چندوظیفه‌ای (بخش ۳-۲) ـ دو لایه	۵۷/۱۰ ± ۰/۴۷	۳۱/۲۷ ± ۰/۸۷	۲۲/۳۲ ± ۰/۴۸	۶۶/۸۳ ± ۱/۵۲

الگوریتم برای مجموعه داده‌های مورد استفاده در همه موارد کمتر از ۳۰ دقیقه بوده است.

جدول ۴-۲ دقت دسته‌بندی چند دسته‌ای با استفاده از این روش را به همراه نتایج سایر روش‌های با رویکرد پیش‌بینی صفت نشان می‌دهد. همان‌طور که مشاهده می‌شود، استفاده از این شبکه عمل‌کرد بهتری نسبت به سایر روش‌های پیش‌بینی صفت داشته است.

۴-۴-۱ استفاده از تابع مطابقت پیشنهادی

همان‌طور که در بخش ۳-۴ عنوان شد تابع مطابقت پیشنهادی در این پژوهش قابلیت اضافه شدن به سایر روش‌های موجود که از دسته‌بند نزدیک‌ترین همسایه یا سنجش مطابقت با ضرب داخلی در یک فضای مشترک استفاده می‌کنند را دارد و می‌تواند نتایج آن‌ها را بهبود دهد. در این بخش به عنوان نمونه این تابع مطابقت را به روش مبتنی بر شبکه عصبی چندوظیفه‌ای ارائه شده اضافه می‌کنیم. این کار به این صورت انجام می‌شود که پس از انجام پیش‌بینی نهایی شبکه عصبی، یک خوشه‌بندی با الگوریتم k-means روی مجموعه داده‌های آزمون انجام می‌شود که در آن $k = ۲n_u$. سپس با استفاده

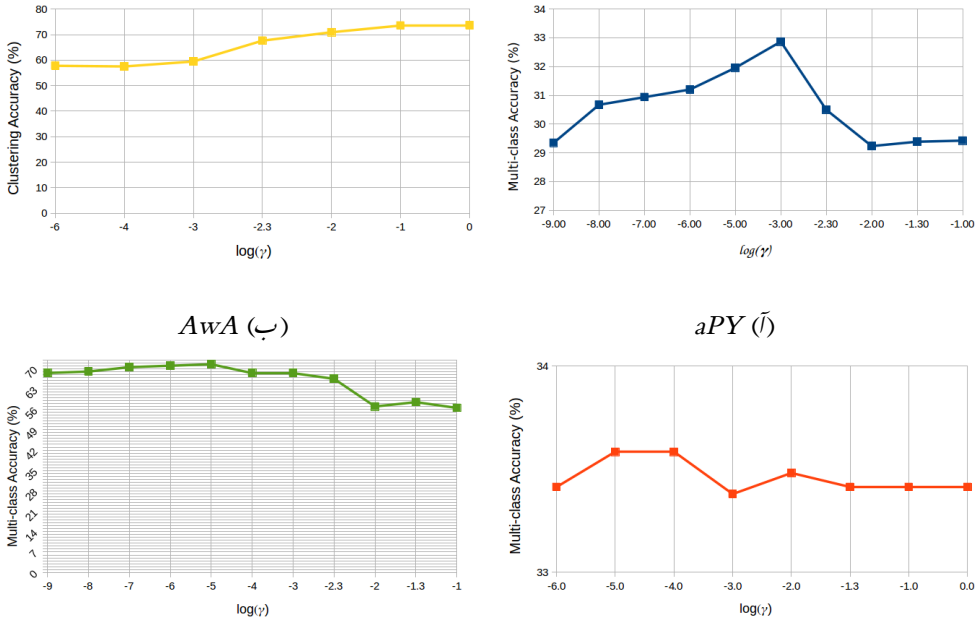
جدول ۴-۳: مقایسه دقت دسته‌بندی (٪) شبکه عصبی پیشنهادی در حالت استفاده از دسته‌بند نزدیکترین همسایه با حالتی که تابع مطابقت پیشنهادی بخش ۳-۴ برای تخصیص برچسب استفاده می‌شود. نتایج ذکر شده برای حالت استفاده از تابع مطابقت دقیقا بر روی پیش‌بینی‌های متناظرشان در حالت استفاده از دسته‌بند نزدیک‌ترین همسایه در سطر بالا اعمال شده‌اند.

روش	AwA	CUB-۲۰۱۱	aPY	SUNA
شبکه چندوظیفه‌ای ـ نزدیکترین همسایه	۷۴/۵۲ ± ۱/۹۳	۳۳/۹۱ ± ۰/۲۱	۳۳/۱۰ ± ۱/۳۶	۶۶/۱۳ ± ۰/۵۰
شبکه چندوظیفه‌ای ـ تابع مطابقت پیشنهادی	۷۴/۶۸ ± ۰/۷۳	۳۳/۹۲ ± ۰/۰۷	۳۸/۲۶ ± ۱/۲۷	۶۷/۵۰ ± ۰/۰۰

از عملیات رای‌گیری روی پیش‌بینی‌های روش قبل، با استفاده از رابطه (۳-۱۴) به هر خوشه یک برچسب تعلق می‌گیرد. حاصل اجرای چنین روندی در جدول ۴-۳ آمده است. سطر اول این جدول دقت دسته‌بندی را در حالت عادی که تنها از دسته‌بند نزدیک‌ترین همسایه برای تخصیص برچسب استفاده می‌شود، نشان می‌دهد. سطر دوم دقت دسته‌بندی را در حالتی که تابع مطابقت پیشنهادی روی همان خروجی‌های مربوط به سطر اول اجرا شده است. همان‌طور که مشاهده می‌شود استفاده از این تابع مطابقت در همه موارد باعث بهبود نتایج شده‌است. دلیل این موضوع استفاده از اطلاعات نیمه‌نظارتی موجود در نمونه‌های آزمون و اجباری شدن هم‌برچسب بودن نمونه‌های مشابه در یک خوشه است. این مسئله با توجه به ساختار غنی موجود در ویژگی‌های ژرف استخراج شده از تصاویر باعث می‌شود نمونه‌هایی که پیش از این با دسته‌بند نزدیک‌ترین همسایه اشتباه دسته‌بندی می‌شدند حال چون اکثریت نمونه‌های موجود در خوشه‌ی آن‌ها برچسب صحیح دریافت کرده‌اند، این نمونه‌ها نیز که همان برچسب را دریافت می‌کنند در دسته‌ی صحیح دسته‌بندی شوند. شبکه مورد استفاده در این آزمایش، حالت یک لایه‌ی همان شبکه معرفی شده در ابتدای این بخش است.

۴-۴-۲ تحلیل پارامتر

در این بخش به تحلیل تاثیر پارامتر γ در رابطه (۳-۲) می‌پردازیم. این پارامتر وزن جمله‌ی اضافه شده به شبکه‌ی پایه که برای تضمین شباهت خروجی شبکه روی نمونه‌های آزمون به بردار توصیف یکی از دسته‌های آزمون به کار می‌رود و مقدار آن در جریان آموزش با اعتبارسنجی تعیین می‌شود. تاثیر مقدار این پارامتر بر دقت نهایی دسته‌بندی در تصویر ۴-۱ آمده است.



(ب) *AwA*

(آ) *aPY*

(د) *SUNA*

(ج) *CUB-۲۰۱۱*

شکل ۴-۱: میزان دقت دسته‌بندی چند دسته‌ای در شبکه چندوظیفه‌ای ارائه شده (نسخه یک لایه) بر حسب $\log_{۱۰}$ *پارامتر* γ *در معادله* (۳-۲).

۴-۵ بررسی خوشه‌بندی نیمه‌نظارتی

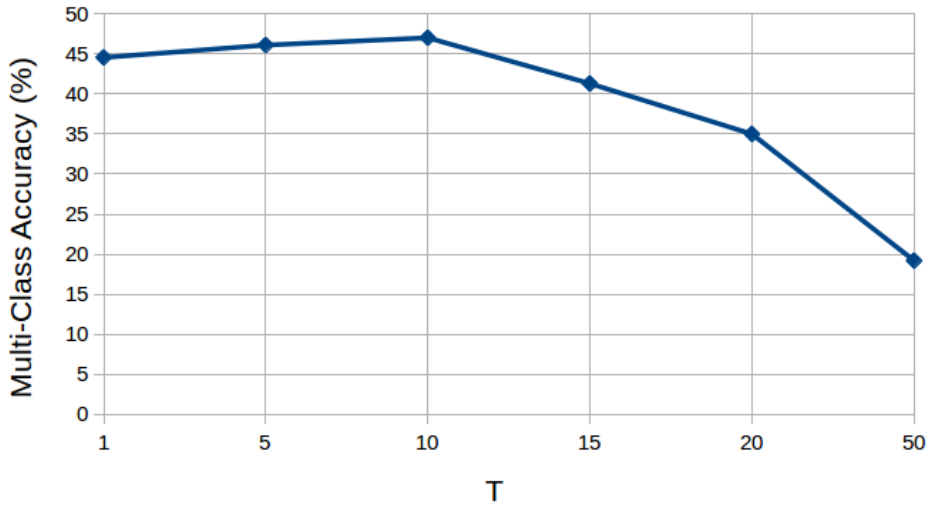
در این بخش به بررسی عمل‌کرد روش خوشه‌بندی نیمه‌نظارتی ارائه شده در بخش ۳-۵ می‌پردازیم. برای این منظور روش ارائه شده را روی هر مجموعه داده اجرا کرده، خوشه‌های مربوط به دسته‌های دیده‌شده را کنار گذاشته و هر یک از خوشه‌های دیگر را به یک دسته از دسته‌های آزمون نسبت می‌دهیم. برای این کار در هر خوشه بر اساس برچسب صحیح نمونه‌ها رای‌گیری می‌شود و برچسبی که بیشتر اعضای آن خوشه آن را دارا هستند به کل اعضای خوشه نسبت داده می‌شود. نتیجه با برچسب‌های صحیح مقایسه شده و دقت دسته‌بندی چنددسته‌ای در جدول ۴-۴ گزارش شده است. برای مقایسه عمل‌کرد، آزمایش مشابهی را با روش k-means اجرا می‌کنیم. به این صورت که الگوریتم k-means را با $k = n_s + ۲n_u$ اجرا کرده و با هر خوشه با رای‌گیری برچسب یکی از دسته‌های دیده نشده را نسبت می‌دهیم. نتایج مربوط به این آزمایش نیز در جدول ۴-۴ گزارش شده است.

جدول ۴-۴: امتیاز معیار دقت (٪) تخصیص خوشه‌ها که با رای‌گیری روی برچسب‌های صحیح به شماره دسته تبدیل شده است؛ بر روی چهار مجموعه داده مورد استفاده در یادگیری صفرضرب. نتایج روش پیشنهادی به صورت میانگین ± انحراف معیار برای سه اجرا گزارش شده‌است.

روش خوشه‌بندی	AwA	CUB-۲۰۱۱	aPY	SUNA
k-means	۶۵/۹۳ ± ۱/۷۳	۳۴/۴۸ ± ۱/۰۰	۶۵/۳۷ ± ۳/۷۳	۱۶/۸۳ ± ۰/۷۶
خوشه‌بندی نیمه‌نظارتی (بخش ۳-۵)	۷۰/۷۴ ± ۰/۳۲	۴۲/۶۳ ± ۰/۰۷	۶۹/۹۳ ± ۳/۴۰	۴۵/۵۰ ± ۱/۳۲

۴-۶ نگاشت به هیستوگرام دسته‌های دیده‌شده با شبکه عصبی

در این بخش به ارائه جزییات پیاده‌سازی و تنظیمات مورد استفاده برای بررسی شبکه‌ی ژرف معرفی شده در بخش ۳ـ ۳ می‌پردازیم. در شبکه‌ی عصبی مورد استفاده در این روش، از چهار لایه با اتصالات کامل بعد از لایه‌های پیچشی برگرفته شده از شبکه vgg استفاده شده است. با توجه به این که این شبکه با معیار دسته‌بندی نمونه‌ها در دسته‌های دیده شده آموزش می‌بیند، اندازه لایه‌ی آخر الزاما باید برابر تعداد دسته‌های دیده شده در هر مجموعه دادگان باشد. اندازه سه لایه‌ی قبل از آن برای هر چهار مجموعه دادگان مورد آزمایش برابر ۱۲۰ عدد درنظر گرفته شده است. برای جلوگیری از بیش‌برازش، میان هر دولایه با اتصالات کامل از یک لایه‌ی حذف تصادفی [۵۷] استفاده شده؛ احتمال حذف تصادفی در این لایه‌ها برابر ۰/۴ در نظر گرفته است. تابع فعال‌سازی لایه‌ی نهایی نسخه‌ای از تابع softmax است که در رابطه (۳-۱۱) معرفی شد. در زمان آموزش این تابع به ازای $T = ۱$ استفاده می‌شود و در زمان آزمون از $T = ۱۵$ استفاده شده است. حساسیت عمل‌کرد شبکه نسبت به مقدار این پارامتر برای مجموعه دادگان aPascal/aYahoo در تصویر ۴-۲ مورد بررسی قرار گرفته است. مشاهده می‌شود که افزایش مقدار T در ابتدا با هموارتر کردن هیستوگرام حاصل باعث افزایش دقت دسته‌بندی شود اما با ادامه افزایش آن مقادیر هیستوگرام حاصل بسیار به یکدیگیر نزدیک شده و اطلاعات موجود در آن از بین می‌رود در نتیجه دقت دسته‌بندی کاهش می‌یابد. در سایر لایه‌ها تابع فعال‌سازی ReLU به کار گرفته شده است. آموزش شبکه مطابق با حالت معمول دسته‌بندی با شبکه‌های عصبی با تابع هزینه آنتروپی متقاطع میان خروجی شبکه و برچسب صحیح (با کدگذاری یکی‌یک) صورت گرفته است. الگوریتم بهینه‌سازی مورد استفاده برای آموزش شبکه، الگوریتم adadelta [۵۹] است. تعداد تکرارها در آموزش شبکه حداکثر ۸۰ تکرار درنظر گرفته شده است. مدت زمان آموزش شبکه با استفاده از پردازنده گرافیکی NVIDIA Geforce Titan Black در تمامی آزمایش‌ها کمتر



شکل ۴-۲: بررسی میزان دقت دسته‌بندی بر حسب پارامتر T در رابطه (۳-۱۱) برای مجموعه دادگان aPascal/aYahoo: افزایش T در ابتدا می‌تواند باعث افزایش دقت شود ولی ادامه افزایش آن باعث نزدیک شدن مقادیر هیستوگرام به یکدیگر و کاهش دقت دسته‌بندی می‌شود.

از ۵ دقیقه بوده است.

نتایج مربوط به این روش در جدول ؟؟ آمده است با عنوان *نگاشت به هیستوگرام* آمده است. همان‌گونه که مشاهده می‌شود این روش با اینکه از روند ساده و همچنین سریعی بخاطر استفاده از الگوریتم‌های بهینه‌سازی تصادفی برخوردار است، به نتایج بهتری نسبت به روش‌های پیشین دست یافته است و تنها از روش بسیار اخیر ارائه شده در [۳۷] دقت کمتری داشته است. این در حالی‌ست که در سایر روش‌های مبتنی بر هیستوگرام ([۳۶، ۳۷] از روندهای بهینه‌سازی همراه با محدودیت استفاده می‌شود که بسیار کندتر هستند. برای مثال حداکثر زمان اجرا در [۳۶] روی چهار مجموعه دادگان مورد بررسی ۳۰ دقیقه اعلام شده است در حالی‌که در آزمایشات انجام شده زمان آموزش شبکه پیشنهادی کمتر از ۵ دقیقه بوده است. همچنین به علت محدودیت‌های روش‌های بهینه‌سازی محدب، این روش‌ها در مجموعه دادگان بزرگ مانند ImageNet قابل استفاده نیستند در حالی‌که روش پیشنهادی دارای قابلیت مقیاس‌پذیری و استفاده در مجموعه دادگان بزرگتر است.

۴-۷ دسته‌بندی با روش خوشه‌بندی و یادگیری نگاشت مجزای نیمه‌نظارتی

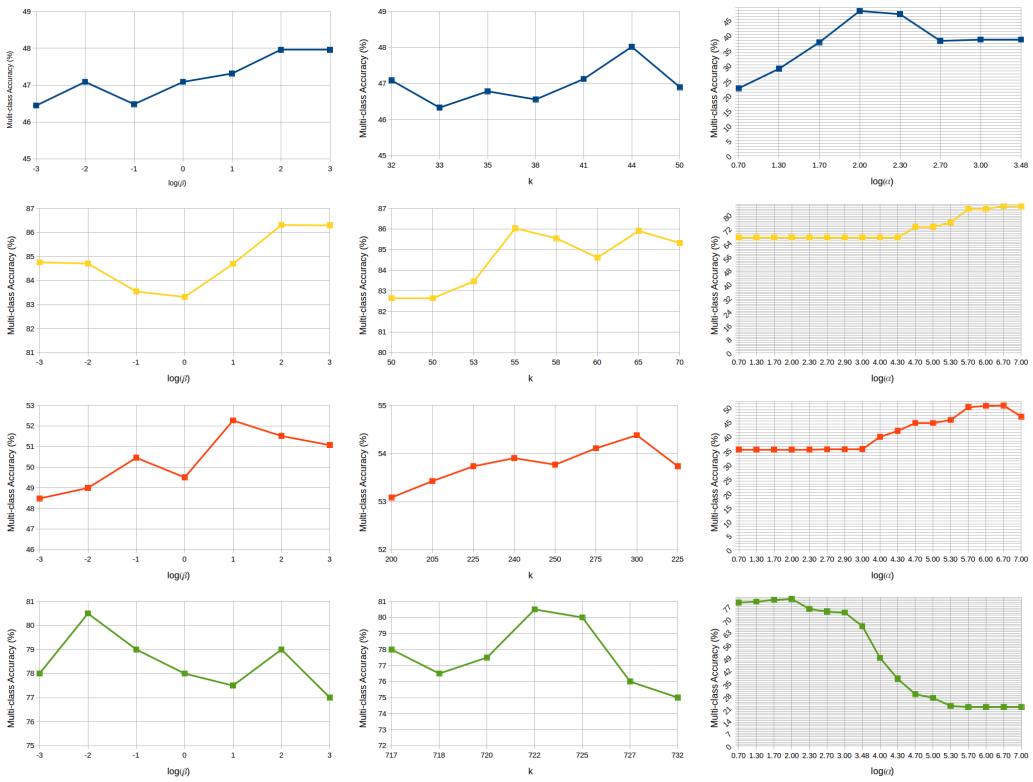
در این بخش به بررسی عملی روش پیشنهادی برای روش خوشه‌بندی و یادگیری نگاشت مجزای نیمه‌نظارتی می‌پردازیم که در بخش ۳-۶ معرفی شد و مراحل آن در الگوریتم ۲ ذکر شده است. این روش مبتنی بر یک خوشه‌بندی روی داده‌های آزمون بود و با استفاده از یک نگاشت خطی از فضای توصیف دسته‌ها به فضای تصاویر، مرکز هر خوشه را به یک دسته‌ی دیده نشده منتسب می‌کرد. بر اساس تابع مطابقت پیشنهادی (بخش ۳-۴)، تمام اعضای هر خوشه همان برچسبی که مرکزشان دریافت کرده را دریافت می‌کند.

این روش با استفاده از دو نوع خوشه‌بندی آزمایش شده است. یکی خوشه‌بندی نیمه‌نظارتی پیشنهادی که نتایج این حالت با عنوان *پیشنهادی* (*خوشه‌بندی نیمه‌نظارتی* + *تابع مطابقت*) در جدول ۴-۵ آمده است. برای بررسی تاثیر خوشه‌بندی ارائه شده یک نسخه دیگر از این روش که در آن از خوشه‌بندی k-means بجای خوشه‌بندی پیشنهادی استفاده شده است نیز مورد آزمایش قرار گرفته است. نتایج مربوط به این روش با عنوان *پیشنهادی* (*تابع مطابقت* + k-*means*) آمده است. نتایج ارائه شده حاصل سه بار اجرا هستند که به صورت *میانگین* $\pm$ *انحراف معیار* بیان شده‌اند. همان‌گونه که از نتایج مشخص است، استفاده از خوشه‌بندی نیمه‌نظارتی ارائه شده همواره نتایج بهتری نسبت به استفاده از خوشه‌بندی k-means تولید خواهد کرد.

تاثیر پارامترهای مورد استفاده در این قسمت در شکل ۴-۳ آمده است. همان‌طور که مشاهده می‌شود پارامتر η در رابطه (۳-۲۳) تاثیر قابل توجهی بر دقت دسته‌بندی نهایی دارد، در نتیجه ما مقدار این پارامتر را با استفاده از روند اعتبارسنجی شرح داده شده در بخش ۴-۲ تنظیم کرده‌ایم. از طرف دیگر مشاهده می‌شود تعداد خوشه‌ها در خوشه‌بندی نیمه‌نظارتی ارائه شده تاثیر قابل توجهی بر دقت دسته‌بندی ندارد، در نتیجه برای سادگی و کاهش زمان روند آموزش ما این پارامتر را همان‌طور که در بخش ۳-۵ شرح داده شد با استفاده از یک قاعده سرانگشتی بر حسب تعداد دسته‌ها تعیین می‌کنیم که تعداد خوشه‌ها برای هر مجموعه داده‌گان برابر $k = n_s + ۲n_u$ در نظر گرفته می‌شود.

۴-۸ خوشه‌بندی و یادگیری نگاشت توام

روش پیشنهادی دوم که در بخش ۳-۷ ارائه شد به خوشه‌بندی و یادگیری نگاشت توام می‌پرداخت و برچسب نمونه‌های آزمون در آن به طور مستقیم در جریان آموزش بدست می‌آید. تنظمیات آزمایش برای روش خوشه‌بندی و نگاشت توام مانند



شکل ۴-۳: تاثیر پارامترهای روش خوشه‌بندی و یادگیری نگاشت مجزای نیمه‌نظارتی. ***سمت چپ:*** *نتیجه دقت دسته‌بندی چند دسته‌ای بدست آمده بر حسب پارامتر α در رابطه (۳-۲۳) که اهمیت جمله منظم‌سازی را نشان می‌دهد. همان‌طور که مشاهده می‌شود، عمل‌کرد روش به این پارامتر حساس است.* ***وسط :*** *نتیجه دقت دسته‌بندی چند دسته‌ای بدست آمده بر حسب تعداد خوشه‌ها در خوشه‌بندی نیمه‌نظارتی. با توجه مقیاس این نمودار مشخص می‌شود که دقت حاصل شده حساسیت کمی نسبت به این پارامتر دارد.* ***سمت راست:*** *نتیجه دقت دسته‌بندی چنددسته‌ای بر حسب پارامتر β در خوشه‌بندی نیمه نظارتی (رابطه (۳-۱۷)).*

برای راحتی مقایسه محور عمودی همه‌ی نمودارها با بازه‌های یک درصدی تقسیم‌بندی شده‌اند.

سطر اول (آبی‌رنگ): مجموعه دادگان aPY . سطر دوم (زرد رنگ): مجموعه دادگان AwA . سطر سوم (قرمز رنگ): مجموعه دادگان $CUB\text{-}2011$. سطر چهارم (سبز رنگ): مجموعه دادگان $SUNA$.

حالت قبل سه بار اجرا و گزارش نتایج به صورت *میانگین* $\pm$ *انحراف معیار* است. دو نوع مقداردهی اولیه انجام شده است. یکی همان‌طور که در بخش ۳-۷ بیان شد، مقداردهی R که با استفاده از الگوریتم ۲ انجام می‌شود. نتایج مربوط به این حالت در جدول ۴-۵ با عنوان *پیشنهادی (توام، مقداردهی R)* آمده‌اند. یک مقداردهی دیگر شروع بهینه‌سازی تناوبی در الگوریتم ۳ با مقداردهی D است که توسط رابطه (۳-۲۳) صورت گرفته است. نتایج مربوط به این حالت با عنوان *پیشنهادی (توام، مقداردهی D)* آمده‌اند. مقایسه نتایج مربوط به این دو نحوه‌ی مقداردهی اولیه نشان می‌دهد

که استفاده از روش پیشنهادی الگوریتم ۲ برای رسیدن به دقت بالا ضروری است، چرا که مشاهده می‌شود که استفاده از مقداردهی اولیه برای R به صورت بیان شده در الگوریتم ۳ به طور متوسط ۶/۸٪ دقت بالاتری در دسته‌بندی نسبت به مقداردهی D با رابطه (۳-۲۳) دارد. دلیل این موضوع همان‌طور که در بخش ۳-۷ بیان شد استفاده از اطلاعات بدون نظارت نمونه‌های آزمون در بدست آوردن مقدار اولیه برای R است در حالیکه در مقداردهی اولیه D تنها نمونه‌های آموزش دخالت دارند.

به علت حساسیت نتایج این روش به پارامترهای آن (مقادیر λ و η در رابطه (۳-۲۵))، مقادیر آن‌ها توسط روند اعتبارسنجی شرح داده شده در بخش ۴-۲ تنظیم می‌شود.

۴-۸-۱ روش‌های مورد مقایسه

در این بخش قصد داریم روش‌های پیشنهادی در بخش‌های ۳-۶ و ۳-۷ را با مطرح‌ترین روش‌های اخیر در حوزه یادگیری صفرضرب مقایسه کنیم. سایر روش‌هایی که در جدول ۴-۵ برای مقایسه آورده شده‌اند، روش‌هایی هستند که بالاترین دقت‌های دسته‌بندی را در دسته‌بندی صفرضرب با استفاده از توصیف‌های به صورت بردار صفت دارا هستند. روش‌های ارائه شده در [۴۰، ۴۴، ۴۱] از این جهت که *نیمه‌نظارتی* هستند، یعنی از نمونه‌های آزمون نیز در زمان آموزش استفاده می‌کنند، با روش‌های ما بیشترین نزدیکی را دارند. البته در [۴۰، ۴۴] از ویژگی‌های کم‌عمق برای تصاویر استفاده شده است که توانایی جداسازی دسته‌ها در آن بسیار پایین‌تر از ویژگی‌های بدست آمده از شبکه‌های عصبی ژرف است که در روش‌های پیشنهادی ما مورد استفاده قرار گرفته است. روش‌های [۲۷، ۳۰] با استفاده از توابع هزینه‌ی بیشترین حاشیه سعی در یادگیری نگاشت از هر دو فضای تصاویر و توصیف دسته‌ها به فضای مشترک دارند. این روش‌ها از ویژگی‌های شبکه‌ی ژرف GoogleNet [۶۲] برای استخراج ویژگی استفاده می‌کنند. ابعاد ویژگی‌های بدست آمده ۱۰۲۴ است که بعد کمتری نسبت به ویژگی‌های ۴۰۹۶ ـ بعدی استخراج شده از شبکه ۱۹ لایه‌ی vgg دارد و توانایی جداسازی دسته‌ها در آن پایین‌تر است. همان‌طور که مشاهده می‌شود استفاده از این ویژگی‌های با بعد بیشتر عملکرد روش ارائه شده در [۲۷] را بهبود داده است.

روش‌هایی که بهترین نتایج را در میان روش‌های رقیب کسب کرده‌اند، روش ارائه شده در [۳۶] و تعمیم آن در [۳۷] هستند. هرچند این روش‌ها نیمه‌نظارتی نیستند و تنها از نمونه‌های آموزش برای یادگیری نمایش تصاویر و توصیف دسته‌ها در یک فضای مشترک، که فضای هیستوگرام دسته‌های دیده شده است استفاده می‌کنند، نتایج بهتری نسبت به

روش‌های نیمه‌نظارتی پیشین در [۴۰، ۴۴، ۴۱] کسب کرده‌اند. این مسئله می‌توان نشانگر یک مسیر مناسب در ترکیب روش پیشنهادی در این پژوهش با فضای مشترک مورد استفاده در آن روش‌ها برای کارهای آتی باشد.

جدول ۴-۵: مقایسه دقت دسته‌بندی چنددسته‌ای روش پیشنهادی با سایر روش‌ها. نتایج بر اساس نوع ویژگی مورد استفاده برای تصاویر دسته‌بندی شده‌اند. جدول شامل دقت دسته‌بندی چنددسته‌ای به صورت (میانگین ± انحراف معیار) است. نتایج سایر روش‌ها از مقالاتی که روش در آن‌ها ارائه شده نقل شده و آزمایش‌ها توسط ما تکرار نشده است. نتایج روش‌های پیشنهادی حاصل سه اجرا هستند.

ویژگی تصاویر	روش	AwA	CUB-۲۰۱۱	aPascal-aYahoo	SUN
کم‌عمق	Li and Guo [۴۰]	۳۸/۲ ± ۲/۳			۱۸/۹ ± ۲/۵
	Li *et al.* [۴۴]	۴۰/۰۵ ± ۲/۲۵		۲۴/۷۱ ± ۳/۱۹	
	Jayaraman and Grauman [۴۳]	۴۳/۰۱ ± ۰/۰۷		۲۶/۰۲ ± ۰/۰۵	۵۶/۱۸ ± ۰/۲۷
GoogleNet	Akata *et al.* [۲۷]	۶۶/۷	۵۰/۱		
	Xian *et al.* [۳۰]	۷۱/۹	۴۵/۵		
VGG-۱۹	Khodirov *et al.* [۴۱]	۷۳/۲	۳۹/۵	۲۶/۵	
	Akata *et al.* [۲۷]	۶۱/۹	۵۰/۱		
	Zhang and Saligrama [۳۶]	۷۶/۳۳ ± ۰/۵۳	۳۰/۴۱ ± ۰/۲۰	۴۶/۲۳ ± ۰/۵۳	۸۲/۵۰ ± ۱/۳۲
	Zhang and Saligrama [۳۷]	۸۰/۴۶ ± ۰/۵۳	۴۲/۱۱ ± ۰/۵۵	۵۰/۳۵ ± ۲/۹۷	۸۳/۸۳ ± ۰/۲۹
	پیشنهادی (نگاشت به هیستوگرام)	۷۶/۵۰ ± ۱/۰۲	۳۳/۲۹ ± ۰/۲۱	۴۷/۴۶ ± ۰/۳۱	۷۹/۸۸ ± ۰/۴۲
	پیشنهادی (خوشه‌بندی و یادگیری نگاشت مجزا)	۸۶/۳۴ ± ۰/۱۳	۵۲/۴۸ ± ۰/۶۰	۴۸/۰۳ ± ۱/۵۶	۷۵/۷۵ ± ۱/۰۶
	پیشنهادی (خوشه‌بندی و یادگیری نگاشت مجزای نیمه‌نظارتی)	۸۶/۳۸ ± ۰/۵۶	۵۳/۱۰ ± ۰/۴۳	۴۸/۵۲ ± ۰/۲۹	۸۰/۶۶ ± ۰/۷۶
	پیشنهادی (توام، مقداردهی D)	۸۳/۰۳	۵۷/۵۵	۴۲/۶۲	۷۲/۵۰
	پشنهادی (توام، مقداردهی R)	۸۸/۶۴ ± ۰/۰۴	۵۸/۸۰ ± ۰/۶۴	۴۹/۷۷ ± ۲/۰۲	۸۶/۱۶ ± ۰/۵۷

۴-۹ تحلیل نتایج

با توجه به جدول ۴-۵ روش پیشنهادی یادگیری توام نگاشت و خوشه‌بندی هنگام مقداردهی اولیه مقادیر R مجموعا به بهترین نتایج دست‌یافته است. این روش روی سه مجموعه داده‌گان از چهار مجموعه که روش‌ها با آن محک زده شده‌اند نتایج بهتری نسبت به سایر روش‌ها دارد و عمل‌کرد پیشگام در حوزه یادگیری صفرضرب را ارتقاء داده است.

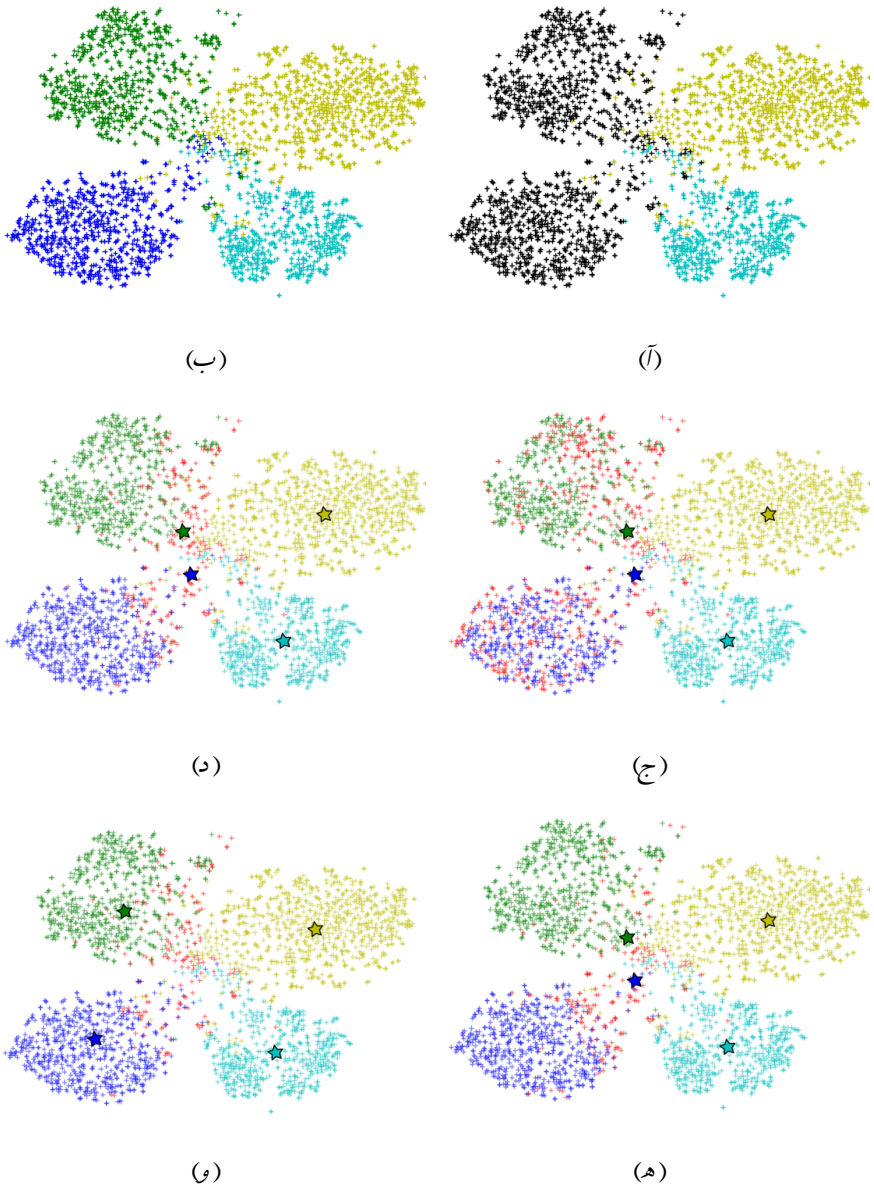
روی مجموعه دادگان aPascal/aYahoo روش ارائه شده در [۳۷] نتایج بهتری کسب کرده است. دلیل این موضوع می‌تواند شباهت بسیار زیاد میان امضای دسته‌ها در این مجموعه دادگان و عدم ایجاد جداسازی بالا میان دسته‌ها توسط این بردارهای توصیف باشد. در روش پیشنهادی یادگیری نگاشت و خوشه‌بندی توام، با توجه به نزدیکی زیاد این بردارهای توصیف، نگاشت آن‌ها در فضای ویژگی تصاویر نیز به یکدیگر نزدیک خواهد بود و جداسازی مناسبی میان نمونه‌های دسته‌های مختلف صورت نمی‌پذیرد؛ ولی در روش ارائه شده در [۳۷] همان‌گونه که در فصل دوم مرور شد، بردارهای توصیف ورودی مستقیماً به کار گرفته نمی‌شوند، بلکه از آن‌ها برای بدست آوردن نمایش دیگری برای دسته‌ها به صورت هیستوگرامی از دسته‌های دیده شده، استفاده می‌شود. وجود این گام می‌تواند مشکل نزدیکی و شباهت زیاد میان امضای دسته‌ها را از بین ببرد. هم‌چنین همان‌طور که در بخش ۳-۷-۱ بیان شد، مقداردهی اولیه مقادیر R با استفاده از روش خوشه‌بندی و تابع مطابقت پیشنهادی عمل‌کرد بهتری نسبت به مقداردهی اولیه D دارد. از این جدول هم‌چنین کارایی روش خوشه‌بندی نیمه‌نظارتی پیشنهادی نسبت به الگوریتم k-means در مسئله یادگیری بدون برد مشخص می‌شود، چرا که در همه‌ی موارد هنگام استفاده از روش خوشه‌بندی نیمه‌نظارتی پیشنهادی در مقایسه با الگوریتم k-means دقت بالاتری در دسته‌بندی حاصل شده است. هر دو حالت این روش ساده که از یک نگاشت خطی و بخش تاثیرگذارتر تابع مطابقت پیشنهادی تشکیل شده‌اند، روی نیمی از چهار مجموعه‌داده‌گان مورد بررسی عمل‌کرد بهتری نسبت به همه‌ی روش‌های پیشین داشته‌اند که نشان‌دهنده کارایی تابع مطابقت پیشنهادی است.

برای تحلیل کارایی روش قسمت‌های مختلف آن و تاثیر هر یک روی یک مجموعه داده واقعی در شکل ۴-۴ نشان داده شده است. نتایج مربوط به اجرای روش روی تمام مجموعه دادگان AwA است، ولی برای این که تغییرات در شکل قابل دنبال کردن باشند تنها چهار دسته در تصویر نشان داده شده‌اند که دو دسته از آن‌ها دسته‌های دیده شده و دو دسته از دسته‌های دیده نشده هستند. در تصویر ۴-۴آ دسته‌های دیده شده به صورت رنگی و دسته‌های دیده نشده با رنگ سیاه مشخص شده‌اند. در تصویر ۴-۴ب برچسب‌های صحیح برای دسته‌های دیده نشده نیز با رنگ مشخص شده است. در تصویر ۴-۴ج توصیف دسته‌ها با استفاده از نگاشت D از رابطه (۳-۲۳) به فضای تصاویر برده شده (نماد ستاره) و سپس نمونه‌های آزمون با استفاده از دسته‌بند نزدیکترین همسایه دسته‌بندی شده‌اند، نمونه‌هایی که رنگ قرمز دارند به دسته‌ای غیر از چهار دسته‌ی موجود در تصویر دسته‌بندی شده‌اند. تصویر ۴-۴د حاصل دسته‌بندی به شیوه‌ی روش خوشه‌بندی و یادگیری نگاشت مجزای نیمه‌نظارتی ارائه شده در بخش ۳-۶ است که در آن از خوشه‌بندی k-means و تابع مطابقت پیشنهادی استفاده شده است. تصویر ۴-۴ه مشابه حالت قبل است با این تفاوت که در آن از خوشه‌بندی نیمه‌نظارتی پیشنهادی به‌جای k-means استفاده شده است. در تصویر ۴-۴و دسته‌بندی و یادگیری نمایش توصیف دسته‌ها در فضای

تصاویر (ستاره‌ها) به صورت توام با روش پیشنهادی بخش ۳-۷ صورت گرفته است. همان‌طور که در تصاویر ۴-۴د و ۴-۴ه مشخص است، استفاده از تابع مطابقت معرفی شده در بخش ۳-۴ برای دسته‌بندی بسیار موفق‌تر از دسته‌بند نزدیک‌ترین همسایه عمل می‌کند و اطلاعات غیر نظارتی موجود در نمونه‌های آزمون دقت دسته‌بندی را بهبود می‌دهد. هم‌چنین برتری روش خوشه‌بندی پیشنهادی در تصویر ۴-۴ه قابل مشاهده است. در تصاویر ۴-۴ج تا ۴-۴ه که از نگاشت (۳-۲۳) برای تصویر کردن توصیف‌ها در فضای تصاویر استفاده شده است، مشکل جابجایی دامنه کاملا قابل رویت است، یعنی برای دسته‌های دیده شده توصیف‌ها به صورت مناسبی در مرکز نمونه‌های آن دسته نگاشته شده‌اند حال آن که برای دسته‌های دیده نشده جابجایی وجود دارد و توصیف‌های آن‌ها از نمونه‌هاشان فاصله گرفته‌اند؛ اما در تصویر ۴-۴و که از روش خوشه‌بندی و یادگیری نگاشت توام استفاده شده است این مشکل برطرف شده است و توصیف‌های دسته‌های دیده نشده نیز مانند دسته‌های دیده شده به مرکز نمونه‌های مربوط به خودشان نگاشته شده‌اند.

۴-۱۰ جمع‌بندی

در این فصل نتایج آزمایشات عملی برای روش‌های مختلف پیشنهادی در فصل قبل ارائه شد. ابتدا در بخش ۴-۱ مجموعه‌دادگان مورد استفاده معرفی شدند. در ادامه در بخش ۴-۴ شبکه عصبی چندوظیفه‌ای پیشنهادی مورد بررسی قرار داده شد و نتایج آن با سایر روش‌های پیش‌بینی صفت و هم‌چنین حالت ساده شده که از نمونه‌های آزمون استفاده نمی‌کند مقایسه شد. همچنین در بخش ۴-۴-۱ تابع مطابقت پیشنهادی به خروجی این شبکه اضافه شد که دقت پیش‌بینی‌های انجام شده را افزایش داد. در بخش ۴-۸ عمل‌کرد روش خوشه‌بندی نیمه‌نظارتی پیشنهادی مورد بررسی قرار گرفت. در بخش ۴-۷ روش دسته‌بندی با نگاشت به فضای تصاویر و استفاده از تابع مطابقت پیشنهادی و خوشه‌بندی نیمه‌نظارتی مورد آزمایش قرار گرفت و در بخش نتایج مربوط ۴-۸ روش یادگیری و خوشه‌بندی توام ارائه شد. در نهایت در بخش ۴-۹ نتایج مورد بررسی و مقایسه قرار گرفتند و علل عمل‌کرد برتر روش‌های پیشنهادی عنوان شد.



شکل ۴-۴: نمایش دوبعدی چهار دسته از مجموعه دادگان AwA با استفاده از نگاشت t-SNE ، دو دسته‌ی دیده شده شامل بزگوزن (فیروزه‌ای) خرس گریزلی (زرد) و دو دسته‌ی دیده نشده شامپانزه (آبی) و پاندا (سبز). تصاویر با نماد بعلاوه و نگاشت توصیف دسته‌ها در فضای تصاویر با ستاره نشان داده شده است. در تصاویر (ب) تا (و) نقطه‌های قرمز نمونه‌هایی که را نشان می‌دهد که دسته‌ای به جز چهار دسته‌ی موجود در شکل برای آن‌ها پیش‌بینی شده است. آ) دسته‌های دیده شده با برچسب صحیح و دیده‌نشده با رنگ مشکی ب) نمایش برچسب صحیح برای تمامی دسته‌ها ج) توصیف‌ها با نگاشت (۳-۲۳) به فضای تصاویر برده شده‌اند و دسته‌بندی با دسته‌بند نزدیک‌ترین همسایه انجام شده است. د) نگاشت مانند حالت قبل و دسته‌بندی با تابع مطابقت پیشنهادی به همراه خوشه‌بند k-means ه) نگاشت مانند حالت قبل و دسته‌بندی با تابع مطابقت پیشنهادی به همراه خوشه‌بند نیمه‌نظارتی پیشنهاد شده و) دسته‌بندی و نگاشت با استفاده از روش پیشنهادی برای یادگیری نگاشت و خوشه‌بندی توام.

فصل ۵

جمع‌بندی

۵-۱ جمع‌بندی

در این پژوهش مسئله یادگیری بدون برد را برای دسته‌بندی تصاویر مورد بررسی قرار دادیم. در این مسئله برای برخی دسته‌ها در زمان آموزش نمونه‌ی برچسب‌داری در اختیار نیست و این دسته‌ها با استفاده از یک نوع اطلاعات جانبی مشخص می‌شوند و برای آن‌ها دسته‌بند ساخته می‌شود. ابتدا یک چهارچوب کلی برای روش‌های موجود در مسئله یادگیری بدون برد ارائه کردیم. این چهارچوب شامل سه گام ۱) نگاشت تصاویر به یک فضای میانی، ۲) نگاشت توصیف‌ها به فضای میانی و ۳) دسته‌بندی در فضای میانی بود. سپس روش‌های پیشین در قالب این چهارچوب مرور شدند. در این مرور مشاهده کردیم که به استفاده از اطلاعات بدون نظارت موجود در ساختار فضای تصاویر کمتر توجه شده است.

در ادامه برای استفاده از اطلاعات موجود در ساختار فضای تصاویر، یک تابع مطابقت مبتنی بر خوشه‌بندی تصاویر بیان کردیم که قابلیت اضافه شدن به روش‌های پیشین و بهبود آن‌ها را داراست. با توجه به تکیه‌ی این تابع مطابقت به یک خوشه‌بندی از تصاویر یک روش خوشه‌بندی نیمه‌نظارتی ارائه دادیم که با ساختار و فرض‌های مسئله یادگیری بدون برد منطبق باشد. با ترکیب تابع مطابقت و خوشه‌بندی نیمه‌نظارتی معرفی شده، یک روش برای مسئله یادگیری بدون برد پیشنهاد کردیم که به نتایجی بهتر از نتایج پیشگام روش‌های پیشین در اکثر آزمایشات دست پیدا کرد. برای رفع نقایص این روش و افزایش بیشتر دقت دسته‌بندی، روش پیشنهادی دوم را تحت عنوان یادگیری نگاشت و خوشه‌بندی توام ارائه کردیم که محدودیت‌های ناشی از جدا بودن این مراحل در روش قبلی را برطرف کرده و دقت دسته‌بندی را افزایش داد.

۵-۲ کارهای آینده

با توجه به این مسئله که روش‌هایی که برای توصیف دسته‌های دیده نشده از هیستوگرام شباهت به دسته‌های دیده شده استفاده می‌کنند، به رغم این‌که از اطلاعات نمونه‌های آزمون استفاده نمی‌کنند، نتایج نزدیکی به روش نیمه‌نظارتی پیشنهاد شده توسط ما نزدیک است، بنظر می‌رسد یک شاخه امیدوارکننده برای ادامه پژوهش ترکیب این دو رویکرد باشد. یعنی نگاشت تصاویر و توصیف‌ها به فضای هیستوگرامی از دسته‌های دیده شده به صورتی که یادگیری این نگاشت‌ها و/یا دسته‌بندی در آن فضای مشترک با توجه و استفاده از نمونه‌های آزمون باشد.

یک شاخه دیگر که برای ادامه می‌تواند در نظر گرفته باشد ترکیب رویکرد شبکه‌های عصبی با روش‌های دیگر ارائه شده است، در این حالت با ویژگی‌های تصویر بکارگرفته شده در روش‌های ارائه شده در بخش‌های ۳-۶ و ۳-۷، به جای این که ثابت فرض شوند می‌توانند در جریان آموزش همراه با سایر پارامترها تعیین شوند.

استفاده از اطلاعات جانبی دیگر مانند نمایش برداری نام دسته‌ها به عنوان یک شاخه دیگر مطرح است که با توجه به ضعیف‌تر بودن اطلاعات نظارتی موجود در این نوع امضای دسته‌ها نسبت به بردار توصیف استفاده شده در این پژوهش، اطلاعات بدون نظارت موجود در نمونه‌های بدون برچسب می‌تواند موثرتر باشند و بهبود بیشتری ایجاد کند.

پیش‌بینی صفت‌های موجود درون تصویر با استفاده از شبکه‌های عصبی بازگشتی یک ایده‌ی قابل پیگیری دیگر است. با توجه به این که این شبکه‌ها امکان مدل‌سازی روابط صفات را دارا هستند، پیش‌بینی ویژگی با استفاده از این شبکه‌ها می‌تواند نتایج بهتری نسبت به مدل‌هایی که صفات را مستقل فرض می‌کنند داشته باشد.

کتاب‌نامه

[1] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions. *IEEE Conference on Computer Vision (ICCV)*, 2015.

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

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

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

[5] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI National Conference on Artificial Intelligence*, pages 646–651, 2008.

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

[7] 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*, pages 1410–1418. 2009.

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

[9] 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*, pages 935–943. 2013.

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

[11] 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.

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

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

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[15] B. Bakker and T. Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

[16] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

[17] E. Bart and S. Ullman. Cross-generalization: learning novel classes from a single example by feature replacement. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 672–679, 2005.

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

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

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

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

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

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

[24] 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*, pages 2121–2129, 2013.

[25] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2015.

[26] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. In *European Conference on Machine Learning (ECML)*, 2010.

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

[28] 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*, pages 3111–3119. 2013.

[29] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[30] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent Embeddings for Zero-shot Classification. pages 69–77, 2016.

[31] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele. Multi-cue zero-shot learning with strong supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS) 25*, pages 1097–1105. 2012.

[33] R. Qiao, L. Liu, C. Shen, and A. van den Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. pages 2249–2257, 2016.

[34] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning Deep Representations of Fine-grained Visual Descriptions. pages 49–58, 2016.

[35] 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.

[36] Z. Zhang and V. Saligrama. Zero-Shot Learning via Semantic Similarity Embedding. In *IEEE Conference on International Computer Vision (ICCV)*, pages 4166–4174, 2015.

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

[38] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation. In *European Confrence on Computer Vision (ECCV)*, volume 8690, pages 584–599, 2014.

[39] B. Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[40] X. Li and Y. Guo. Max-margin zero-shot learning for multi-class classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 626–634, 2015.

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

[42] Y. Fu and L. Sigal. Semi-supervised Vocabulary-informed Learning. 2016.

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

[44] D. Schuurmans and A. B. Tg. Semi-Supervised Zero-Shot Classification with Label Representation Learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4211–4219, 2015.

[45] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1717–1724, June 2014.

[46] H. El Khiyari, H. Wechsler, et al. Face recognition across time lapse using convolutional neural networks. *Journal of Information Security*, 7(03):141, 2016.

[47] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2014.

[48] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[49] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is np-hard. In *International Workshop on Algorithms and Computation*, pages 274–285. Springer, 2009.

[50] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[51] D. Hoiem, S. K. Divvala, and J. H. Hays. Pascal voc 2008 challenge, 2008.

[52] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

[53] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

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

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[57] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[58] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[59] M. D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[60] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[61] F. Chollet. Keras. https://github.com/fchollet/keras, 2015 (last visited June 2016).

[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference onComputer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

واژه‌نامه انگلیسی به فارسی

A

تابع فعال‌سازی ..... Activation Function

تناوبی ..... Alternative

صفت ..... Attribute

پیش‌بینی صفت ..... Attribute Prediction

B

پس‌انتشار ..... Back Propagation

کیسه‌ی کلمات ..... Bag of Words

اندازه دسته ..... Batch Size

شبکه بیزی ..... Baysian Network

دوخطی ..... Bi-Linear

C

شروع سرد ..... Cold Start

محدب ..... Convex

پیچش ..... Convolution

پیچشی ..... Convolutional

آنتروپی متقاطع ..... Cross Entropy

D

پیش‌بینی صفت مستقیم ..... Direct Attribute Prediction

F

انتخاب ویژگی ..... Feature Selection

صافی ..... filter

لایه با اتصالات کامل ..... fully connected layer

H

پارامتر ..... Parameter

I

پیش‌بینی صفت غیرمستقیم .... Indirect Attribute Prediction

L

راستی‌نمایی ..... Likelihood

ویژگی‌های محلی ..... local features

رگرسیون لجستیک ..... Logistic Regression

M

بیشینه حاشیه ..... Max Margin

دقت دسته‌بندی چنددسته‌ای. Mulit-Class Accuracy

O

یادگیری تک‌ضرب ..... One-shot Learning

بیش‌برازش ..... Over Fitting

P

افراز ..... Partitioning

تکه‌تکه خطی ..... Piece-wise Linear

ادغام ..... Pooling

R

تابع رتبه‌بند ..... Ranking Function

سامانه توصیه‌گر ..... Recommender System

بازگشتی ..... Recurrent

S

یادگیری نیمه‌نظارتی ... Semi-supervised Learning

امضا ..... Signature

سادک ..... Simplex

ایستا ..... stationary

یادگیری ساختار ..... Structure Learning

T

مدل‌سازی موضوع ..... Topic Modeling

انتقال یادگیری ..... Transfer Learning

واژه‌نامه فارسی به انگلیسی

ا

آنتروپی متقاطع ..... Cross Entropy

ادغام ..... Pooling

افراز ..... Partitioning

امضا ..... Signature

انتخاب ویژگی ..... Feature Selection

انتقال یادگیری ..... Transfer Learning

اندازه دسته ..... Batch Size

ایستا ..... stationary

ب

بازگشتی ..... Recurrent

بیش‌برازش ..... Over Fitting

بیشینه حاشیه ..... Max Margin

پ

پس‌انتشار ..... Back Propagation

پیچش ..... Convolution

پیچشی ..... Convolutional

پیش‌بینی صفت ..... Attribute Prediction

پیش‌بینی صفت غیرمستقیم .... Indirect Attribute Prediction

پیش‌بینی صفت مستقیم ..... Direct Attribute Prediction

ت

تابع رتبه‌بند ..... Ranking Function

تابع فعال‌سازی ..... Activation Function

تکه‌تکه خطی ..... Piece-wise Linear

تناوبی ..... Alternative

د

دقت دسته‌بندی چنددسته‌ای. Mulit-Class Accuracy

دوخطی ..... Bi-Linear

ر

راستی‌نمایی ..... Likelihood

رگرسیون لجستیک ..... Logistic Regression

س

سادک ..... Simplex

سامانه توصیه‌گر ..... Recommender System

ش

شبکه بیزی ..... Baysian Network

شروع سرد ..... Cold Start

ص

صافی ..... filter

صفت ..... Attribute

ف

پارامتر ..... Parameter

ک

کیسه‌ی کلمات ..... Bag of Words

ل

لایه با اتصالات کامل ..... fully connected layer

م

محدب ..... Convex

مدل‌سازی موضوع ..... Topic Modeling

و

ویژگی‌های محلی ..... local features

ی

یادگیری تک‌ضرب ..... One-shot Learning

یادگیری ساختار ..... Structure Learning

یادگیری نیمه‌نظارتی ... Semi-supervised Learning

یشترین حاشیه ..... Max Margin

Abstract In some of object recognition problems, labeled data may not be available for all categories. Zero-shot learning utilizes auxiliary information (also called signatures) describing each category in order to find a classifier that can recognize samples from categories with no labeled instance. On the other hand, with recent advances made by deep neural networks in computer vision, a rich representation can be obtained from images that discriminates different categorizes and therefore obtaining a unsupervised information from images is made possible. However, in the previous works, little attention has been paid to using such unsupervised information for the task of zero-shot learning. In this work, we first propose a multi-task neural network to predict attributes from images while exploiting this unsupervised information in order to mitigate the so called *domain shift problem* in predictions on unseen data. We also propose a novel semi-supervised zero-shot learning method that works on an embedding space corresponding to abstract deep visual features. We seek a linear transformation on signatures to map them onto the visual features, such that the mapped signatures of the seen classes are close to labeled samples of the corresponding classes and unlabeled data are also close to the mapped signatures of one of the unseen classes. We use the idea that the rich deep visual features provide a representation space in which samples of each class are usually condensed in a cluster. The effectiveness of the proposed method is demonstrated through extensive experiments on four public benchmarks improving the state-of-the-art prediction accuracy on three of them.

Keywords: Zero-shot Learning, Semi-supervised Learining, Deep Learning, Representation Learning.

Sharif University of Technology

Department of Computer Engineering

M.Sc. Thesis

Artificial Intelligence

Deep Zero-shot Learning

By:

Seyed Mohsen Shojaee

Supervisor:

Dr. Mahdieh Soleymani

Summer 2016