دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد

گرایش هوش مصنوعی

عنوان:

یادگیری بدون برد با شبکه‌های عمیق

نگارش:

سیدمحسن شجاعی

استاد راهنما:

دکتر مهدیه سلیمانی

تابستان ۱۳۹۵


بسم الله الرحمن الرحیم

سپاس

سپاس‌گزارم.

چکیده

کلیدواژه‌ها: زمان‌بندی کارکنان، زمان‌بندی مدرسه، جستجوی خلاق، برنامه درسی.

فهرست مطالب

فهرست شکل‌ها

فصل ۱

مقدمه

فصل ۲

روش‌های پیشین

در این فصل ابتدا یک چارچوب کلی برای روش‌های مورد استفاده در یادگیری بدون برد توصیف می‌شود. سپس روش‌های موجود طبق این چارچوب دسته‌بندی شده و مرور خواهند شد. پیش از تعریف و بیان رسمی مسئله یادیگری بدون برد، استفاده از اشتراک و تمایز برخی ویژگی‌ها میان دسته‌های مختلف در بینایی ماشین مورد بررسی قرار گرفته است [۲، ۳، ۴] اما این روش‌ها به شناسایی دسته‌های کاملا جدید از روی این ویژگی‌ها توجه نشان نداده‌اند. مسئله‌ی یادگیری تک‌ضرب[۱] هم یک مسئله نزدیک به یادگیری بدون برد است که پیش‌تر مورد بررسی بوده است [۵]. در حقیقت می‌توان یادگیری تک‌ضرب را حالت خاصی از یادگیری بدون برد در نظر گرفت که در آن توصیف دسته‌های دیده نشده به صورت یک نمونه از آن دسته ارائه شده است [۶].

پدیده شروع سرد[۲] در سامانه‌های توصیه‌گر[۳] را نیز می‌توان از حالت‌های خاص یادگیری بدون برد در نظر گرفت که در آن برای یک کاربر یا مورد جدید پیشنهاد صورت می‌گیرد.

بیان مسئله یادگیری بدون برد به طور رسمی برای اولین بار در [۶] صورت گرفت. در آن‌جا دو رویکرد کلی برای حل مسئله یادگیری بدون برد بیان می‌شود. یک روش که رویکرد فضای ورودی[۴] نامیده می‌شود، سعی در مدل کردن نگاشتی با دو ورودی دارد. یکی نمونه‌ها و دیگری توصیف دسته‌ها. این نگاشت برای نمونه‌ها و توصیف‌های مربوط به یک

[۱] One-shot Learning

[۲] cold start

[۳] Recommender Systems

[۴] input space view

دسته امتیاز بالا و برای نمونه‌ها و توصیفاتی که متعلق به دسته‌ی 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 رویکرد فضای مدل[۵] 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 معرفی یک چارچوب کلی برای انجام یادگیری بدون برد لازم است. دو رویکرد فوق نیز در این چارچوب قابل بیان هستند، این موضوع در بخش ؟؟ که مثال‌هایی از این رویکردها مرور می‌شود، روشن‌تر خواهد شد.

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

۱. یادگرفتن نگاشتی از فضای تصاویر به فضای مشترک که آن را با ϕ نشان می‌دهیم.

۲. نگاشت توصیف‌ها به فضای مشترک که آن را با θ نشان می‌دهیم.

۳. اختصاص برچسب به تصاویر

۲-۱ نمادگذاری

برای این که توصیف دقیق روش‌های پیشین 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.

تصاویر را با $x \in \mathbb{R}^d$ نشان می‌دهیم که d ابعاد داده را نشان می‌دهد. توصیف‌ها را با $c \in \mathbb{R}^a$ نمایش می‌دهیم 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 c^y که در آن $y \in \mathcal{U} \cup \mathcal{S}$ بردار توصیف دسته y را نشان می‌دهد.

[۵] model space view

فرض می‌کنیم در زمان آموزش $\{(x^i, y^i)\}_{i=1}^{N_s}$ شامل N_s تصویر از دسته‌های دیده شده به همراه برچسب موجود است. $X_s \in \mathbb{R}^{N_s \times d}$ مجموعه تصاویر و Y_s برچسب‌های داده‌های آموزش با نمایش یکی یک [۶] است. هم‌چنین توصیف‌های هر کدام از دسته‌های آموزش، $C_s \in \mathbb{R}^{s \times a}$ نیز موجود است. X_u و C_u بطور مشابه برای دسته‌های آزمون تعریف می‌شوند. $(X)_i$ سطر iم از ماتریس X و x_i درایه‌ی iم از بردار x را نشان می‌دهد. ضرب داخلی با نماد $\langle ., . \rangle$ نشان داده شده است.

در ادامه به بررسی روش‌های ارائه شده برای مسئله یادگیری بدون برد با استفاده از چارچوب ارائه شده خواهیم پرداخت.

۲-۲ پیش‌بینی ویژگی

این دسته از روش‌ها عموما به حالتی از مسئله یادگیری بدون برد تعلق دارند که توصیف دسته‌ها از نوع بردار ویژگی باشد. در این حالت فضای مشترک همان فضای ویژگی‌ها در نظر گرفته می‌شود. به عبارت دیگر نگاشت θ نگاشت همانی فرض شده و یادگرفته نخواهد شد. روش‌های اولیه ارائه شده برای یادگیری بدون برد از نوع پیش‌بینی ویژگی[۷] 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.

۲-۲-۱ پیش‌بینی ویژگی مستقیم و غیر مستقیم

در [۷] با فرض این که ویژگی‌ها به صورت مستقل از یکدیگر قابل پیش‌بینی هستند دو رویکرد برای این کار ارائه می‌کند. پیش‌بینی ویژگی مستقیم[۸] و پیش‌بینی ویژگی غیر مستقیم[۹]. مدل گرافی مورد استفاده در این دو رویکرد در تصویر ؟؟ آمده است. در پیش‌بینی ویژگی مستقیم برچسب‌ها به شرط دانستن ویژگی‌های درون تصویر، از تصویر مستقل هستند. در این روش برای هر یک ویژگی‌ها یک دسته‌بند یاد گرفته می‌شود. با توجه به این که ویژگی‌ها برای تصاویر آزمون معین هستند این کار با استفاده از یک دسته‌بند احتمالی برای هر ویژگی قابل انجام است. در نهایت احتمال تعلق هر یک از برچسب‌های $u \in \mathcal{U}$ با استفاده از رابطه زیر بدست خواهد آمد.

$$P(z_u|x) = \sum_{c \in \{0,1\}^a} P(u|c)p(c|x) \qquad (۱-۲)$$

[۶] One-Hot Encoding

[۷] Attribute Prediction

[۸] Direct Attribute Prediction

[۹] Indirect Attribute Prediction

از با توجه به فرض استقلال ویژگی داریم $P(c|x) = \prod_{n=1}^{a} P(c_m|x)$. برای محاسبه جمله $P(z_u|a)$ از قانون بیز استفاده می‌کنیم:

$$P(u|c) = \frac{P(u)P(c|u)}{P(a^u)} = \frac{P(u)\mathbb{1}(c = c^u)}{P(c^u)}$$

با جایگذاری در رابطه (۲-۱) خواهیم داشت:

$$P(u|x) = \frac{P(u)}{P(c^u)} \prod_{n=1}^{a} P(a_n^u|x) \tag{۲-۲}$$

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

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

$$P(c_i|x) = \sum_{k=1}^{n_u} P(y_k|x)\mathbb{I}(c_i = c_i^{y_k}) \tag{۲-۳}$$

در نهایت در هر دو روش برچسب نهایی با تخمین MAP[۱۰] از رابطه زیر تعیین می‌شود:

$$\hat{y} = \arg\max_{u \in \mathcal{U}} P(u|x) = \arg\max_{u \in \mathcal{U}} \prod_{i=1}^{a} \frac{P(c_i^u|x)}{P(c_i^u)} \tag{۲-۴}$$

روش ارائه شده در [۸] مشابه همین روش است با این تفاوت که احتمال مشاهده هر کدام ویژگی‌ها 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 ویژگی‌ها از یکدیگر است؛ چرا که این فرض در مسائل واقعی معمولا بر قرار نیست. برای مثال زمانی که ویژگی آبزی بودن برای یک موجود در نظر گرفته می‌شود احتمال ویژگی پرواز کردن برای آن بسیار کاهش می‌یابد.

۲-۲-۲ مدل‌سازی احتمالی روابط بین ویژگی‌ها

مدل‌های گرافی برای در نظر گرفتن وابستگی‌های میان ویژگی‌ها به کار گرفته شده‌اند. نویسندگان [۹] برای در نظر گرفتن ارتباط بین ویژگی‌ها و ارتباط ویژگی‌ها با برچسب نهایی روش‌های مدل‌سازی موضوع [۱۱] را از حوزه یادگیری در متن اقتباس می‌کنند. همچنین نویسندگان [۱۰] برای این کار یک چارچوب بر اساس مدل‌های گرافی احتمال معرفی می‌کنند.

[۱۰] Maximum a Posteriori

[۱۱] Topic Modeling

در این چارچوب یک شبکه بیزی[۱۲] برای مدل کردن این روابط در نظر گرفته می‌شود و ساختار آن که نشان‌دهنده وابستگی یا استقلال ویژگی‌ها با هم یا با برچسب است، با کمک روش‌های یادگیری ساختار[۱۳] شناخته می‌شود.

۲-۳ نگاشت به فضای توصیف‌ها

در برخی موارد توصیف‌های داده شده از جنسی غیر از ویژگی هستند ولی فضای مشترک همان فضای توصیف‌ها در نظر گرفته می‌شود و سعی می‌شود تصاویر به این فضا نگاشته شوند. روش ConSE[۱۴] [؟] از چنین نگاشتی استفاده می‌کند. ابتدا یک شبکه عصبی کانولوشنال برای دسته‌بندی نمونه‌های دسته‌های دیده‌شده آموزش داده می‌شود. این یادگیری یک مسئله دسته‌بندی عادی است و شبکه‌ها در اکثر موارد از قبل به صورت پیش‌آموزش دیده شده وجود دارند. تابع فعال‌سازی[۱۵] لایه‌ی آخر این شبکه به این صورت تعریف می‌شود:

$$softmax(z)_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \quad j = ۱, \dots, n_s. \tag{۵-۲}$$

تابع بالا به ازای هر j، امتیاز تعلق نمونه به دسته‌ی jم 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 بدون برد، هنگامی که یک نمونه از دسته‌های آزمون را به شبکه می‌دهیم، خروجی بدست آمده از رابطه (؟؟) می‌تواند به عنوان میزان شباهت آن نمونه به هر یک دسته‌های آموزش در نظر گرفته شود. فرض کنید که برای هر نمونه $\hat{y}(x,n)$، nمین عنصر بزرگ $softmax(x)$ را نشان دهد، یعنی nمین برچسب محتمل برای x از میان دسته‌های آموزش. حالا برای پیش‌بینی برچسب x از میان دسته‌های آموزش از این رابطه استفاده می‌کنیم:

$$\phi(x) = \frac{۱}{Z} \sum_{n=۱}^{T} P(\hat{y}(x,n)|x) \cdot c_{\hat{y}(x,n)}, \tag{۶-۲}$$

که T یک فراپارامتر مدل $Z = \sum_{n=۱}^{T} P(\hat{y}(x,n)|x)$ 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[۱۶] [۴۰] نیز از رویکرد مشابهی استفاده می‌کند.

[۱۲] Baysian Network
[۱۳] Structure Learning
[۱۴] Convec combination of Semantic Embeddings
[۱۵] Activation Function
[۱۶] Co-Occurance Statistics

در این روش همانند رابطه (۲-۶)، 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.

۲-۴ نگاشت‌های دو خطی

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^TW تصویر x را به فضای توصیف‌ها نگاشته و Wc توصیف c را به فضای تصاویر می‌نگارد.در نهایت تابع مطابقت میان یک توصیف و تصویر به صورت زیر تعریف می‌شود:

$$F(x,c) = \phi(x)^T W \theta(y) \tag{۲-۷}$$

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 غیر خطی قابل حل خواهند بود.

۲-۴-۱ یادگیری با توابع رتبه‌بند

یک انتخاب متداول برای تابع هزینه، توابع رتبه‌بند[۱۷] 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 یادیگری نگاشت بهتر از مجموع مربعات است که تنها فاصله نقاط از برچسب خودشان را در نظر می‌گیرد [۱۱].

در [۱۲] تابع هزینه رتبه‌بند WSABIE [۱۳] که برای حاشیه‌نویسی تصاویر پیشنهاد شده، به مسئله یادگیری بدون برد انطباق می‌دهد. تابع هزینه WSABIE به این صورت تعریف شده است:

[۱۷] ranking function

$$L(x_s, Y_s; W, \theta) = \frac{1}{N_s} \sum_{n=1}^{N_s} \lambda_{r_\Delta(x_n, y_n)} \sum_{y \in \mathcal{Y}} \max(0, l(x_n, y_n, y)) \quad (۸-۲)$$

$$l(x_n, y_n, y) = \mathbb{1}(y \neq y_n) + \phi(x_n)^T W \theta(y) - \phi(x_n)^T W \theta(y_n) \quad (۹-۲)$$

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 ویژگی‌ها را این گونه جریمه می‌کند که به ازای برچسب نادرستی که رتبه بالاتری از برچسب صحیح در دسته‌بندی دریافت کرده، جریمه‌ای متناسب با امتیاز برچسب ناصحیح در نظر گرفته می‌شود.ضریب نزولی λ_k میزان جریمه را برای برچسب‌های غلط در رتبه‌های بالا، بیشتر در نظر می‌گیرد. در انطباق برای یادگیری بدون برد، بهینه‌سازی تنها روی نگاشت W انجام شده و تابع θ دانسته فرض می‌شود: $\theta(y) = c_y$.

ایده‌ی بالا در [۱۴] ادامه داده شده و نگاشت شباهت ساخت‌یافته SJE[۱۸] نامیده شده است. ، در این حالت تابع مطابقت بین توصیف‌ها و تصاویر از رابطه (۲-۷) تعریف می‌شود. تابع هزینه ساده‌تر از حالت قبل به صورت

$$\frac{1}{N_s} \sum_{n=1}^{N_s} \max_{y \in \mathcal{Y}} (0, l(x_n, y_n, y)) \quad (۱۰-۲)$$

در نظر گرفته شده که l همانند رابطه (۲-۹) است. هم‌چنین برای استفاده از چند توصیف به صورت هم‌زمان، تعریف تابع مطابقت به صورت زیر تعمیم داده می‌شود:

$$F(x, y; \{W\}_{1 \ldots K}) = \sum_k \alpha_k \theta(x)^T W_k \phi_k(y) \quad (۱۱-۲)$$

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

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 بر روی آن‌ها تست انجام شده است شامل بردار ویژگی‌های دودویی یا پیوسته تعیین شده توسط انسان و نمایش برداری متون دائره‌المعارفی با روش‌های word2vec [۱۵] و GloVe [۱۶] 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.

[۱۸] Structured Joint Embedding

روش SJE در [۱۷] برای برخی نگاشت‌های غیرخطی نیز تعمیم داده شده است. در این روش که LatEm[۱۹] نام دارد تابع هزینه مانند حالت قبل (رابطه (۲-۱۰)) تعریف شده است با این تفاوت که تابع مطابقت میان توصیف و تصویر بجای رابطه دوخطی (۲-۷) از این رابطه تبعیت می‌کند:

$$F(x,y) = \max_{1 \leqslant i \leqslant L} \phi(x)^T W \theta(y) \qquad (۱۲-۲)$$

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

در [۱۱] نیز که برای اولین بار توصیف تنها نام برچسب دسته‌ها در نظر گرفته شده، از نگاشت دو خطی استفاده شده است. در این روش نام برچسب‌ها با استفاده از مدل نهان‌سازی کلمات word2vec کلمات به بردارهایی نگاشته می‌شوند. ابعاد فضای نهان‌سازی کلمات یک فراپارامتر است که در این مقاله با اعتبار سنجی تعیین شده است. استخراج ویژگی از تصاویر با استفاده از شبکه عصبی کانولوشنال [۱۸] که روی دسته‌های دیده شده آموزش داده شده، انجام می‌شود. در نهایت یک تابع بیشترین حاشیه[۲۰] برای یادگیری نگاشت دو خطی پیشنهاد می‌شود.

$$L((x_n, y_n); W) = \sum_{y \neq y_n} \max(0, \xi - x_n W c_{y_n} + x_n W c_y) \qquad (۱۳-۲)$$

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

۲-۴-۲ روش‌های مبتنی بر خطای مجموع مربعات

یک نحوه‌ی استفاده دیگر از نگاشت‌های دو خطی، دسته‌بندی مستقیم با این نگاشت است.

$$\underset{W \in \mathbb{R}^{d \times a}}{\text{minimize}} \left\| X_s^T W C_s - Y \right\|_{Fro} + \Omega(W) \qquad (۱۴-۲)$$

که در آن Ω یک جمله منظم‌سازی است. در این حالت اگر تبدیل را از فضای تصاویر به فضای ویژگی‌ها نگاه کنیم، نگاشت W باید تصاویر را به زیرفضایی عمود به تمامی بردار ویژگی‌های 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

[۱۹] Latent Embedding Model

[۲۰] Max margin

برای دسته‌بندی به شمار نمی‌آید، به جمله منظم سازی آن نسبت داده می‌شود. جمله منظم‌سازی Ω به این صورت تعریف می‌شود:

$$\Omega(W) = \lambda \|WC_s\|_{Fro} + \gamma \left\|X_s^T W\right\|_{Fro} + \lambda\gamma \|W\|_{Fro} \qquad (۱۵-۲)$$

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 ویژگی است در نتیجه طبیعی است که پارامترهای این دو دسته‌بند با نرم فروبنیوس آن‌ها کنترل شود تا از بیش‌برازش[۲۱] 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 [۱۹] 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 ویژگی وجود دارد، ماتریس تبدیل W به دو ماتریس تجزیه می‌شود:

$$W = V_x^T V_c \qquad (۱۶-۲)$$

با این تجزیه از افزایش شدید تعداد پارامترها در اثر افزایش بعد بردار توصیف‌ها جلوگیری می‌شود. (دقت کنید که بعد C برابر $d \times a$ است) علاوه بر این V_c می‌تواند برای استخراج ویژگی‌های مفید و حذف نویز از C 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:

$$\min_{V_x, V_c} \left\|X_s^T + V_x^T V_c C\right\|_{Fro} + \lambda_۱ \left\|V_x^T V_c C\right\|_{Fro} + \lambda_۲ \left\|V_c^T\right\|_{۲,۱} \qquad (۱۷-۲)$$

که $\left\|M^T\right\|_{۲,۱} = \sum_i \left\|M_{(i)}\right\|_۲$ و این نوع منظم‌سازی، ستون‌های ماتریس V_c را به سمت تنک بودن سوق خواهد داد. در واقع اگر $\lambda_۲$ بزرگ انتخاب شود، V_c نقش یک ماتریس انتخاب ویژگی[۲۲] را خواهد داشت. جمله‌های منظم سازی دیگر در (۲-۱۵) به دلیل تاثیر اندکشان در آزمایشات عملی حذف شده‌اند.

[۲۱] overfitting

[۲۲] feature selection

۲-۵ 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 [۲۰] ارائه شده است. در این روش که برای تصاویر مجموعه متون بزرگتری نسبت به دادگان قبلی جمع‌آوری و استفاده شده است. این ازدیاد در داده‌ها امکان آموزش مدل‌های پیچیده‌تر و پیشرفته‌تر را برای یادگیری نگاشت از فضای تصاویر فراهم می‌کند و فاصله میان عمل‌کرد یادگیری بدون برد هنگام استفاده از توصیف‌های متنی و توصیف‌های به صورت بردار ویژگی را کمتر کرده است. در این حالت فرض می‌شود که داده‌های آموزش به صورت $\{(v_n, t_n, y_n), n = ۱, ..., N\}$ 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 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 مجوعه 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 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

$$F(v, t) = \theta(v)^T \phi(t) \tag{۱۸-۲}$$

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

$$f_v(v) = \arg\max_{y \in \mathcal{Y}} \left(\mathbb{E}_{t \sim \mathcal{T}(y)}[F(v, t)]\right) \tag{۱۹-۲}$$

$$f_t(t) = \arg\max_{y \in \mathcal{Y}} \left(\mathbb{E}_{v \sim \mathcal{V}(y)}[F(v, t)]\right). \tag{۲۰-۲}$$

یادگیری تابع F با تابع هزینه‌ی زیر صورت می‌گیرد:

$$\frac{1}{N}\sum_{n=1}^{N}\ell_v(v_n,t_n,y_n)+\ell_t(v_n,t_n,y_n), \tag{۲۱-۲}$$

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 این تابع هزینه با رابطه (۲-۱۰) در اضافه شدن جمله‌ی دوم است. در رابطه (۲-۱۰) 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 باید به ویژگی بصری خود نزدیک‌تر باشد تا سایر ویژگی‌های بصری. نگاشت θ مانند سایر روش‌ها یک شبکه عصبی عمیق کانولوشنال است که از قبل با داده‌های ImageNet آموزش داده شده‌است. برای هر تصویر قسمت‌های بصری مختلف با بریدن قسمت‌های متفاوت از تصویر حاصل می‌شود. نگاشت ϕ برای متون با سه شبکه عصبی مختلق کانولوشنال، بازگردنده و کانولوشنال بازگردنده (CNN-RNN) مدل شده است. استفاده از این شبکه‌ها برای نگاشت متن در این روش نخستین بار در این روش رخ داده‌است. جمع‌آوری مجموعه دادگان متنی بزرگتر، آموزش چنین شبکه‌هایی را ممکن کرده است.

در [۲۱] 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:

$$y^*=\arg\max_y \phi(c^y)^T x \tag{۲۲-۲}$$

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

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

۲. تابع مطابقت: نگاشت دو خطی 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 (۲-۲۲)، $c^T W$ 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:

$$\phi(c) = \arg\min_{w,\zeta_i} w^T w - \alpha c^T D w - \beta\, ln(P_{reg}(w|c)) + \gamma \sum \zeta_i \qquad (۲۳\text{-}۲)$$

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

$$c^T D c \geqslant l$$

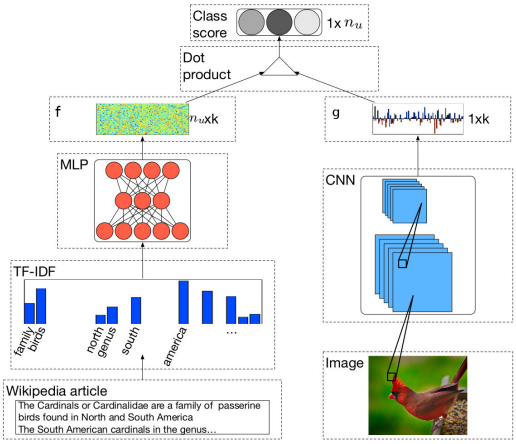
که α, β, γ, l فراپارامترهای مدل هستند. جمله اول در این تابع هزینه، منظم‌سازی دسته‌بند خطی w است. جمله دوم مشابهت w با $c^T D$ را الزام می‌کند و جمله سوم احتمال بالا در رگرسیون را در نظر می‌گیرد. محدودیت $-(w^T x_i) \geqslant \zeta_i$ بر اساس فرض عدم تعلق نمونه‌های آزمون به کلاس‌های دیده‌شده تعریف شده است 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 [۲۳] برای دسته‌بندهای غیرخطی نیز توسعه داده‌اند [۲۲].

۲-۶ 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.

در [۱] از شبکه‌های عصبی عمیق برای یادگیری توام نگاشت‌های ϕ و θ استفاده شده است. نمای کلی شبکه مورد

[۲۳] kernel trick



شکل ۲-۱: شبکه مورد استفاده برای یادگیری توام نگاشت تصاویر و توصیف‌ها که یک شبکه عصبی عمیق با دو ورودی است. ورودی اول از نوع تصویر است و ابتدا با یک شبکه کانولوشنال سپس با چند لایه چگال به فضایی $k-$بعدی می‌رود. ورودی دوم که یک مقاله از ویکی‌پدیای انلگیسی است پس از تبدیل به نماش 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].

استفاده در این روش در تصویر ۲-۶ 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 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 غیر خطی از توصیف کلاس y است: $w^y = f(c^y)$ که f شبکه عصبی مخصوص متن است (نیمه‌ی چپ تصویر ۲-۶.) استخراج ویژگی غیر خطی از تصاویر نیز با یک شبکه عصبی که تابع آن را g می‌نامیم، انجام شده است(نیمه‌ی راست تصویر ۲-۶.) در نهایت دسته‌بندی با تابع زیر انجام می‌شود:

$$y^* = \arg\max_y \ w^{yT} g(x). \qquad (۲۴-۲)$$

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

فرض کنید b خروجی یک لایه‌ی کانولوشنال 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 کانولشنال فورمول‌بندی می‌شود. ابتدا یک کاهش ابعاد غیر خطی 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 کانولوشن $w^y = f'(c^y)$ 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 کانولوشنال به صورت زیر تعریف می‌شود:

$$\text{score}(x, y) = o\Big(\sum_{i=1}^{K'} w_i^{y'} \check{*} a'_i\Big), \tag{۲۵-۲}$$

$\text{score}(x, y)$ امتیاز تعلق x به دسته‌ی y است؛ $o(\cdot)$ یک تابع ادغام[۲۴] به صورت $o : \mathbb{R}^{l \times h} \mapsto \mathbb{R}$ و $\check{*}$ نشانگر عمل کانولوشن 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 کانولوشنال پیشنهاد می‌شود که در 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:

$$\text{score}(x, y) = w^{yT} g(x) + o\Big(\sum_{i=1}^{K'} w_i^{y'} \check{*} g'(a)_i\Big). \tag{۲۶-۲}$$

در این حالت پارامترهای مربوط به g, g', f, f' به صورت همزمان یادگرفته می‌شوند. یادگیری در شبکه بر اساس خطای تنها خروجی که نشان می‌دهد آیا این متن و توصیف هم‌دسته هستند یا نه صورت می‌گیرد. در این پژوهش دو تابع هزینه

[۲۴] pooling

برای خطا در نظر گرفته شده ۱) آنتروپی تقاطعی [۲۵] ۲) تابع هزینه لولا[۲۶]. بررسی عمل‌کرد این دو نوع تابع هزینه نشان می‌دهد که بر اساس معیار ارزیابی نهایی هر کدام می‌توان عمل‌کرد بهتری نسبت به دیگری داشته باشد. اگر معیار ارزیابی دقت دسته‌بندی در k انتخاب اول[۲۷] باشد تابع هزینه لولا بهتر عمل می‌کند و اگر معیار مساحت زیر نمودار صحت و بازیابی[۲۸] باشد، آنتروپی متقاطع عمل‌کرد بهتری دارد.

در [۳۰] روشی برای ساخت بردارهای ویژگی برای تصاویر، برای دسته‌بندی بهتر آن‌ها، در حالت عادی دسته‌بندی تصاویر، ارائه شده است. این روش برای هر دسته یک بردار ویژگی و برای هر یک از ویژگی‌ها یک دسته‌بند یاد می‌گیرد. این روش برای یادگیری بدون برد 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 ویژگی برای دسته‌ها جزو 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 ویژگی تولید می‌کند. فرض کنید در کل n دسته موجود باشد و و قصد داشته باشیم بردار ویژگی‌های l بعدی تولید کنیم (l یک فراپارامتر 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 ویژگی‌هاست. در نهایت یک نمونه با استفاده از رابطه زیر قابل دسته‌بندی خواهد بود:

$$y^* = \arg\min_i \|A_{(i)} - f(x)^T\| \qquad (۲۷-۲)$$

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

- ایجاد تمایز: بردار ویژگی 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: ویژگی‌ها باید با خطای کم از روی تصاویر قابل پیش‌بینی باشند. یک روش برای ایجاد چنین حالتی این است که ویژگی‌ها 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

[۲۵] Cross Entropy
[۲۶] hinge loss
[۲۷] top-k accuracy
[۲۸] Precision Recall Area Under the Curve

f خواهد داشت. برای یادگیری A طوری که دو خاصیت فوق را داشته باشد تابع هزینه

$$\max_{A} \sum_{i,j} \left\| A_{(i)} - A_{(j)} \right\|_۲^۲ - \lambda \sum_{i,j} S_{ij} \left\| A_{(i)} - A_{(j)} \right\|_۲^۲ \qquad (۲۸-۲)$$

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 ویژگی‌های بصری مشابه داشته باشند تا بتوان این ویژگی‌ها 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 بدون برد، 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 (۲-۲۷) انجام می‌شود.

۲-۶-۱ 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 اقدلیدسی در این فضا) در زمینه‌ی یادگیری بدون برد 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 [۴۱] 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:

$$\theta(\mathbf{c}) = \underset{\boldsymbol{\alpha} \in \Delta^{|\mathcal{S}|}}{\arg\min} \left\{ \frac{\gamma}{۲} \|\boldsymbol{\alpha}\|^۲ + \frac{۱}{۲} \|\mathbf{c} - \sum_{y \in \mathcal{S}} \mathbf{c}_y \alpha_y\|^۲ \right\}, \qquad (۲۹-۲)$$

که در آن $\Delta^{|\mathcal{S}|}$ سیمپلکس 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[۲۹] 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:

$$\text{INT:} \quad \phi_y(\mathbf{x}) = \min(\mathbf{x}, \mathbf{v}_y), \tag{۳۰-۲}$$

$$\text{ReLU:} \quad \phi_y(\mathbf{x}) = \max(۰, \mathbf{x} - \mathbf{v}_y), \tag{۳۱-۲}$$

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:

$$\phi(x) = \big(w^T\psi_۱(x), w^T\psi_۲(x), \ldots, w^T\psi_{n_s}(x)\big) \tag{۳۲-۲}$$

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

$$y^* = \arg\max_{y\in\mathcal{Y}} \langle \phi(x), \theta(c^y)\rangle. \tag{۳۳-۲}$$

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

$$\min_{\mathcal{V},\mathbf{w},\boldsymbol{\xi},\boldsymbol{\epsilon}} \frac{۱}{۲}\|\mathbf{w}\|^۲ + \frac{\lambda_۱}{۲}\sum_{\mathbf{v}\in\mathcal{V}}\|\mathbf{v}\|^۲ + \lambda_۲\sum_{y,s}\epsilon_{ys} + \lambda_۳\sum_{i,y}\xi_{iy} \tag{۳۴-۲}$$

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

$$\sum_{i=۱}^{N} \frac{\mathbb{I}_{\{y_i=y\}}}{N_y}\Big[f(\mathbf{x}_i, y) - f(\mathbf{x}_i, s)\Big] \geqslant \Delta(y, s) - \epsilon_{ys}, \tag{۳۵-۲}$$

$$f(\mathbf{x}_i, y_i) - f(\mathbf{x}_i, y) \geqslant \Delta(y_i, y) - \xi_{iy}, \tag{۳۶-۲}$$

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

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

[۲۹] Rectified Linear Unit

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 بیشترین حاشیه است با این تفاوت که علاوه بر محدودیت بیشترین حاشیه (رابطه(۲-۳۶)) یک محدودیت برای دسته‌بندی صحیح به صورت میانگین هم در رابطه (۲-۳۵) اضافه شده است. این محدودیت جدید می‌تواند باعث شود که داده‌ها به گونه‌ای نگاشته شود که نه تنها دسته‌بندی صحیح صورت گیرد بلکه یک توزیع با مرکز $\theta(c^y)$ ایجاد کنند. این حالت باعث اینجاد 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 [۴۴] با یادگیری توامان نگاشت توصیف‌ها و تصاویر توسعه داده‌اند. علاوه بر یادگیری توامان پارامترهای نگاشت‌ها، برای داده‌های تست، 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.

۲-۷ 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 [؟] برای نخستین بار مشکل جابجایی دامنه[۳۰] معرفی شد. این مشکل که در شکل ؟؟ قابل مشاهده است به متفاوت بودن خواص متفاوت ویژگی‌ها برای دسته‌های مختلف اشاره می‌کند. برای مثال ویژگی 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 ویژگی در تصویر ببر ضعیف خواهد بود. در [؟] برای حل این مشکل دو تکنیک به کار گرفته شده است. ابتدا یافتن نمایش مشترک برای سه دامنه‌ی تصاویر، بردار ویژگی و بردار نام دسته‌ها به

[۳۰] Domain shift problem

صورت توامان با استفاده از CCA[۳۱][؟] سپس برچسب‌گذاری داده‌های بدون برچسب در این فضای مشترک با استفاده از یک تکنیک انتشار برچسب[۳۲] بیزی.

در [۴۳] مسئله به صورت یک دسته‌بندی روی دسته‌های دیده شده و خوشه‌بندی روی دسته‌های دیده‌نشده به صورت توام مدل شده است. در این روش یک دسته‌بند خطی روی تصاویر یادگرفته می‌شود که این دسته‌بند ترکیبی از پارامترهای مدل و توصیف‌هاست. به صورت دقیق‌تر چهارچوب یادگیری برابر خواهد بود با:

$$\min_{Y,U,W,\boldsymbol{\xi}} \quad \frac{\beta}{۲}\left\|W\right\|_{Fro}^{۲}+\frac{\beta}{۲}\left\|U\right\|_{Fro}^{۲}+\mathbf{۱}^{T}\boldsymbol{\xi} \tag{۳۷-۲}$$

$$s.t. \quad diag\big((Y-\mathbf{۱}\mathbf{۱}_{k}^{T})\big)UWX^{T})\geqslant(\mathbf{۱}-Y\mathbf{۱}_{k})-\boldsymbol{\xi},\ \forall k\in\mathcal{Y} \tag{۳۸-۲}$$

$$Y\in\{۰,۱\}^{(N_s+N_u)\times(n_s+n_u)},\quad BY=Y_s, \tag{۳۹-۲}$$

$$Y\mathbf{۱}=\mathbf{۱},\quad l\mathbf{۱}\leqslant Y^{T}\mathbf{۱}\leqslant h\mathbf{۱} \tag{۴۰-۲}$$

که در این صورت‌بندی فوق، U را می‌توان توصیف‌های موجود برای هر دسته در نظر گرفت، $\mathbf{۱}$ یک بردار تمام یک و $\mathbf{۱}_k$ یک بردار که عنصر kم آن یک و سایر عناصر آن صفر است را نشان می‌دهند.

[۳۱] Canonical Correlation Analysis

[۳۲] Label Propagation

فصل ۳

روش پیشنهادی

فصل ۴

نتایج

فصل ۵

جمع‌بندی

۵-۱ جمع‌بندی

۵-۲ کارهای آینده

Bibliography

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

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

[3] 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.

[4] E. Bart and S. Ullman. Cross-generalization: learning novel classes from a single example by feature replacement. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 672–679, 2005.

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

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

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

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

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

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent Embeddings for Zero-shot Classification. mar 2016.

[18] 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.

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

[20] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning Deep Representations of Fine-grained Visual Descriptions. 2016.

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

[22] 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.

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

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

[25] 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.

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

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

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

[29] 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.

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

[31] G. Tsoumakas and Katakis. Multi Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

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

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

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

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

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

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

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

[39] 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.

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

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

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

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

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

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

[46] 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.

Abstract

Keywords: Timetabling, School Timetabling Problem, Personnel Scheduling

Sharif University of Technology

Department of Computer Engineering

M.Sc. Thesis

Artificial Intelligence

Deep Zero-shot Learning

By:

Seyed Mohsen Shojaee

Supervisor:

Dr. Mahdaieh Soleymani

Summer 2017