



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



سپاس

سپاس گزارم.

چکیده

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

فهرست مطالب

۱	۱ مقدمه
۲	۲ روش‌های پیشین
۳	۱-۲ نمادگذاری
۴	۲-۲ پیش‌بینی ویژگی
۴	۱-۲-۲ پیش‌بینی ویژگی مستقیم و غیر مستقیم
۵	۲-۲-۲ مدل‌سازی احتمالی روابط بین ویژگی‌ها
۶	۳-۲ نگاشت به فضای توصیف‌ها
۷	۴-۲ نگاشت‌های دو خطی
۷	۱-۴-۲ یادگیری با توابع رتبه‌بند
۹	۲-۴-۲ روش‌های مبتنی بر خطای مجموع مربعات
۱۱	۵-۲ نگاشت به فضای تصاویر
۱۳	۶-۲ نگاشت به یک فضای میانی
۱۶	۱-۶-۲ نگاشت به فضای دسته‌های دیده شده
۱۷	۳ روش پیشنهادی

۴	نتایج	۱۸
۵	جمع‌بندی	۱۹
۱-۵	جمع‌بندی	۱۹
۲-۵	کارهای آینده	۱۹

فهرست شکل‌ها

۱-۲ 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 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 [؟]. ۱۴

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 یادگیری بدون برد، استفاده از اشتراک و تمایز برخی ویژگی‌ها میان دسته‌های مختلف در بینایی ماشین مورد بررسی قرار گرفته است [۱، ۲، ۳]. اما این روش‌ها به شناسایی دسته‌های کاملاً جدید از روی این ویژگی‌ها توجه نشان نداده‌اند. مسئله یادگیری تک‌ضرب^۱ هم یک مسئله نزدیک به یادگیری بدون برد است که پیش‌تر مورد بررسی بوده است [۴]. 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 [۵].

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

^۱One-shot Learning

^۲cold start

^۳Recommender Systems

^۴input 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 ψ نشان می‌دهیم.

۲. نگاشت توصیف‌ها به فضای مشترک که آن را با ϕ 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 $\mathcal{Y} = \mathcal{U} \cup \mathcal{S}$ نشان می‌دهیم که n_s تعداد دسته‌های آموزش را با n_u 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.

^۵model 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 i م از ماتریس X و x_i درایه i م از بردار x را نشان می‌دهد. ضرب داخلی با نماد $\langle \cdot, \cdot \rangle$ 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 عموماً 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 ψ 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.

۱-۲-۲ پیش‌بینی ویژگی مستقیم و غیر مستقیم

در [۱۰] 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 گرافیکی 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.

$$P(z_u|x) = \sum_{c \in \{0,1\}^a} P(u|c)p(c|x) \quad (1-2)$$

^۶One-Hot Encoding

^۷Attribute Prediction

^۸Direct Attribute Prediction

^۹Indirect Attribute Prediction

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

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

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

$$P(u|x) = \frac{P(u)}{P(c^u)} \prod_{n=1}^a P(a_n^u|x) \quad (2-2)$$

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:

$$P(c_i|x) = \sum_{k=1}^{n_u} P(y_k|x)\mathbb{I}(c_i = c_i^{y_k}) \quad (3-2)$$

در نهایت در هر دو روش برچسب نهایی با تخمین MAP^{۱۱} از رابطه زیر تعیین می‌شود:

$$\hat{y} = \arg \max_{u \in \mathcal{U}} P(u|x) = \arg \max_{u \in \mathcal{U}} \prod_{i=1}^a \frac{P(c_i^u|x)}{P(c_i^u)} \quad (4-2)$$

روش ارائه شده در [۴] 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 معمولاً برقرار نیست. برای مثال زمانی که ویژگی آیزی 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 [۴] 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 [۴] برای این کار یک چارچوب بر اساس مدل‌های گرافی احتمال معرفی می‌کنند. در

^{۱۱}Maximum a Posteriori

^{۱۱} Topic Modeling

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 برجسب است، با کمک روش‌های یادگیری ساختار^{۱۳} شناخته می‌شود.

۳-۲ 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} [۴] 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:

$$\text{softmax}(z)_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \quad j = 1, \dots, n_s. \quad (5-2)$$

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 بدون برد، هنگامی که یک نمونه از دسته‌های آزمون را به شبکه می‌دهیم، خروجی بدست آمده از رابطه (۴۴) 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 $\text{softmax}(x)$ را نشان دهد، یعنی n مین برجسب محتمل برای x از میان دسته‌های آموزش. حالا برای پیش‌بینی برجسب x از میان دسته‌های آموزش از این رابطه استفاده می‌کنیم:

$$\phi(x) = \frac{1}{Z} \sum_{n=1}^T P(\hat{y}(x, n)|x) \cdot c_{\hat{y}(x, n)}, \quad (6-2)$$

که T یک فراپارامتر مدل $Z = \sum_{n=1}^T P(\hat{y}(x, n)|x)$ 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} [۴] نیز از رویکرد مشابهی استفاده می‌کند.

^{۱۲} Bayesian Network

^{۱۳} Structure Learning

^{۱۴} Convec combination of Semantic Embeddings

^{۱۵} Activation Function

^{۱۶} Co-Occurance Statistics

در این روش همانند رابطه (۶-۲)، 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 برجسب‌ها پیشنهاد شده است.

۴-۲ 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 $x^T W$ 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:

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

در این حالت، این که فضای مشترک در حقیقت کدام یک از فضاهای تصاویر یا توصیفات هستند، جواب روشی 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.

۱-۴-۲ یادگیری با توابع رتبه‌بند

یک انتخاب متداول برای تابع هزینه، توابع رتبه‌بند^{۱۷} هستند. با توجه به این که عموماً 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 که تنها فاصله نقاط از برجسب خودشان را در نظر می‌گیرد [۹].

در [۹] تابع هزینه رتبه‌بند WSABIE [۹] 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

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

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

که در آن $r_\Delta(x_n, y_n) = \sum_{y \in \mathcal{Y}} \mathbb{I}(l(x_n, y_n, y) > \cdot)$ 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 [۴] ادامه داده شده و نگاشت شباهت ساخت‌یافته SJE^{۱۸} نامیده شده است. ، در این حالت تابع مطابقت بین توصیف‌ها و تصاویر از رابطه (۷-۲) تعریف می‌شود. تابع هزینه ساده‌تر از حالت قبل به صورت

$$\frac{1}{N_s} \sum_{n=1}^{N_s} \max(\cdot, l(x_n, y_n, y)) \quad (۱۰-۲)$$

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

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

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

که $\phi_k(y)$ توصیف‌های مختلف از دسته‌ی y را نشان می‌دهد و W_1, \dots, 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 دایره‌المعارفی با روش‌های word2vec [۴] و GloVe [۵] 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 [۴] 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 دوطرفه (۲-۷) از این رابطه تبعیت می‌کند:

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

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

در [۴] 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 فرایارامتر است که در این مقاله با اعتبار سنجی تعیین شده است. استخراج ویژگی از تصاویر با استفاده از شبکه عصبی کانولوشنال [۴] که روی دسته‌های دیده شده آموزش داده شده، انجام می‌شود. در نهایت یک تابع بیشترین حاشیه^{۲۰} برای یادگیری نگاشت دو خطی پیشنهاد می‌شود.

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

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.

$$\underset{W \in \mathbb{R}^{d \times a}}{\text{minimize}} \|X_s^T W C_s - Y\|_{Fro}^2 + \Omega(W) \quad (۱۴-۲)$$

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 هزینه مناسبی

^{۱۹}Latent Embedding Model

^{۲۰}Max margin

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

$$\Omega(W) = \lambda \|WC_S\|_{Fro}^2 + \gamma \|X_s^T W\|_{Fro}^2 + \lambda\gamma \|W\|_{Fro}^2 \quad (۱۵-۲)$$

این جمله منظم سازی 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 منظم سازی 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 [۹] 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:

$$W = V_x^T V_c \quad (۱۶-۲)$$

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:

$$\min_{V_x, V_c} \|X_s^T + V_x^T V_c C\|_{Fro}^2 + \lambda_1 \|V_x^T V_c C\|_{Fro}^2 + \lambda_2 \|V_c^T\|_{2,1} \quad (۱۷-۲)$$

که $\|M^T\|_{2,1} = \sum_i \|M_{(i)}\|_2$ و این نوع منظم سازی، 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^{۲۲} را خواهد داشت. جمله‌های منظم سازی دیگر در (۱۵-۲) به دلیل تاثیر اندکشان در آزمایشات عملی حذف شده‌اند.

^{۲۱}overfitting

^{۲۲}feature selection

۵-۲ نداشت به فضای تصاویر

در برخی از روش‌ها فضای مشترک فضای ویژگی‌های تصویر است و نداشتی 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 نداشت به فضای تصاویر بررسی می‌کنیم.

یک تعمیم از SJE در [۹] 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 نداشت 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 $\{(v_n, t_n, y_n), n = 1, \dots, N\}$ 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 معرفی شده 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 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 $F: \mathcal{V} \times \mathcal{T} \rightarrow \mathbb{R}$ میان تصاویر و توصیف‌هاست. که به صورت

$$F(v, t) = \theta(v)^T \phi(t) \quad (18-2)$$

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:

$$f_v(v) = \arg \max_{y \in \mathcal{Y}} (\mathbb{E}_{t \sim \mathcal{T}(y)} [F(v, t)]) \quad (19-2)$$

$$f_t(t) = \arg \max_{y \in \mathcal{Y}} (\mathbb{E}_{v \sim \mathcal{V}(y)} [F(v, t)]). \quad (20-2)$$

یادگیری تابع F با تابع هزینه‌ی زیر صورت می‌گیرد:

$$\frac{1}{N} \sum_{n=1}^N \ell_v(v_n, t_n, y_n) + \ell_t(v_n, t_n, y_n), \quad (21-2)$$

که توابع ℓ_t و ℓ_v این گونه تعریف شده‌اند:

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

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

تفاوت این تابع هزینه با رابطه (۱۰-۲) در اضافه شدن جمله‌ی دوم است. در رابطه (۱۰-۲) 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 مختلف 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 داده است. جمع‌آوری مجموعه دادگان متنی بزرگتر، آموزش چنین شبکه‌هایی را ممکن کرده است.

در [۴] 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:

$$y^* = \arg \max_y \phi(c^y)^T x \quad (22-2)$$

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 $c^T D x$ 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:

$$\phi(c) = \arg \min_{w, \zeta_i} w^T w - \alpha c^T D w - \beta \ln(P_{reg}(w|c)) + \gamma \sum \zeta_i \quad (2-23)$$

$$s.t. : -(w^T x_i) \geq \zeta_i, \quad \zeta_i \geq 0, \quad i = 1, \dots, N_s$$

$$c^T D c \geq l$$

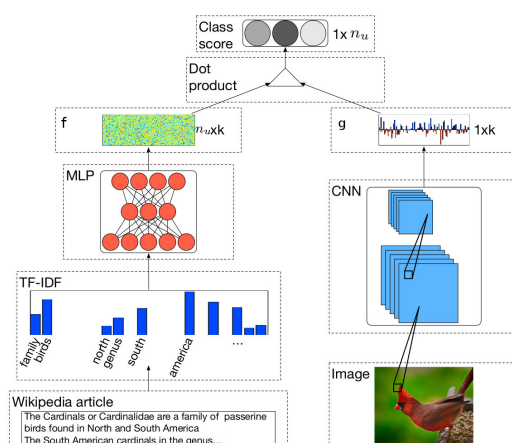
که α, β, γ, l فراپارامترهای مدل هستند. جمله اول در این تابع هزینه، منظم‌سازی دسته‌بند خطی w است. جمله دوم شباهت w با $c^T D$ را الزام می‌کند و جمله سوم احتمال بالا در رگرسیون را در نظر می‌گیرد. محدودیت $-(w^T x_i) \geq \zeta_i$ بر اساس فرض عدم تعلق نمونه‌های آزمون به کلاس‌های دیده‌شده تعریف شده است و 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 [۹].

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

^{۲۳}kernel trick



شکل ۲-۱: شبکه مورد استفاده برای یادگیری توأم 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 انگلیسی است پس از تبدیل به نمایش برداری به صورت tf-idf با چند لایه 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 [۴].

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 توأم 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:

$$y^* = \arg \max_y w^{yT} g(x). \quad (24-2)$$

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 یادگیری فیلتر 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 کانولوشنال 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 $K' \ll M$. 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: $w_c^y \in \mathbb{R}^{K' \times m \times m}$. همانند حالت قبل، f' با یک شبکه عصبی چند لایه مشخص می‌شود. در نهایت دسته‌بند کانولوشنال به صورت زیر تعریف می‌شود:

$$\text{score}(x, y) = o\left(\sum_{i=1}^{K'} w_i^{y'} \ast a'_i\right), \quad (25-2)$$

$\text{score}(x, y)$ 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 \ast 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:

$$\text{score}(x, y) = w^{yT} g(x) + o\left(\sum_{i=1}^{K'} w_i^{y'} \ast g'(a)_i\right). \quad (26-2)$$

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 [۹] 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 کاملاً متفاوت است. در این روش بردار ویژگی برای دسته‌ها جزو خروجی‌های روش است نه ورودی‌های آن. در این جا 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 قصد داشته باشیم بردار ویژگی‌های l بعدی تولید کنیم (l یک فرایارامتر 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:

$$y^* = \arg \min_i \|A_{(i)} - f(x)^T\| \quad (27-2)$$

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

$$\max_A \sum_{i,j} \|A_{(i)} - A_{(j)}\|_p^2 - \lambda \sum_{i,j} S_{ij} \|A_{(i)} - A_{(j)}\|_p^2 \quad (2-28)$$

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. (مثلاً بر اساس ضرب داخلی یا فاصله اقلیدسی 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.

فصل ۳

روش پیشنهادی

فصل ۴

نتایج

فصل ۵

جمع بندی

۱-۵ جمع بندی

۲-۵ کارهای آینده

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