علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۱- ۱۱
مقاله عادی

دسته‌بندی و حاشیه‌نویسی همزمان تصویر با استفاده از مدل‌های احتمالاتی موضوع و کدگذاری LLC کلمات بصری

سید نوید محمدی فومنی احمد نیک‌آبادی

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

تاکنون تلاش‌های زیادی به منظور استفاده از مدل‌های موضوعی نظیر مدل احتمالاتی LDA جهت دسته‌بندی و حاشیه‌نویسی همزمان تصاویر صورت گرفته است. اخیراً مدل‌های موضوع دیگری بر مبنای شبکه‌های عصبی احتمالاتی نظیر SupDocNADE معرفی شده‌اند که نتایج خوبی در مدل‌کردن داده‌های چندمقداری ارائه داده‌اند. در این مدل‌ها کلمات حاشیه‌نویسی نیز در کنار کلمات بصری تعبیه شده و به عنوان بردار ویژگی برای شبکه در نظر گرفته می‌شود. عدم‌تعادل در تعداد کلمات بصری و حاشیه‌نویسی سبب می‌شود تا سهم کلمات حاشیه‌نویسی برای بازنمایی در لایه پنهان شبکه‌عصبی مورد استفاده در این مدل، بسیار کمتر از کلمات بصری باشد. برای حل این مشکل در این مقاله، کلمات حاشیه‌نویسی در هیستوگرام بردار ویژگی وزن‌دهی می‌شوند. با افزودن قابلیت وزن‌دهی ورودی‌ها می‌توان از کدگذار LLC که چندین کلمه مشابه در فرهنگ لغت را بصورت وزن‌دار در ساخت بردار ویژگی دخیل می‌کند، برای تولید کلمات بصری استفاده نمود. با آزمایش مدل پیشنهادی بر روی پایگاه داده‌های UIUC_Sports و LabelMe، بهبود ۵ درصدی در معیار F در کلمات حاشیه‌نویسی و بهبود ۱ درصدی در دقت دسته‌بندی نسبت به مدل‌های موجود مشاهده می‌شود.

کلمات کلیدی: دسته‌بندی و حاشیه‌نویسی تصویر، مدل‌های موضوع، مدل احتمالاتی، شبکه عصبی، کدگذار LLC.

۱- مقدمه

گسترش روزافزون داده‌ها و ذخیره‌ی آنها بصورت دیجیتال مدیریت و سازماندهی آنها را به یک امر مهم تبدیل کرده است. در این میان، طبقه‌بندی این داده‌ها و اطلاعات به طور خودکار به یک چالش مهم تبدیل شده است. تصاویر و متون حجم عظیمی از این داده‌ها را به خود اختصاص داده‌اند. دسته‌بندی و حاشیه‌نویسی[۱] تصاویر از جمله مسائل پرکاربرد در حوزه پردازش تصاویر هستند. در مسأله دسته‌بندی تصاویر، هدف تعیین دسته یک تصویر ورودی از میان مجموعه‌ای از دسته‌های از قبل تعریف شده است. در مسأله حاشیه‌نویسی سعی می‌شود که مجموعه‌ای از کلمات کلیدی مرتبط با یک تصویر استخراج شود.

مدل‌های احتمالاتی موضوع[۲] یا تخصیص پنهان دیریکله[۳] را می‌توان به عنوان یکی از معروف‌ترین رویکردها در مدل کردن داده‌های چندمقداری[۴] دانست [۱، ۲، ۳] تخصیص پنهان دیریکله که به اختصار LDA نامیده می‌شود، مدل مولدی است که ابتدا در زمینه پردازش زبان‌های طبیعی معرفی شد. در این مدل هر سند بصورت توزیع چندجمله‌ای روی موضوعات تعریف شده است. از طرفی هر موضوع یک توزیع چندجمله‌ای روی کلمات می‌باشد. لازم به ذکر است که موضوعات بین همه سندها مشترک بوده ولی توزیع موضوعات برای هر سند، خاص آن سند می‌باشد [۴].

استفاده از مدل‌های احتمالاتی موضوع در بینایی ماشین با تعریف "کلمات بصری[۵]" آغاز شد [۵]. با استخراج کلمات بصری، هر تصویر تبدیل به یک سند می‌شود و حال می‌توان با آموزش LDA روی کیسه‌ای از کلمات بصری، تصاویر را مدل کرد. LDA یک مدل بدون ناظر بوده و برای کاربردهایی مانند دسته‌بندی مناسب نمی‌باشد. از این‌رو مدل تخصیص پنهان دیریکله با ناظر sLDA معرفی شد که با مدل کردن برچسب هر کلاس به همراه کلمات بصری، سبب استخراج ویژگی تمایزی می‌شود [۶]. اولین مدل جدی که دسته‌بندی و حاشیه‌نویسی را بطور

همزمان انجام داد را می‌توان تخصیص پنهان دیریکله با ناظر چند کلاسه[۶] (MC_sLDA) دانست [۷]. در این مدل و مدل‌های مشابه تغییرات جدی در مدل پایه‌ای LDA انجام شده تا بتواند یک فضای مشترک بین برچسب تصاویر و متن حاشیه برقرار کند [۸، ۹، ۱۰].

یکی از معایب این مدل‌ها عدم‌وجود استنتاج دقیق[۷] و قابل محاسبه[۸] می‌باشد. در این مدل‌ها برای بدست آوردن توزیع پسین[۹] مجبور به استفاده از روش‌های تقریبی هستیم که به مراتب کندتر بوده و بار محاسباتی بالایی دارند. برای حل مشکلات فوق، مدلی بر پایه ماشین بولتزمن محدود با نام سافت‌مکس تکراری[۱۰] معرفی شد [۱۱]. استنتاج برروی اسناد بازنمایی شده توسط این مدل، بسیار کاراتر از مدل‌های موضوع پیشین است. مهمترین مشکل مدل‌هایی که بر پایه ماشین بولتزمن محدود معرفی می‌شوند قابل محاسبه نبودن تابع نرمالیزه‌کننده یا همان تابع تقسیم[۱۱] است [۱۲]. عیب دیگر این روش‌ها ناتوانی در مدل کردن حالتی است که احتمال وقوع یک کلمه در هر موضوع خاص کم اما در ترکیبی از موضوعات زیاد است. به عبارت دیگر این کلمات توسط موضوعات استخراج شده، قابل تخمین نمی‌باشد. در مدل سافت‌مکس تکراری روش نمونه‌برداری مبتنی بر تابکاری (AIS[۱۲]) مورد استفاده قرار گرفته شده است که روشی مناسب جهت تخمین نسبت تابع تقسیم است [۱۳]. مهمترین عیب مدل سافت‌مکس تکراری وجود پیچیدگی زمانی خطی برابر با اندازه فرهنگ لغت، در زمان بروزرسانی (یادگیری) پارامترها می‌باشد.

شبکه عصبی تخمین‌گر توزیع اتورگرسیو[۱۳] (NADE) جایگزین مناسبی برای RBM[۱۴]ها است [۱۴]. این مدل بیشتر شبیه اتوانکدرها[۱۵] است که اندازه واحدهای ورودی و خروجی یکسان می‌باشد. در این مدل‌ها بدون نیاز به تخمین و با استفاده از روش‌های گرادیان، پارامترهای مدل آموزش داده می‌شود که این مزیت اصلی آنها به شمار می‌رود.

در [۱۵] مدل مولدی برای اسناد ارائه شده که از ترکیب NADE و سافت‌مکس تکراری الهام گرفته شده است. این مدل که با نام DocNADE شناخته می‌شود، توزیع توام را با استفاده از قاعده زنجیره‌ای تجزیه کرده و هر احتمال شرطی را مانند NADE توسط یک گره از شبکه مدل می‌کند. از این مدل علاوه بر استفاده در داده‌های متنی می‌توان در داده‌های چند مقداری نیز استفاده نمود. به دلیل بدون ناظر بودن این مدل و نیاز به استخراج ویژگی‌های تمایزی بیشتر، مدل باناظر آن با عنوان SupDocNADE معرفی شد [۱۶]. مدل SupDocNADE به‌دلیل با ناظر بودن یادگیری آن، توانایی استخراج ویژگی‌های تمایزی بیشتری دارد از این‌رو جهت مدل کردن توام برچسب کلاس، کلمات حاشیه‌نویسی و کلمات بصری استفاده شده است.

در تمامی مدل‌های ذکر شده جهت دسته‌بندی و حاشیه‌نویسی تصاویر از روش کیسه ویژگی‌ها [۱۶] استفاده شده است که به اختصار BoF نیز گفته می‌شود [۱۶،۹،۷]. در این روش هر تصویر توسط هیستوگرام تعداد رخداد الگوها بازنمایی می‌شود و مهمترین مزیت این روش‌ها کم بودن بارمحاسباتی آن است [۱۷، ۱۸].

در روش سبد ویژگی‌ها، اطلاعات مکانی مورد توجه قرار نمی‌گیرد و برای این منظور روش تطبیق هرم مکانی[۱۷] مطرح شد که به اختصار SPM گفته می‌شود. در روش SPM چند سطح مختلف از تصویر جهت بازنمایی نهایی استفاده می‌شود [۱۹، ۲۰]. این روش نیازمند یک دسته‌بند غیرخطی می‌باشد که محاسبات لازم جهت دسته‌بندی را افزایش می‌دهد. از طرفی محاسبه توصیف‌گر در سطوح مختلف دارای پیچیدگی زمانی بسیاری است. برای حل مشکل سرعت محاسبات در روش SPM و دقت پایین روش BoF روش کدگذاری خطی با محدودیت محلی بیان شد که به اختصار LLC گفته می‌شود. این روش کدگذاری جای چندی‌سازی[۱۸] برداری را در روش‌های گفته شده می‌گیرد [۲۱]. روش چندی‌سازی برداری در روش‌های ساخت کلمات بصری BoF و SPM استفاده شده است. چندی‌سازی در این روش‌ها یعنی تخصیص تنها نزدیک‌ترین کلمه بصری در فرهنگ لغات جهت بازنمایی هر تکه از تصویر. حال با استفاده از روش LLC چندین کلمه بصری بصورت وزن‌دار در بازنمایی تکه‌های استخراج شده از تصویر شرکت می‌کنند.

در مدل SupDocNADE کلمات حاشیه‌نویسی و کلمات بصری کنار هم تعبیه شده و بردار ورودی را تشکیل می‌دهند. از طرفی کلمات بصری بسیار بیشتر از کلمات حاشیه‌نویسی می‌باشند و این عدم‌تعادل سبب می‌شود تا مدل کارایی مناسبی در حاشیه‌نویسی تصاویر نداشته باشد. برای حل این مشکل، در این مقاله پیشنهاد می‌شود که کلمات حاشیه‌نویسی در هیستوگرام بردار ویژگی وزن‌دهی شود. این وزن‌دهی سبب می‌شود تا بردار کلمات حاشیه‌نویسی تاثیر مناسبی در بازنمایی پنهان مدل داشته باشند. از طرفی با افزودن قابلیت وزن‌دار شدن ورودی‌ها می‌توان از کدگذارهایی مانند LLC به جای روش‌های سنتی کوانتیزاسیون برداری استفاده نمود.

در بخش ۲ ابتدا مدل پایه‌ای DocNADE و نحوه‌ی مدل کردن اسناد متنی شرح داده شده و سپس نحوه‌ی گسترش این مدل جهت کار با داده‌های برچسب‌دار معرفی می‌شود. در ادامه این بخش روش‌های مختلف کدکردن کلمات بصری شرح داده شده است. همانطور که پیش‌تر بیان شد عدم‌تعادل بین تعداد کلمات بصری و کلمات حاشیه‌نویسی سبب مشکلاتی در یادگیری کلمات حاشیه می‌شود که در بخش ۳ روش پیشنهادی جهت حل این مشکلات مطرح شده است. در بخش ۴ آزمایشات مختلفی جهت بررسی کارایی مدل انجام شده و نتایج آن مقایسه شده است. در بخش پایانی جمع‌بندی روی مطالب گفته شده و روش پیشنهادی انجام شده است.

۲- مبانی نظری

در این بخش ابتدا مدل‌های موضوع بر پایه شبکه‌های عصبی شرح داده شده است و سپس دو روش مورد استفاده در ساخت کلمات بصری شرح داده می‌شود و در انتها نحوه‌ی کد کردن با استفاده از روش LLC بیان می‌شود.

۲-۱- مدل‌های DocNADE و SupDocNADE

در این بخش ابتدا به شرح شبکه عصبی تخمین‌گر توزیع اتورگرسیو برای اسناد (DocNADE) پرداخته شده است. سپس مدل باناظر DocNADE که برچسب کلاس را نیز مدل می‌کند شرح داده می‌شود. این مدل با دخیل کردن برچسب داده‌ها در روند آموزش ویژگی‌ها، تمایز بیشتری توسط ویژگی‌های پنهان مدل کسب می‌کند. در ادامه نحوه‌ی بهره‌برداری از اطلاعات مکانی توسط این مدل شرح داده شده و در انتها نحوه‌ی مدل کردن کلمات حاشیه به همراه سایر ویژگی‌ها در مدل SupDocNADE شرح داده شده است.

۲-۱-۱- DocNADE

DocNADE به عنوان رویکردی برای مدل کردن اسناد معرفی شده است. برای استفاده از این مدل زمانی که نوع ورودی تصویر باشد لازم است که هر تصویر به صورت کیسه کلمات بصری بازنمایی شود. ابتدا تصویر به تکه‌های مشخص تقسیم می‌شود سپس برروی هر تکه توصیف‌گر SIFT اعمال می‌شود. حال با اعمال الگوریتم خوشه‌بندی نماینده‌هایی از این توصیف‌گرها به عنوان کلمات بصری در نظر گرفته می‌شوند (برای اطلاعات بیشتر به بخش ۴-۲ مراجعه شود). حال هر تصویر می‌تواند بصورت کیسه‌ای از کلمات بصری $v = [v_1, v_2, \dots, v_D]$ بازنمایی

شود که v_i اندیس نزدیکترین خوشه به iامین توصیف‌گر SIFT می‌باشدو همچنین D تعداد توصیف‌گرهای استخراج شده از تصویر را مشخص می‌کند [۱۵].

توزیع توام کلمات بصری $p(v)$ در DocNADE بصورت احتمالات شرطی $p(v_i|v_{<i})$ بازنویسی می‌شود:

$$p(v) = \prod_{i=1}^{D} p(v_i|v_{<i}) \quad (۱)$$

که در آن بردار $v_{<i}$ شامل همه v_jهای می‌شود که $j < i$ است. لازم به ذکر است که رابطه (۱) برای تمامی توزیع‌ها براساس قانون زنجیره‌ای احتمال صادق است. در واقع مهمترین فرض DocNADE این است که احتمالات شرطی می‌توانند توسط یک شبکه عصبی پیشخور مدل شوند.

یکی از معماری‌هایی که می‌توان برای مدل کردن $p(v_i|v_{<i})$ در نظر گرفت به صورت زیر است:

$$h_i(v_{<i}) = g\left(c + \sum_{k<i} W_{:,v_k}\right) \quad (۲)$$

$$p(v_i = w|v_{<i}) = \frac{\exp\left(b_w + V_{w,:}h_i(v_{<i})\right)}{\sum_{w'} \exp\left(b_{w'} + V_{w',:}h_i(v_{<i})\right)} \quad (۳)$$

که $g(.)$ تابع فعالیت درایه به درایه غیرخطی است، $W \in \mathbb{R}^{H\times K}$ و $V \in \mathbb{R}^{K\times H}$ ماتریس اتصال، $b \in \mathbb{R}^{K}$ و $c \in \mathbb{R}^{N}$ پارامترهای بایاس و H، K به ترتیب تعداد واحدهای پنهان (موضوعات) و اندازه فرهنگ‌لغت می‌باشد.

محاسبه توزیع $p(v_i = w|v_{<i})$ نیازمند محاسبات خطی از مرتبه K بوده و تکرار این کار به تعداد کلمات بصری D بسیار زمان‌بر است. برای حل این مشکل از درخت دودویی استفاده شده است. استفاده از درخت دودویی محاسبات را از مرتبه خطی K به لگاریتمی تبدیل می‌کند. برای این کار هر برگ درخت به صورت تصادفی به یک اندیس کلمه تخصیص پیدا می‌کند و احتمال آن کلمه برابر با ضرب احتمالات موجود در مسیر ریشه تا آن برگ تعریف می‌شود. احتمال انتقال به چپ/راست در درخت توسط تابع رگرسیون لاجستیک دودویی مدل شده که ورودی آن $h_i(v_{<i})$ می‌باشد.

فرض کنید $l(v_i)$ نشان‌دهنده‌ی دنباله گره‌های درخت از ریشه به برگ v_i و $\pi(v_i)$ نشان‌دهنده دنباله دودویی انتخاب چپ/راست گره در مسیر باشد. به عنوان مثال $l(v_i)_1$ همواره گره‌ی ریشه و $\pi(v_i)_1$ برابر ۱ است اگر v_i در زیردرخت چپ باشد و برابر صفر در حالت عکس آن. حال $V \in \mathbb{R}^{T\times H}$ ماتریسی شامل وزن‌های رگرسیون لاجستیک است و $b \in \mathbb{R}^{T}$ بردار بایاس که T تعداد گره‌های داخلی درخت دودویی و H تعداد واحدهای مخفی می‌باشد. احتمال $p(v_i = w|v_{<i})$ بصورت زیر مدل شده است:

$$p(v_i = w|\mathbf{v}_{<i}) = \prod_{k=1}^{|\pi(v_i)|} p(\pi(v_i)_k|\mathbf{v}_{<i}) \quad (۴)$$

که خروجی رگرسیون لاجستیک گره‌های میانی

$$p(\pi(v_i)_k = 1|v_{<i}) = \mathrm{sigm}\left(b_{l(v_i)_m} + V_{l(v_i)_m,:}h_i(v_{<i})\right) \quad (۵)$$

و $\mathrm{sigm}(x) = 1/(1+\exp(-x))$ تابع سیگموئید و $m \in \{1,2,\dots,|\pi(v_i)|\}$ است. تابع رگرسیون لاجستیک رابطه (۴) از مرتبه $O(\log K)$ است.

بنابر آنچه گفته شد، با ترکیب روابط (۲، ۴ و ۵) می‌توان احتمال $p(v) = \prod_{i=1}^{D} p(v_i|v_{<i})$ را برای هر سندی محاسبه کرد. یادگیری پارامترهای $\{W, V, b, c\}$ مدل توسط بهینه کردن درست‌نمایی داده‌های آموزشی با استفاده از رویکرد نزول در امتداد گرادیان حاصل می‌شود.

پس از آموزش مدل، بازنمایی پنهان سند جدید v^* بصورت زیر استخراج می‌شود:

$$h_y(v^*) = g\left(c + \sum_{i}^{D} W_{:,v_i^*}\right) \quad (۶)$$

این بازنمایی می‌تواند به عنوان ورودی یک دسته‌بند جهت اعمال باناظر بینایی ماشین در نظر گرفته شود. در واقع اندیس y نشان‌دهنده استفاده این بازنمایی در تخمین برچسب کلاس تصویر می باشد [۸].

SupDocNADE -۲-۱-۲

مشاهده شده است که استخراج ویژگی‌های بصری از تصویر، با استفاده از مدل‌های احتمالاتی موضوع مانند LDA نمی‌توانند نتایج مناسبی برای کاربردهایی همچون دسته‌بندی داشته باشند. یکی از دلایل این امر را می‌توان در نوع آموزش ویژگی‌ها دانست. ویژگی‌هایی که به صورت بدون ناظر از تصاویر استخراج می‌شوند برای توصیف ساختار آماری تصویر آموزش داده شده‌اند و این ویژگی‌ها نمی‌توانند ساختاری را استخراج کنند تا تمایز بین کلاسی حداکثر شود. این موضوع سبب ابداع انواع مختلف LDA مانند sLDA شده است [۶، ۷]. DocNADE نیز یک مدل موضوع بدون ناظر می‌باشد از این رو SupDocNADE معرفی شد تا ویژگی‌های مناسبی را جهت کاربرد دسته‌بندی استخراج نماید [۱۶].

الف) آموزش مدل

بطور خاص اگر تصویر $v = \{v_1, v_2, \dots, v_D\}$ و برچسب کلاس $y \in \{1, \dots, C\}$ به عنوان ورودی مدل باشد آنگاه توزیع توام SupDocNADE بصورت زیر تعریف می‌شود:

$$p(\mathbf{v}, y) = p(y|\mathbf{v}) \prod_{i=1}^{D} p(v_i|\mathbf{v}_{<i}) \quad (۷)$$

و مانند DocNADE احتمالات شرطی توسط واحدهای شبکه عصبی مدل می‌شود. در این مدل نیز از معماری مشابه DocNADE برای $p(v_i|v_{<i})$ استفاده شده است و تنها نیازمند تعریف مدلی برای $p(y|\mathbf{v})$ می‌باشد. از آنجایی که $h_y(v)$ بازنمایی تصویر است، می‌توان از آن جهت دسته‌بندی استفاده کرد. از اینرو $p(y|v)$ بصورت یک رگرسیون لجستیک چند کلاسه مدل می‌شود که نحوه‌ی محاسبه آن از روی $h_y(v)$ بصورت زیر است:

$$p(y|v) = \mathrm{softmax}\left(d + Uh_y(v)\right)_y \quad (۸)$$

که تابع $\mathrm{softmax}(a)_i = \exp(a_i) / \sum_{j=1}^{C} \exp(a_j)$ و $d \in \mathbb{R}^{C}$ پارامتر بایاس برای لایه با ناظر و $U \in \mathbb{R}^{C\times H}$ ماتریس اتصال بین لایه پنهان h_y و برچسب کلاس است.

به عبارت دیگر $p(y|v)$ به صورت یک شبکه عصبی چند کلاسه مدل می‌شود که ورودی آن کیسه‌ای از کلمات بصری است. تفاوت اساسی این مدل با شبکه‌های عصبی در این است که برخی پارامترهای آن (پارامترهای پنهان W و c) نیز جهت مدل کردن احتمال شرطی کلمات بصری $p(v_i|v_{<i})$ استفاده می‌شوند. برای بیشینه کردن درست‌نمایی مدل باید تابع

$$-\log p(v, y) = -\log p(y|v) + \sum_{i=1}^{D} -\log p(v_i|v_{<i}) \quad (۹)$$

روی تصاویر آموزشی کمینه شود که این به عنوان یادگیری از نوع مولد شناخته می‌شود [۱۱]. عبارت اول در رابطه فوق بطور کامل تمایزی است، در حالی که عبارت دوم بدون ناظر بوده و می‌تواند به عنوان تنظیم کننده در نظر گرفته شود. این عبارت تنظیم کننده بصورت بدون ناظر ساختار آماری بین کلمات بصری را مدل می‌کند. در عمل این تنظیم‌کننده می‌تواند جواب را به سمتی بایاس کند که خاصیت تمایزی زیادی نداشته باشد و سبب تعمیم‌پذیری بهتر مدل شود. همانند کارهای قبلی می‌توان ترکیبی از یادگیری مولد/تمایزی را در نظر گرفت که این امر با اعمال وزن‌دهی صورت می‌گیرد.

$$-\log p(v,y) = -\log p(y|v) + \lambda \sum_{i=1}^{D} -\log p(v_i|v_{<i}) \qquad (۱۰)$$

که λ به عنوان پارامتر تنظیم کننده رفتار می‌کند. جهت بهینه کردن رابطه (۱۰) بر روی مجموعه آموزشی از روش نزول در امتداد گرادیان تصادفی، و از قانون پس انتشار خطا برای محاسبه مشتق پارامترها استفاده شده است.

ب) موقعیت مکانی ویژگی‌های بصری

اطلاعات مکانی همواره نقش مهمی در درک تصاویر داشته است. به عنوان مثال آسمان همواره در قسمت بالای تصویر دیده می‌شود و یا ماشین همواره در نیمه پایینی تصویر به چشم می‌خورد. بسیاری از کارهای پیشین از این اطلاعات بطور موثر در کار خود استفاده کرده‌اند [۱۶].

فرض کنید تصویر به چند ناحیه مجزا $R = \{R_1, R_2, \dots, R_M\}$ تقسیم شودکه M تعداد نواحی مختلف می‌باشد. حال یک تصویر می‌تواند به صورت زیر بازنمایی شود

$$v^R = [v_1^R, v_2^R, \dots, v_D^R] = [(v_1, r_1), (v_2, r_2), \dots, (v_D, r_D)] \qquad (۱۱)$$

که $r_i \in R$ ناحیه‌ای است که کلمه بصری v_i از آن استخراج شده است. برای مدل کردن احتمال توام کلمات بصری ابتدا توزیع بصورت $\prod_i p((v_i, r_i)|v_{<i}^R)$ تجزیه می‌شود و با هر یک از $K \times M$ جفت کلمه/ناحیه بصورت یک کلمه بصری مجزا رفتار می‌شود. این امر به این معنی است که به ازای هر حالت یک برگ به درخت دودویی افزوده می‌شود و از طرفی چون محاسبات بصورت لگاریتمی با اندازه برگ‌های درخت رشد می‌کند مشکلی با افزایش تعداد مناطق نخواهیم داشت.

ج) افزودن کلمات حاشیه‌نویسی

در این بخش به شرح چگونگی مدل کردن کلمات حاشیه‌نویسی پرداخته می‌شود. فرض کنید A فرهنگ‌لغت از پیش تعریف شده برای کلمات حاشیه‌نویسی باشد. برای یک سند خاص کلمات حاشیه‌نویسی بصورت $a = [a_1, a_2, \dots, a_L]$ نمایش داده شود که $a_i \in A$ و L تعداد کلمات حاشیه‌نویسی می‌باشند. حال یک تصویر به همراه کلمات حاشیه‌نویسی آن بصورت ترکیبی از کلمات بصری و کلمات حاشیه‌نویسی می‌توانند بازنمایی شوند:

$$\begin{aligned} v^A &= [v_1^A, v_2^A, \dots, v_D^A, v_{D+1}^A, \dots, v_{D+L}^A] \\ &= [v_1^R, v_2^R, \dots, v_D^R, a_1, a_2, \dots, a_L] \end{aligned} \qquad (۱۲)$$

بطور خاص در این مدل کلمات حاشیه‌نویسی بصورت توام با کلمات بصری اندیس‌دهی می‌شوند و همچنین برگ‌هایی از درخت دودویی برای اندیس کلمات حاشیه‌نویسی تخصیص داده می‌شود.

۲-۲- ساخت کلمات بصری

در این بخش سه روش کد کردن سبد کلمات، روش SPM و LLC جهت ساخت کلمات بصری شرح داده شده است.

۲-۲-۱- سبد کلمات

این روش که از روش‌های مبتنی بر فرهنگ لغت محسوب می‌شود به صورت زیر عمل می‌کند:

۱- استخراج نقاط کلیدی مربوط به همه تصاویر آموزشی (از همه دسته‌ها)

۲- حصول توصیف‌گر حول تمام نقاط کلیدی

۳- خوشه‌بندی تمام توصیف‌گرهای بدست آمده با روش k-means برای حصول فرهنگ‌لغت. اعضای این فرهنگ لغت همان کلمات بصری هستند که فرض می‌شود هر تصویر با استفادهاز این کلمات بصری ایجاد شده است (در پردازش زبان‌ها طبیعی تفسیر به این گونه است که هر جمله از تعدادی کلمه در فرهنگ لغت استفاده می‌کند و در اینجا هر تصویر از تعدادی کلمه بصری).

پس از حصول فرهنگ لغت، زمان آن می‌رسد که بازنمایی جدیدی برای هر یک از تصاویر موجود در مجموعه دادگان مورد استفاده بدست آید. این روند برای هر تصویر در ذیل آورده شده است:

۱- استخراج نقاط کلیدی مربوط به تصویر و حصول توصیف‌گرهای حول تمام این نقاط کلیدی (همانند فاز یادگیری فرهنگ لغت)

۲- مرحله کد کردن: یافتن نزدیک‌ترن کلمه بصری موجود در لغت نامه به هر توصیف‌گر و تخصیص عدد مربوط به آن کلمه بصری (یا همان مرکز خوشه) به آن توصیف‌گر .

۳- ایجاد هیستوگرام مبتنی بر تعداد کلمات بصری موجود در تصویر.

هیستوگرام به دست آمده از روش فوق بازنمایی جدید آن تصویر را نمایش می‌دهد [۱۷]. در ادامه روش‌های دیگری بیان می‌شود که ساختار دیگری را برای ساخت کلمات بصری بیان می‌کنند.

۲-۲-۲- روش تطبیق هرمی مکانی (SPM)

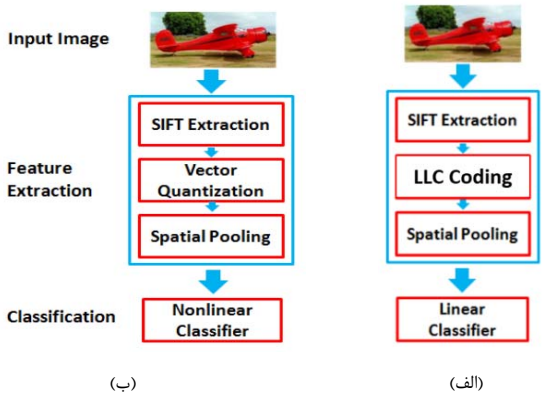
در روش سبد کلمات، اطلاعات مکانی نقاط کلیدی مورد توجه قرار گرفته نمی‌شود (این امر از نام سبد کلمات نیز بر می‌آید). بنابر این، یک منبع اطلاعاتی ارزشمند که می‌توانست در دسته‌بندی تصاویر کارا باشد مورد استفاده قرار نمی‌گرفت. در روش تطبیق هرمی مکانی، فاز یادگیری فرهنگ لغت همچون روش اصلی سبد کلمات انجام می‌گرفت، اما این بار با یک ایده ساده، چند سطح مختلف برای تصویر در نظر گرفته شده است که در هر سطح، تصویر به چند بخش با اندازه مساوی تقسیم می‌شود و برای هر بخش یک هیستوگرام مبتنی بر کلمات بصری ایجاد می‌شود.

در انتها هیستوگرام‌های حاصل با اعمال یک ضریب (ضریبی که برای سطوح بالاتر، بزرگتر و برای سطوح پایین‌تر، کوچکتر است تا اهمیت هیستوگرام‌هایی که اطلاعات مکانی بیشتر دارند افزایش یابد)، به همدیگر الحاق شده تا بازنمایی نهایی را شکل دهند. پس از حصول بازنمایی نهایی در روش تطبیق هرم مکانی، از یک دسته‌بندی کننده قوی همچون ماشین بردار پشتیبان برای دسته‌بندی تصویر استفاده شده است. این ایده ساده توانست دقت را به نسبت روش سبد کلمات تا حد بسیارخوبی افزایش دهد [۲۰].

۲-۲-۳- کد کردن LLC

روش اولیه تطبیق هرم مکانی برای عملکرد خوب نیاز داشت تا از یک دسته‌بندی کننده غیرخطی (برای مثال SVM با کرنل توابع پایه گوسی[۱۹]) استفاده کند [۲۱]. با توجه به اینکه مرتبه زمانی آموزش یک دسته‌بند SVM غیرخطی از درجه $O(n^2)$ تا $O(n^3)$ بود (n برابر تعداد تصاویر آموزشی است)، بنابراین توسعه‌پذیری روش سنتی تطبیق هرم مکانی پایین بود. روش LLC با جزئیاتی که در ادامه شرح داده می‌شود، با ایجاد یک کد جدید برای هر یک از توصیف‌گرهای صحنه، موفق شد تا به یک بازنمایی جدید از تصویر برسد که این بازنمایی با یک دسته‌بندی کننده خطی که با مرتبه زمانی $O(n)$ هم قابل آموزش بود، موفق شد که زمان آموزش را به شکل خوبی کاهش، و دقت حاصل را افزایش دهد.

پیش از پرداختن به نحوه حصول کد توسط روش LLC بهتر است تا دقیقاً جایگاه تغییر ایجاد شده در روش تطبیق هرم مکانی را مشخص کنیم. تفاوت این دو روش در شکل ۱ نمایش داده شده است.

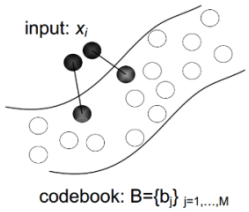


شکل ۱- محل ایجاد تغییر در LLC به نسبت روش تطبیق هرم مکانی (اقتباس از [۲۲]). الف) LLC خطی ب) SPM غیرخطی

همانطور که در شکل مشاهده می‌شود، در روش LLC که در ستون راست تصویر واقع شده، کد کردن LLC جای کوانتیزاسیون برداری[۲۰] را گرفته است. روش کوانتیزاسیون برداری که در سبد کلمات و تطبیق هرم مکانی وجود داشت به این صورت بود که هنگام تخصیص یکی از کلمات بصری فرهنگ لغت به یکی از توصیف‌گرها، تنها نزدیک‌ترین کلمه بصری در نظر گرفته شده و مابقی کلمات بصری نادیده گرفته می‌شدند. به زبان ریاضی، اگر فرض کنیم هر تصویر دارای N نقطه کلیدی است که توصیف‌گرهایش را بدست آورده‌ایم، و هر توصیف‌گر قرار است با یک کد M بعدی جایگزین شود (M تعداد کلمات بصری فرهنگ لغت است)، می بایست رابطه بهینه‌سازی زیر حل شود:

$$arg\ \min_{c} \sum_{i=1}^{N} \|x_i - Bc_i\|^2$$
$$s.t.\ \|c_i\|_{l^0} = 1\ .\ \|c_i\|_{l^1} = 1\ .\ c_i \geq 0.\forall i \qquad (۱۳)$$

که در رابطه فوق، $B = \{b_j\}_{j=1,\dots,M}$ نمایش دهنده فرهنگ لغت و c_i کد جدیدی بوده که قرار است جایگزین توصیف‌گر x_i شود. در صورت حل رابطه بهینه‌سازی فوق که شرایط نرم l_0 و نرم l_1 آن آورده شده است، درواقع گویی برای توصیف‌گر ورودی x_i به دنبال نزدیک‌ترین عضو فرهنگ لغت بوده‌ایم که مفهوم آن در شکل ۲ به تصویر کشیده شده است.



شکل ۲- تخصیص نزدیکترین عضو فرهنگ لغات به توصیف‌گر ورودی در روش کوانتیزاسیون برداری [۲۱]

اما روش LLC برای یافتن کد مربوط به هر توصیف‌گر سعی در بهینه‌سازی رابطه دیگری دارد. این رابطه بهینه‌سازی در ذیل آورده شده است:

$$\arg\min_{c} \sum_{i=1}^{N} \|x_i - Bc_i\|^2 + \lambda \|d_i \odot c_i\|^2$$
$$s.t.\ 1^T c_i = 1.\forall i \qquad (۱۴)$$

که در رابطه فوق:

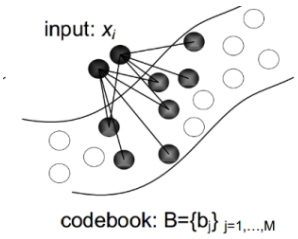
$$d_i = \exp\left(\frac{dist(x_i.B)}{\sigma}\right)$$
$$dist(x_i.B) = [dist(x_i.b_1).\dots.dist(x_i.b_M)]^T \qquad (۱۵)$$

ترم اول این رابطه بهینه‌سازی با قیدی که ذکر شده است، در صدد است تا به جای اینکه همچون روش کوانتیزاسیون برداری که یک کد M بعدی به صورت $[0,0,\dots,1,0,\dots,0]$ که عنصر مربوط به نزدیک‌ترین کلمه بصری آن ۱ است را تولید می‌کند، یک کد M بعدی تولید کند که خطای کوانتیزاسیون را با دخیل کردن تمام کلمات بصری کمینه می‌کند. اما ترم دوم رابطه بهینه‌سازی سعی دارد تا معیار مجاورت[۲۱] را ارضا کند و این امر باعث می‌شود تا ضریب مربوط به کلمات بصری نزدیک‌تر در کد حاصل بزرگتر شود. پارامتر σ نحوه محاسبه فاصله بین توصیف‌گر و اعضای فرهنگ لغت را کنترل می‌کند و به ازای یک σ ثابت نیز، λ میزان وزن‌دهی به مجاورت را مشخص می‌کند. هر چه λ بزرگتر شود وزن مجاورت نیز بزرگتر می‌شود. به این معنا که تنها به همسایه‌های نزدیک توصیف‌گر x_i در فرهنگ لغت ضرایب غیر صفر می‌دهیم. در شکل ۳ مثالی از نحوه کد کردن روش LLC آورده شده است. یال‌های متصل به اعضای فرهنگ لغت نشان دهنده این هستند که این اعضا به طور خطی و با وزن‌هایی که برایشان طی حل مسأله بهینه‌سازی بدست می‌آید، توصیف‌گر x_i را بازسازی می‌کنند.

یکی دیگر از نکات جذاب در مورد روش LLC این است که رابطه بهینه‌سازی ارائه شده یک راه حل تخمینی از مرتبه $O(M+K^2)$ به جای $O(M^2)$ دارد که M تعداد کلمات بصری در فرهنگ لغت و K تعداد همسایه‌های در نظر گرفته شده برای وزن دهی در LLC است که $K \ll M$ است. برای این کار بجای بهینه کردن رابطه (۱۴) می‌توان از روش k نزدیکترین همسایه جهت پیدا کردن توصیف‌گرهای مجاور استفاده نمود و رابطه زیر که بسیار ساده تر است را حل نمود.

$$\min_{\tilde{c}} \sum_{i=1}^{N} \|x_i - \tilde{c}_i B_i\|^2$$
$$s.t.\ 1^T \tilde{c}_i = 1,\ \forall i \qquad (۱۶)$$

این موضوع، استفاده از این روش در کاربردهای بلادرنگ را نیز امکان‌پذیر کرده است.



شکل ۳- وزن‌دهی بیشتر به همسایه‌های نزدیک توصیف‌گر در لغت نامه در روش LLC [۲۱]

آخرین مطلبی که باید ذکر شود، این است که پس از حصول کدهای مربوط به روش LLC برای هر توصیف‌گر، بازنمایی نهایی براساس انباشت این کدها صورت می‌گیرد که در زیر نحوه‌ی انباشت حداکثری که بهترین روش انباشت در LLC است نشان داده شده است.

انباشت حداکثر: اگر c_{ij} نشان‌دهنده عنصر jام مربوط به کد توصیف‌گر iام باشد، انباشت حداکثر به صورت ذیل است:

$$\text{finalCode}_j = \max(c_{1j}, c_{2j}, \dots, c_{Nj}) \qquad (۱۷)$$

۳-روش پیشنهادی

در شکل ۴ بلاک دیاگرام مربوط به مدل مورد نظر جهت دسته‌بندی و حاشیه‌نویسی تصویر نشان داده شده است. در قسمت پیش‌پردازش ابتدا هر تصویر تکه‌تکه شده و به توصیف‌گر داده می‌شود. پس از حصول خروجی توصیف‌گرهای مربوط به تصویر توسط کدگذار بازنمایی مربوط به هر تصویر تولید می‌شود. در بخش آموزش کلمات حاشیه‌نویسی در کنار کلمات بصری تعبیه شده‌اند و به عنوان بردار ویژگی به مدل SupDocNADE داده شده است. در قسمت آزمایش مدل فقط خروجی حاصل از توصیف‌گر تکه‌های تصویر به مدل داده شده و برچسب کلاس به همراه کلمات حاشه نویسی توسط مدل تولید می‌شود.

در عمل تعداد کلمات بصری استخراج شده از تصویر بسیار بزرگتر از کلمات حاشیه‌نویسی است. به عنوان مثال از یک تصویر با اندازه ۳۰۰×۴۰۰ حدود ۲۰۰۰ کلمه بصری استخراج می‌شود (شکل ۴ قسمت پیش‌پردازش) این در حالی است که کلمات حاشیه‌نویسی بین ۵ الی ۲۰ عدد متغیر می‌باشند.

عدم‌تعادل بین کلمات بصری و کلمات حاشیه‌نویسی ممکن است سبب برخی مشکلات شود. به عنوان مثال سهم کلمات حاشیه‌نویسی برای بازنمایی در لایه پنهان بسیار کمتر از کلمات بصری است. از طرفی هر کلمه به نسب تعداد تکرار خود به کل کلمات استخراج شده بر روی گرادیان خطا تاثیر می‌گذارد. پس گرادیانی که از کلمات حاشیه تولید می‌شود بسیار کوچک بوده تا بتواند تاثیر با معنایی در افزایش احتمال شرطی حاصل از کلمات حاشیه داشته باشد.

برای حل این مشکل، پیشنهاد می‌شود که کلمات حاشیه‌نویسی در هیستوگرام بردار ویژگی $v_{<i}$ وزن‌دهی شوند. برای این منظور یک بردار ω به همراه v در نظر گرفته می‌شود. حال بردار ورودی به صورت زیر تعریف می‌شود:

$$\tilde{\mathbf{v}} = \left[\left(v_1, \omega(v_1)\right), \left(v_2, \omega(v_2)\right), \dots, \left(v_{D+L}, \omega(v_{D+L})\right)\right] \qquad (۱۸)$$

بردار ω برای مولفه‌های متناظر با کلمات بصری برابر ۱ و برای مولفه‌های متناظر با کلمات حاشیه برابر ρ می‌باشد. حال رابطه (۱۰) بصورت زیر بازنویسی می‌شود:

$$-\log p(v,y) = -\log p(y|v) - \lambda \sum_{i=1}^{D} \omega(v_i) \log p(v_i|v_{<i})) \qquad (۱۹)$$

و در رابطه (۲) نیز بجای v از $\tilde{v}$ استفاده می‌شود.

با وزن‌دهی کلمات حاشیه، مدل با توجه بیشتر به این کلمات سبب کاهش مشکلات عدم‌تعادل بین کلمات بصری و کلمات حاشیه‌نویسی خواهد شد. در عمل مقدار وزن ρ توسط اعتبارسنجی متقابل[۲۲] مشخص می‌شود. در الگوریتم ۱ شبه کد بروزرسانی گرادیان پارامترهای رابطه (۱۹) براساس روش پیشنهادی نشان داده شده است.

الگوریتم ۱- نحوه محاسبه گردایان مدل برروی داده‌های آموزشی

Input: training vector $\mathbf{v}$, training weight vector ω_D
unsupervised learning weight λ
Output: gradient of Equation 14 w. r. t parameters
$f(\mathbf{v}) \leftarrow \text{softmax}(\mathbf{d} + \mathbf{U}\mathbf{h}^{c}(\mathbf{v}))$
$\delta\mathbf{d} \leftarrow (f(\mathbf{v}) - 1_y)$
$\delta\mathbf{act} \leftarrow (\mathbf{U}^T\delta\mathbf{d}) o 1_{h_y>0}$
$\delta\mathbf{c} \leftarrow 0, \delta\mathbf{b} \leftarrow 0, \delta\mathbf{V} \leftarrow 0, \delta\mathbf{W} \leftarrow 0$
for i from D to 1 do
  $\delta\mathbf{h}_i \leftarrow 0$
for m from 1 to $|\pi(v_i)|$ do
  $\delta t \leftarrow \frac{\lambda\omega(v_i)(p(\pi(v_i)_m|\mathbf{v}_{<i}) - \pi(v_i)_m)}{D}$
  $\delta b_{l(v_i)_m} \leftarrow \delta b_{l(v_i)_m} + \delta t$
  $\delta\mathbf{V}_{l(v_i)_m,:} \leftarrow \delta\mathbf{V}_{l(v_i)_m,:} + \delta t\mathbf{h}_i^T$
  $\delta\mathbf{h}_i \leftarrow \delta\mathbf{h}_i + \delta t\mathbf{V}_{l(v_i)_m,:}^T$
end for
  $\delta\mathbf{act} \leftarrow \delta\mathbf{act} + \delta\mathbf{h}_i o 1_{\mathbf{h}_i>0}$
  $\delta\mathbf{c} \leftarrow \delta\mathbf{c} + \delta\mathbf{h}_i o 1_{\mathbf{h}_i>0}$
  $\delta\mathbf{W}_{:,v_i} = \omega(v_i)\delta\mathbf{W}_{:,v_i} + \delta\mathbf{act}$
end for

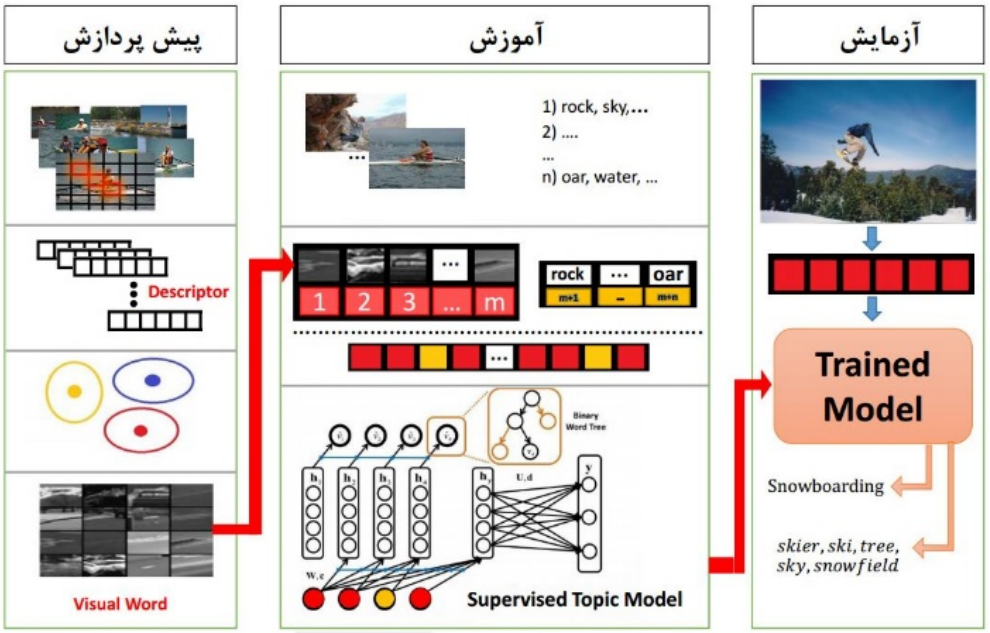
با افزودن قابلیت وزن‌دار شدن ورودی‌های مدل، می‌توان از LLC استفاده نمود. همانطور که در بخش ۲-۲-۳ بیان شد روش LLC برخلاف روش‌های سنتی که تنها از یک کلمه جهت بازنمایی استفاده می‌نمایند، چندین کلمه مشابه در فرهنگ لغت را بصورت وزن‌دار در ساخت بردار ویژگی دخیل می‌کند. حال هر تکه از تصویر توسط یک بردار وزن بازنمایی شده است. پس از اعمال توصیف‌گر و حصول بازنمایی، انباشت حداکثری روی بردار نهایی طبق رابطه (۱۷) اعمال می‌شود. بردار حاصل از انباشت به عنوان کلمات بصری بوده و مقادیر درایه‌های آن وزن ورودی می‌باشد. این کلمات بصری با کلمات حاشه‌نویسی و وزن‌های آن به عنوان ورودی به مدل داده می‌شود.

۴- نتایج و آزمایشات

برای ارزیابی عملکرد مدل پیشنهادی در دسته‌بندی و حاشیه‌نویسی تصاویر از دو پایگاه‌دادهی LabelMe [۲۳] و UIUC-Sports [۱] استفاده شده است. این پایگاه داده‌ها به عنوان مجموعه داده‌های پایه جهت حاشیه‌نویسی و دسته‌بندی می‌باشند که هر تصویر به همراه کلمات حاشیه‌نویسی موجود است. مدل پیشنهادی با مدل‌های SupDocNADE [۱۶]، DocNADE [۱۵] و Mc_sLDA [۷] مقایسه شده است.

۴-۱- معرفی مجموعه داده‌ها

در این مقاله از دو مجموعه داده استفاده شده است. مجموعه داده اول UIUC-Sports می‌باشد که شامل ۱۷۹۲ تصویر در ۸ دسته ورزشی تقسیم شده است.



شکل ۴- بلوک دیاگرام مربوط به مدل جهت دسته‌بندی و حاشیه‌نویسی تصویر

برچسب هر کلاس به علاوه تعداد تصاویر بصورت ۱- بدمینتون (۳۱۳ تصویر) ۲- بوچی (۱۳۷ تصویر) ۳- کراکت (۳۳۰ تصویر) ۴- چوگان بازی (۱۸۳ تصویر) ۵- صخره نوردی (۱۹۴ تصویر) ۶- قایقرانی (۲۵۵ تصویر) ۷- موج‌سواری (۱۹۰ تصویر) ۸- اسکی (۱۹۰ تصویر) است. که تصاویر مطابق [۱۶] به عرض ۴۰۰ پیکسل تغییر اندازه داده شده است.

پایگاه داده LabelMe بطور آنلاین و با استفاده از جعبه ابزارهای موجود آن قابل دسترس است. در این مقاله از مجموعه یکسان با [۱۶] استفاده شده است. در این مجموعه داده ۸ کلاس: بزرگراه، درون‌شهری، ساحل، جنگل، ساختمان بلند، خیابان‌ها، برون‌شهری با تعداد ۲۰۰ تصویر به ازای هر کلاس موجود می‌باشد. در هر دو این مجموعه داده‌ها کلماتی که کمتر از ۳ بار در متن حاشیه‌نویسی ظاهر شده‌اند از پایگاه داده حذف شده است.

۴-۲- پارامترهای مدل و معیارهای ارزیابی

با توجه به [۷] از توصیف‌گر سیفت ۱۲۸ بعدی جهت استخراج کلمات بصری استفاده شده است. اندازه هر تکه‌ی[۲۳] تصویر که عملگر سیفت روی آن اعمال می‌شود برابر ۱۶×۱۶ پیکسل می‌باشد و اندازه قدم برای در نظر گرفتن تکه بعدی ۸ پیکسل است. از هر تصویر که در داده‌های آموزشی قرار دارد ۱۰۰ خروجی به تصادف انتخاب شده و به خوشه‌بند k-means داده می‌شود. در این مقاله ۲۴۰ مرکز خوشه به عنوان فرهنگ‌لغت کلمات بصری در نظر گرفته شده است. هر تصویر به شبکه‌های ۲×۲ تقسیم شده و تعداد ۲۴۰×۲×۲=۹۶۰ کلمه بصری مختلف برای فرهنگ لغات را تشکیل می‌دهند.

از دقت دسته‌بندی جهت ارزیابی دسته‌بند مدل و از میانگین معیار F- پنج کلمه محتمل برای ارزیابی حاشیه‌نویسی مدل استفاده شده است. معیار F- بصورت زیر تعریف می‌شود.

$$\text{F} - \text{measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (۱۵)$$

در این مقاله همانند [۱۶] از ماشین بردار پشتیبان جهت دسته‌بندی تصاویر استفاده شده است. لایه مخفی h_y که توسط کلمات بصری و کلمات وزن‌دار حاشیه‌نویسی آموزش داده شده به عنوان ورودی به دسته‌بند ماشین بردار پشتیبان داده می‌شود. پارامترهای این دسته‌بند توسط اعتبارسنجی متقابل محاسبه می‌شود.

۴-۳- نتایج آزمایشات

در این بخش به بررسی تاثیر پارامترهای مختلف روش پیشنهادی بر دقت آن پرداخته شده و راهکارهایی برای تنظیم این پارامترها ارائه می‌گردد. در ادامه نتایج حاصل از آزمایشات مختلف شرح داده شده است.

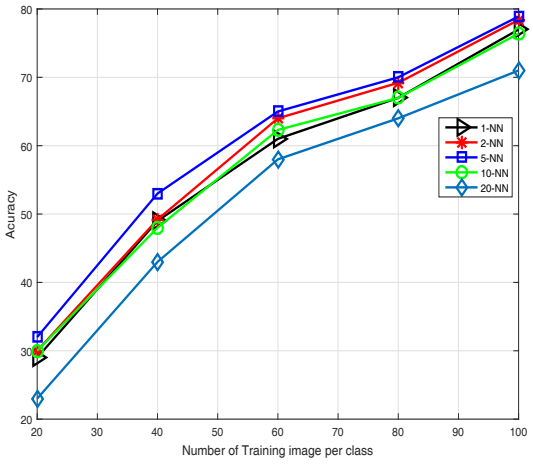
۴-۳-۱- تنظیم پارامترها

ابتدا جهت بدست آوردن مقدار k همسایه مناسب جهت بازنمایی هر تکه از تصویر در LLC تخمینی، دقت دسته‌بندی به ازای تعداد K مختلف مورد ارزیابی قرار گرفته است. مقادیر ۱، ۲، ۵، ۱۰ و ۲۰ همسایه مختلف برای K در نظر گرفته شده که در شکل ۵ و ۶ به ترتیب برای پایگاه داده‌ی UIUC_Sports و LabelMe دقت دسته‌بندی به ازای تعداد تصاویر در نظر گرفته شده به ازای هر کلاس جهت آموزش مورد ارزیابی قرار گرفته است. لازم به ذکر است زمانی که تعداد همسایه‌ها برابر یک باشد روش LLC جواب یکسانی با روش کیسه کلمات دارد.

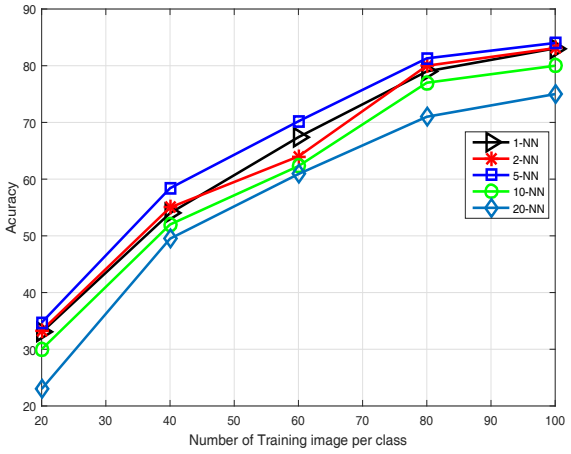
همانطور که مشاهده می شود افزایش تعداد همسایه از یک تا ۵ باعث افزایش دقت روش پیشنهادی می‌شود اما افزایش بیشتر آن باعث کاهش کارآیی شده است. از سویی افزایش تعداد نمونه‌های آموزشی تا حدود ۸۰ نمونه به ازاء هر کلاس باعث افزایش دقت شده است. از آنجا که افزایش تعداد همسایه‌ها باعث افزایش بار محاسباتی و کاهش سرعت اجرای برنامه می‌شود، مقادیر کمتر این پارامتر که دقت مناسبی را فراهم سازند توصیه می‌شود.

برای بدست آوردن وزن مناسب کلمات حاشیه‌نویسی، از روش اعتبارسنجی متقابل استفاده شده است. الگوریتم به ازای وزن‌های مختلف با در نظر گرفتن

روش LLC با ۵ و ۱ همسایه اجرا شده که در شکل ۷ و شکل ۸ نتایج معیار F برروی داده‌های تست نشان داده شده است. همانطور که در این نمودارها مشخص است برای پایگاه داده‌ی UIUC_Sports وزن ۵۰۰ و در حالتی که از کدگذار LLC استفاده می‌شود وزن ۲۰۰ مناسب است. برای پایگاه داده LabelMe وزن ۳۰۰ و در حالتی که از کدگذار LLC استفاده شده است وزن ۱۰۰ مناسب می‌باشد. با استفاده از وزن‌های بدست آمده در مرحله قبل یادگیری مدل برروی داده‌های آموزشی صورت می‌گیرد.



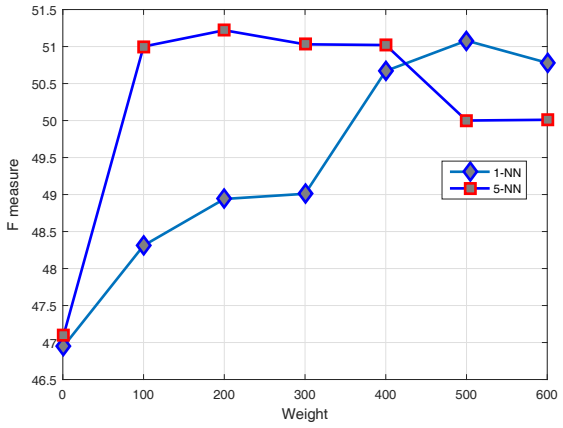
شکل ۵- نمودار دقت دسته‌بندی پایگاه داده UIUC_Sports به ازای تعداد همسایه‌های در نظر گرفته شده در روش LLC



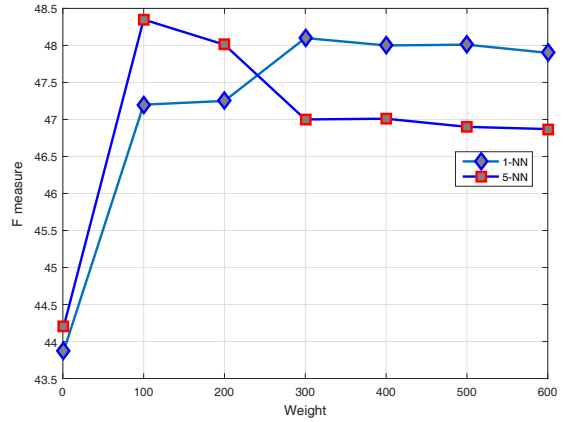
شکل ۶- نمودار دقت دسته‌بندی پایگاه داده LabelMe به ازای تعداد همسایه‌های در نظر گرفته شده در روش LLC

۴-۳-۲- ارزیابی نتایج

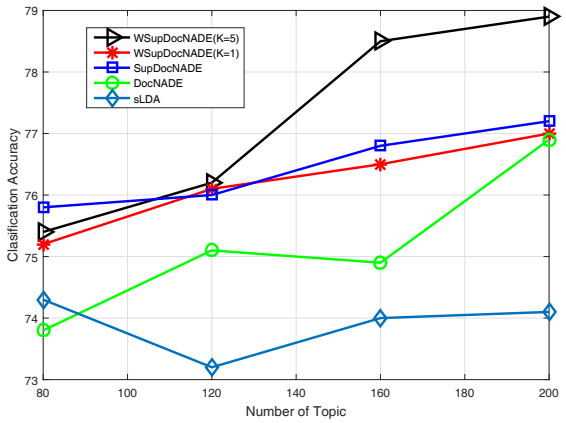
نتایج حاصل از دسته‌بندی برروی پایگاه داده‌های موردنظر در شکل ۹ و شکل ۱۰ نشان داده شده است. همانطور که مشاهده می‌شود با افزایش تعداد موضوعات مختلف برای روش‌های مختلف دسته‌بندی و حاشیه‌نویسی دقت دسته‌بندی به مراتب افزایش یافته است. زمانی که از روش کیسه کلمات بدون LLC استفاده شود دقت دسته‌بندی بهبودی نسبت به مدل SupDocNADE ندارد ولی زمانی که LLC با در نظر گرفتن ۵ همسایه جهت کدگذاری استفاده شده است شاهد بهبود حداقل ۱ درصدی در دقت دسته‌بندی نسبت به قبل هستیم.



شکل ۷- نمودار معیار F پایگاه داده UIUC_Sports به ازای وزن‌دهی‌های مختلف

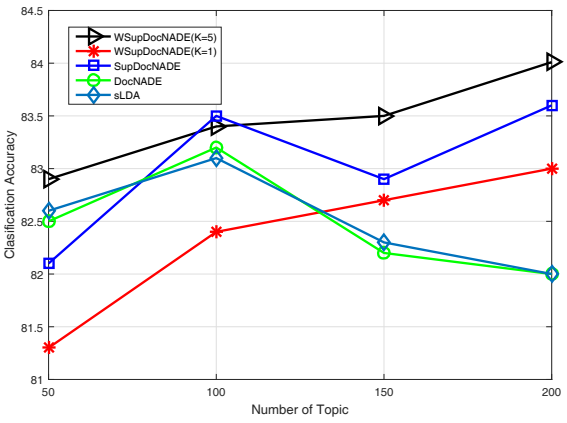


شکل ۸- نمودار معیار F برای پایگاه داده LabelMe به ازای وزن‌دهی‌های مختلف

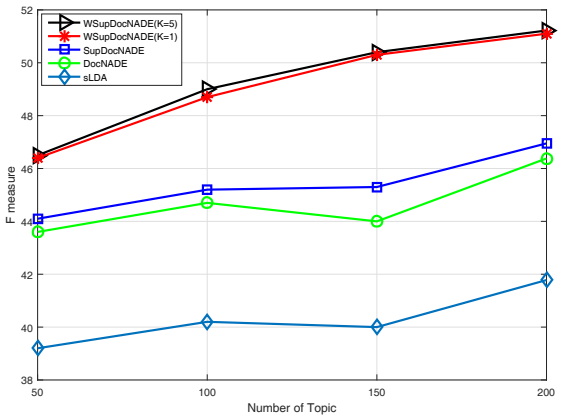


شکل ۹- نمودار دقت دسته‌بندی پایگاه داده UIUC_Sports به ازای تعداد واحدهای پنهان (تعداد عناوین)

شکل ۱۱ و شکل ۱۲ نتایج حاصل از حاشیه‌نویسی مدل را نشان داده است. همانطور که مشاهده می‌شود مدل پیشنهادی معیار F ۵۱.۲۲ درصد برای داده‌های UIUC_Sports و برای دادهای LabelMe مقدار ۴۸.۳۵ را کسب کرده است. با توجه به این نمودارها بهبود ۵ درصدی در معیار F کلمات حاشیه‌نویسی مشاهده می‌شود. جدول ۱ بهترین نتایج حاصل از دسته‌بندی و حاشیه‌نویسی مدل‌های مختلف را نشان می‌دهد.



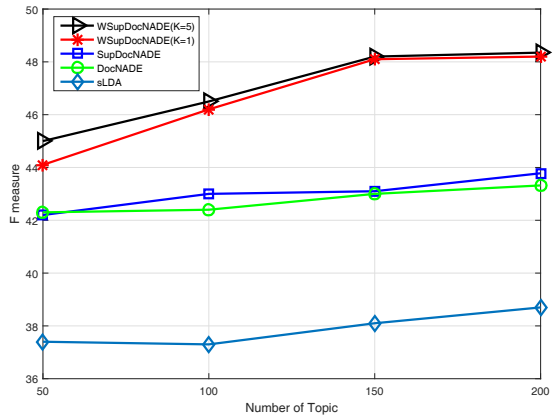
شکل ۱۰- نمودار دقت دسته‌بندی پایگاه داده LabelMe به ازای تعداد واحدهای پنهان (تعداد عناوین)



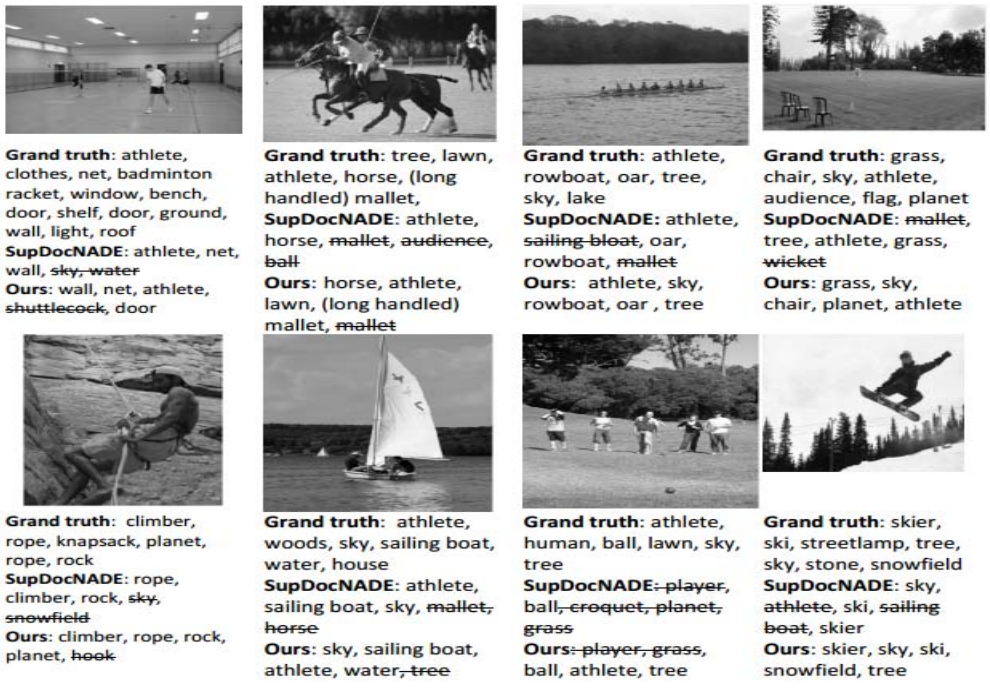
شکل ۱۱- نمودار معیار F برای پایگاه داده UIUC_Sports به ازای تعداد واحدهای پنهان (تعداد عناوین)

جدول ۱- مقایسه کارایی مدل پیشنهادی با روش‌های دیگر

مدل	UIUC-Sport		LabelMe	
	F-mesure	ACC	F-mesure	ACC
sLDA [۷]	٪ ۳۸/۰	٪ ۷۶/۸۷	٪ ۳۸/۷	٪ ۸۱/۸۷
DocNADE [۱۵]	٪ ۴۶/۳۸	٪ ۷۴/۲۳	٪ ۴۳/۳۲	٪ ۸۱/۹۷
SupDocNADE [۱۶]	٪ ۴۶/۹۵	٪ ۷۷/۲۹	٪ ۴۳/۸۷	٪ ۸۳/۶
روش پیشنهادی (k=1)	٪ ۵۱/۰۸	٪ ۷۷/۰	٪ ۴۸/۲	٪ ۸۳/۰۱
روش پیشنهادی (k=5)	**٪ ۵۲/۲۲**	**٪ ۷۸/۹**	**٪ ۴۸/۳۵**	**٪ ۸۴/۰۱**



شکل ۱۲- نمودار معیار F برای پایگاه داده LabelMe به ازای تعداد واحدهای پنهان (تعداد عناوین)



شکل ۱۳- برای هر تصویر کلمات حاشیه درست (Grand truth)، کلمات حاصل از روش SupDocNADE و کلمات حاصل از روش پیشنهادی این مقاله بیان شده است. کلمات خط خورده کلمات پیشنهادی اشتباه هستند

در شکل ۱۳ نمونه‌ای از حاشیه‌نویسی کلمات توسط مدل پیشنهادی و مدل [۱۶] نشان داده شده است. همانطور که مشاهده می‌شود کیفیت کلمات حاشیه‌نویسی در مدل پیشنهادی بهبود یافته است. بر خلاف روش SupDocNADE روش پیشنهادی از تکرار کلمات حاشیه‌ای نظیر mallet و athlete پرهیز کرده و در کلیه تصاویر بررسی شده تعداد کلمات حاشیه‌ای پیشنهاد شده‌ی اشتباه کمتری داشته است.

۵- نتیجه‌گیری

در این مقاله ابتدا مدل موضوعی SupDocNADE معرفی شد که نتایج خوبی در مدل‌کردن داده‌های چند مقداری مانند دسته‌بندی و حاشیه‌نویسی تصاویر ارائه داده است. در این مدل کلمات حاشیه‌نویسی در کنار کلمات بصری تعبیه شده و به عنوان بردار ویژگی برای شبکه در نظر گرفته می‌شود. در عمل تعداد ویژگی‌های استخراج شده از تصویر بسیار بزرگتر از ویژگی‌هایی است که از کلمات حاشیه‌نویسی بدست می‌آیند. عدم‌تعادل بین کلمات بصری و حاشیه‌نویسی سبب می‌شود تا سهم کلمات حاشیه‌نویسی برای بازنمایی در لایه پنهان شبکه عصبی مورد استفاده در این مدل، بسیار کمتر از کلمات بصری باشد. در این مقاله برای حل مشکل فوق، از وزن‌دهی ویژگی‌های ورودی شبکه استفاده شده است به این صورت که به کلمات حاشیه‌نویسی شده وزن بیشتری نسبت به کلمات بصری استخراج شده از تصویر داده شود تا فراوانی کم این کلمات نسبت به کلمات بصری جبران شود. از طرفی با افزودن قابلیت وزن‌دار کردن ورودی از روش کدگذاری LLC به جای روش سنتی کوانتیزاسیون برداری استفاده شد. نتایج نشان‌دهنده بهبود ۵ درصدی در معیار F مدل در حاشیه‌نویسی تصاویر و بهبود حداقل ۱ درصدی در دقت دسته‌بندی تصاویر است. در ادامه می‌توان از روش‌های پیشرفته‌تری به منظور استخراج ویژگی‌های تصویر استفاده کرد و همچنین می‌توان مفهوم توجه[۲۴] را در تولید کلمات حاشیه‌نویسی مد نظر قرار داد.

مراجع

[1] L.-J. Li, and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Computer Vision IEEE 11th International Conference on*, pp. 1–8, 2007.

[2] D. M. Blei, and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134, 2003.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] D. M. Blei, "Probabilistic topic models," in *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[5] L. Fei-Fei, and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition IEEE Computer Society Conference on*, vol. 2, pp. 524–531, 2005.

[6] J. D. Mcauliffe, and D. M. Blei, "Supervised topic models," in *Advances in neural information processing systems*, pp. 121–128, 2008.

[7] W. Chong, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition IEEE Conference on*, pp. 1903–1910, 2009.

[8] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Computer Vision and Pattern Recognition IEEE Conference on*, pp. 2036–2043, 2009.

[9] X. LI, C. SUN, L. U. Peng, X. WANG, and Y. ZHONG, "Simultaneous image classification and annotation based on probabilistic model," in *Journal of China Universities of Posts and Telecommunications,* vol. 19, no. 2, pp. 107–115, 2012.

[10] Y. Wang, and G. Mori, "Max-margin Latent Dirichlet Allocation for Image Classification and Annotation," in *BMVC*, vol. 2, no. 6, pp. 7, 2011.

[11] G. E. Hinton, and R. R. Salakhutdinov, "Replicated softmax: an undirected topic model," in *Advances in neural information processing systems*, pp. 1607–1614, 2009.

[12] R. Salakhutdinov, and I. Murray, "On the quantitative analysis of deep belief networks," in *Proceedings of the 25th international conference on Machine learning*, pp. 872–879, 2008.

[13] R. M. Neal, "Annealed importance sampling," *Statistics and computing*, vol. 11, no. 2, pp. 125–139, 2001.

[14] H. Larochelle, and I. Murray, "The neural autoregressive distribution estimator," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 29–37, 2011.

[15] H. Larochelle, and S. Lauly, "A neural autoregressive topic model," in *Advances in Neural Information Processing Systems*, pp. 2708–2716, 2012.

[16] Y. Zheng, Y.-J. Zhang, and H. Larochelle, "Topic modeling of multimodal data: an autoregressive approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1370–1377, 2014.

[17] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1–22, pp. 1–2, 2004.

[18] J. Sivic, and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Computer Vision Ninth IEEE International Conference on*, pp. 1470, 2003.

[19] K. Grauman, and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image

features," in *Computer Vision Tenth IEEE International Conference on*, vol. 2, pp. 1458–1465, 2005.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision and pattern recognition IEEE computer society conference on*, vol. 2, pp. 2169–2178, 2006.

[21] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition IEEE Conference on*, pp. 3360–3367, 2010.

[22] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition IEEE Conference on*, pp. 1794–1801, 2009.

[23] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1, pp. 157–173, 2008.

سید نوید محمدی فومنی مدرک کارشناسی و کارشناسی‌ارشد خود را به ترتیب از دانشگاه زنجان در سال ۹۳ و دانشگاه صنعتی امیرکبیر در سال ۹۵ دریافت نموده است. پایان‌نامه ایشان در مورد استفاده از مدل‌های عنوان جهت دسته‌بندی و حاشیه‌نویسی تصاویر می‌باشد که از رویکرد شبکه‌های عصبی کانولوشنی جهت استخراج ویژگی استفاده شده است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

navid_foumani@aut.ac.ir

احمد نیک‌آبادی عضو هیئت علمی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیرکبیر است. وی مدرک کارشناسی‌ارشد و دکترای خود را در زمینه هوش مصنوعی از دانشگاه صنعتی امیرکبیر دریافت کرده است. زمینه‌های تحقیقاتی مورد علاقه ایشان مدل‌های احتمالاتی گرافی، پردازش تصویر و هوش محاسباتی است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

nickabadi@aut.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۰/۲۸

تاریخ اصلاح: ۱۳۹۵/۱۱/۱۵

تاریخ قبول شدن: ۱۳۹۵/۱۱/۳۰

نویسنده مرتبط: دکتر احمد نیک‌آبادی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران.

[1]Annotation
[2]Probabilistic Topic Models
[3]Latent Dirichlet Allocation
[4]Multimodal
[5]Visual Word
[6]Multi-Class Supervised Latent Dirichlet Allocation
[7]Exact Inference
[8]Intractable
[9]Posterior
[10]Replicated Softmax
[11]Partition Function
[12]Annealed Importance Sampling
[13]Neural Autoregressive Distribution Estimator
[14]Restricted Boltzmann Machine
[15]Autoencoder
[16]Bag-of-Feature
[17]Spatial Pyramid Matching
[18]Quantization
[19]Radial Basis Function
[20]Vector Quantization
[21]Locality
[22]Cross Validation
[23]Patch
[24]Attention