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

عنوان:
دسته‌بندی ریزدانه‌ای تصاویر
Fine-grained Image Classification

نگارش:
یاسر سوری
۹۲۲۰۴۷۴۴

استاد راهنما:
دکتر شهره کسایی

استاد ممتحن داخلی:
دکتر محمد تقی منظوری شلمانی

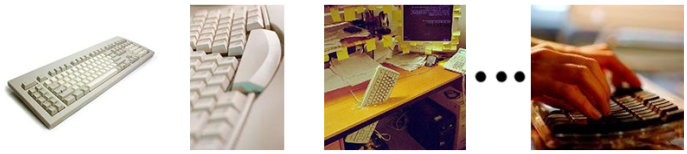
شهریور ۹۳

چکیده: دسته‌بندی تصاویر ریزدانه‌ای عبارت است از دسته‌بندی تصاویر در حالتی که دسته‌های مورد نظر همگی زیر دسته‌ی یک دسته‌ی کلی‌تر هستند. برای مثال برای زیر دسته‌ی کلی پرندگان ما می‌توانیم گونه‌های مختلف پرندگان را در نظر بگیریم. در این حالت خاص مسئله دسته‌ها معمولاً از نظر ظاهری بسیار به یکدیگر شبیه هستند به گونه‌ای که افراد غیر متخصص نمی‌توانند دسته‌ها را از یکدیگر تمایز دهند. در چنین شرایطی راه حل‌های ارائه شده برای مسئله دسته‌بندی تصاویر معمولی اکثراً نتایج خوبی کسب نمی‌کنند. لذا ارائه روش‌هایی جدید برای حل این مسئله لازم است. در این گزارش ابتدا به مرور روش‌های مهم در دسته‌بندی تصاویر معمولی و سپس به مرور روش‌های ارائه شده برای دسته‌بندی تصاویر ریزدانه‌ای می‌پردازیم. سپس روش انتخاب شده و دلایل انتخاب آن را بررسی می‌کنیم.

واژه‌های کلیدی: بینایی کامپیوتری، بازشناسی شیء، دسته‌بندی تصاویر، بازشناسی ریزدانه‌ای، دسته‌بندی تصاویر ریزدانه‌ای.

۱ مقدمه

در دسته‌بندی تصویر [۱] هر تصویر با توجه به محتوایش دسته‌بندی می‌شود. برای مثال آیا تصویر شامل شیء خودرو هست یا خیر. معمولاً در بینایی کامپیوتری مسئله بدین صورت است که تعدادی دسته مشخص را در نظر می‌گیریم (مثلاً انسان، خودرو، ساختمان، تلویزیون، صندلی، اسب و ...) سپس طبق چارچوب معمول یادگیری ماشین، توسط تعدادی تصویر شامل یکی از دسته‌ها (نمونه‌های مثبت) و تعدادی تصویر بدون شیء‌ای از آن دسته (نمونه‌های منفی) یادگیری برای آن دسته انجام می‌شود. در نهایت پس از یادگیری تمام دسته‌ها در مواجهه با تصویر جدید لازم است تشخیص دهیم که آیا شیء‌ای از هر کدام از آن دسته‌های مورد بررسی در تصویر وجود دارد یا خیر. برای نمونه به شکل ۱,۱ توجه کنید. در این شکل داده‌های آموزشی و آزمایشی برای دسته‌بند دسته‌ی «صفحه کلید [۲] » از پایگاه داده Caltech256 [۱] نشان داده شده است.



(آ) *نمونه‌های مثبت، شامل شیء صفحه کلید*

(ب) *نمونه‌های منفی، بدون شیء صفحه کلید*

(ج) *تصویر جدید آزمایش*

شکل ۱,۱: *نمونه‌ای از تصاویر آموزشی و آزمایشی سامانه دسته‌بند تصاویر برای دسته‌ی «صفحه کلید». انتخاب شده از پایگاه داده Caltech256 [۱]. سامانه لازم است با مشاهده نمونه‌های مثبت ۱,۱آ و نمونه‌های منفی ۱,۱ب یادگیری را انجام داده و بتواند پاسخ دهد که در تصویر جدید آزمایشی ۱,۱ج آیا صفحه کلید وجود دارد یا خیر. برای این نمونه پاسخ مثبت است.*

(آ) *هواپیمای جنگنده*

(ب) *هواپیمای مسافربری*

شکل ۲,۱: *نمونه‌ای از دو دسته‌ی متفاوت ولی شبیه به هم از پایگاه داده Imagenet [۲]. مدل دسته‌بند علاوه بر توانایی مدل‌سازی تفاوت‌های داخل دسته‌ای، باید توانایی تمایز بین دسته‌های گاها شبیه به یکدیگر را داشته باشد.*

چالش‌های اصلی این مسئله تنوع زیاد اشیاء درون هر کدام از دسته‌ها، نحوه‌ی عکس برداری و وجود اشیاء دیگر در تصویر است که باعث ایجاد تصاویری با تنوع بالا می‌شود. مدل کردن این تنوع مربوط به اشیاء هر دسته باید همزمان با توانایی تمایز بین دسته‌های مختلف باشد. برای نمونه شکل ۲,۱ تصویری از دو دسته مختلف مربوط به پایگاه داده Imagenet [۲] را نمایش می‌دهد. مدل دسته‌بند باید توانایی تمایز بین این دو دسته شبیه به هم را داشته باشد.

اگر در دسته‌بندی تصویر، دسته‌های مورد بررسی زیر دسته‌ی [۳] یک دسته‌ی کلی‌تر باشند(مانند گونه‌های مختلف پرندگان، مدل‌های مختلف خودروهای سواری و انواع مختلف هواپیماها)، آنگاه مسئله را «دسته‌بندی ریزدانه‌ای تصویر [۴]» می‌نامند. در دسته‌بندی ریزدانه‌ای تصویر معمولاً شباهت دسته‌ها به یکدیگر بسیار زیاد است به نحوی که افراد غیر متخصص نمی‌توانند به راحتی این دسته‌ها را بازشناسی نمایند. برای نمونه در شکل ۳,۱ چند گونه‌ی مختلف از پرستوی دریایی [۵] متعلق به پایگاه داده CUB-200-2011 [۳] نمایش داده شده است. همانگونه که می‌بینید با اینکه این نمونه‌ها به نحوی انتخاب شده‌اند که وضعیت مشابهی دارند، هنوز هم پیدا کردن ویژگی‌های تمایز دهنده بین گونه‌های مختلف کار بسیار سختی است و نیاز به تخصص دارد.

(آ) *The Caspian tern*

(ب) *The Artic tern*

(ج) *The common tern*

(د) *The elegent tern*

شکل ۳,۱: *چهارگونه‌ی مختلف از پرستوهای دریایی متعلق به پایگاه داده CUB-200-2011 [۳]. شباهت بسیار زیاد بین دسته‌های مختلف کار را حتی برای افراد غیر متخصص بسیار سخت می‌کند.*

روش‌های دسته‌بندی تصویر معمولی در مسایل دسته‌بندی ریزدانه‌ای اکثراً موفق نیستند (به بخش فولان مراجعه شود). دلیل اصلی این عدم موفقیت وجود ویژگی‌های بسیار اندک و شدیداً محلی تمایزدهنده [۶] برای دسته‌های ریزدانه ایست. برای مثال دو گونه‌ی elegent tern در شکل ۳,۱د و common tern در شکل ۳,۱ج فقط در رنگ پا و شکل تاج با یکدیگر تفاوت دارند و در سایر اجزا غیر قابل تمایز هستند.

مسئله دسته‌بندی تصویر را می‌توان به دلیل کاربردهای زیاد آن یکی از اساسی‌ترین مسائل بینایی کامپیوتری دانست که امروزه مورد علاقه محققین در سطح جهان است. نتایج بهترین روش‌های دسته‌بندی تصاویر بر روی بزرگ‌ترین پایگاه داده‌های دسته‌بندی تصویر نشان داده است که اکثر خطای این روش‌ها مربوط به دسته‌هایی است که از نظر ظاهری به یکدیگر بسیار نزدیک هستند. به عبارت دیگر خطای این روش‌ها اکثراً خطای ریزدانه ایست (به بخش فولان مراجعه شود). لذا برای تقویت روش‌های دسته‌بندی تصویر، تمرکز بر روی دسته‌بندی ریزدانه‌ای اهمیت زیادی دارد. از جمله کاربردهای مسئله دسته‌بندی ریزدانه‌ای تصویر می‌توان به موتورهای جستجو و بازیابی محتوا محور تصاویر [۷] (برای مثال پیدا کردن مدل خاصی از یک خودرو در

شکل ۴,۲: *نمونه خروجی سامانه‌های دسته‌بندی شیء و صحنه. برای دسته‌بندی شیء اگر شیءهای هواپیما، اتوبوس، خودرو سواری و اسب را مد نظر داشته باشیم، خروجی سامانه باید بدین صورت باشد: «هواپیما در تصویر وجود دارد. اوتوبوس در تصویر وجود دارد. خودرو سواری و اسب در تصویر وجود ندارد». برای دسته‌بندی صحنه اگر صحنه‌های فرودگاه، فروشگاه و مسجد را در نظر بگیریم، خروجی سامانه باید بدین صورت باشد: «تصویر صحنه فرودگاه را نشان می‌دهد».*

بین انبوهی از تصاویر)، سامانه‌های کمک آموزشی (آموزش گونه‌های مختلف حیوانات) و سامانه‌های نظارتی [۸] (کنترل ترافیک و تشخیص مدل‌های مختلف خودرو) اشاره کرد.

در ادامه ابتدا در بخش ۲ به معرفی سامانه‌های دسته‌بند تصویر معمولی و کارهای پیشین در این زمینه خواهیم پرداخت. سپس در بخش ۳ به معرفی روش‌های دسته‌بندی تصاویر ریزدانه‌ای و پایگاه داده‌های مرتبط با آن خواهیم پرداخت. در بخش فولان چند آزمایش و ایده برای ادامه کار را مطرح خواهیم کرد و نتایج ابتدایی را گزارش می‌دهیم. در انتها در بخش فولان نتیجه‌گیری‌ها و جمع‌بندی را مطرح می‌کنیم.

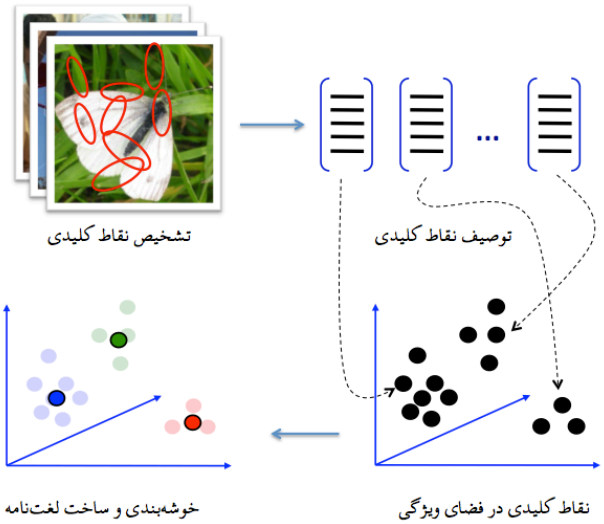
۲ دسته‌بندی تصاویر معمولی

در این گزارش دسته‌بندی غیر ریزدانه‌ای تصاویر را معمولی خطاب می‌کنیم. همانطور که در مقدمه (بخش ۱) بیان شد، در دسته‌بندی تصویر، باید تصویر را با توجه به محتوایش دسته‌بندی کنیم. برای این منظور دسته‌ها را می‌توان به چندین صورت تعریف کرد. دسته‌های به صورت سنتی به دو دسته تقسیم می‌شدند: دسته‌بندی شیء [۹] (برای مثال Caltech101/256 [۴، ۱] و PASCAL VOC [۵]) و دسته‌بندی صحنه [۱۰] (برای مثال MIT 67 Scene [۶]). در دسته‌بندی شیء لازم است که وجود و یا عدم وجود شیای از دسته‌های مورد نظر در تصویر را تشخیص دهیم و لازم نیست که مکان آن شیء را در تصویر مشخص نماییم. در دسته‌بندی صحنه لازم است از بین صحنه‌های مورد بررسی تشخیص دهیم که تصویر متعلق به کدام صحنه است. برای مثال تصویر ۴,۲ را در نظر بگیرید. فرض کنید که شیءهای هواپیما، اتوبوس، خودرو سواری و اسب را در مد نظر داریم و می‌خواهیم دسته‌بندی شیء را برای تصویر ۴,۲ انجام دهیم. در این حالت خروجی سامانه باید بدین صورت باشد: «هواپیما در تصویر وجود دارد. اوتوبوس در تصویر وجود دارد. خودرو سواری و اسب در تصویر وجود ندارد». حالا اگر صحنه‌های فرودگاه، فروشگاه و مسجد را در نظر بگیریم و بخواهیم دسته‌بندی صحنه را برای تصویر انجام دهیم، خروجی سامانه باید بدین صورت باشد: «تصویر صحنه فرودگاه را نشان می‌دهد».

در سال‌های اخیر محققین به دسته‌بندی‌های دیگری نیز روی آورده‌اند. برای نمونه می‌توان به بازشناسی صفت [۱۱][۷] و بازشناسی افعال [۱۲] در تصاویر [۸] اشاره کرد. در این گزارش تمرکز بر دسته‌بندی شیء خواهد بود. در ادامه برخی روش‌های دسته‌بندی شیء در تصاویر را بررسی خواهیم کرد.

۱,۲ روش‌های مبتنی بر لغت‌نامه

روش‌های مبتنی بر لغت‌نامه از قدیمی‌ترین روش‌ها در دسته‌بندی تصویر هستند که هنوز هم مورد استفاده قرار می‌گیرند. ایده اولیه این روش‌ها از تحقیقات پردازش متن گرفته شده است. در پردازش متن، یک لغت‌نامه داریم و یک متن و می‌خواهیم متن را دسته‌بندی کنیم. اگر فرکانس لغات داخل متن را محاسبه کنیم، می‌بینیم که لغات پر تکرار در متون ورزشی مشترک هستند. این مشاهده باعث شد که روش‌های مبتنی بر لغت‌نامه برای دسته‌بندی متن مورد استفاده قرار بگیرند.



شکل ۵,۲: *روش ساخت لغت‌نامه در روش کیسه‌ای از لغات*

محققان بررسی کردند که همین مشاهدات در مورد تصاویر نیز برقرار است. برای مثال فرض کنیم که لغات را در تصاویر به صورت تکه‌ای از تصویر تعریف کنیم. حال اگر لغت معادل دهان، چشم و بینی را در تصویری مشاهده کنیم می‌توانیم نتیجه بگیریم که به احتمال زیاد تصویر مربوط به یک چهره است. البته مسئله برای تصویر از متن پیچیده‌تر است، زیرا لغت‌نامه‌ای در دسترس نیست و همچنین پیدا کردن لغت‌ها کار مشکلی است.

کلیت این روش‌ها بر پایه‌ی دسته‌بندی در یادگیری ماشین است. بدین صورت که تعدادی تصویر آموزشی و دسته‌های آن‌ها (برچسب) را در اختیار داریم. ابتدا یک مدل دسته‌بند مثل SVM [۹] را انتخاب می‌کنیم و با داده‌های آموزشی مدل را آموزش می‌دهیم. حال برای داده‌های آزمایشی با استفاده از مدل برچسب را تخمین می‌زنیم. برای استفاده از این چارچوب دسته‌بندی باید نمایش برداری مناسبی از تصاویر داشته باشیم که در آموزش و آزمایش طول یکسانی داشته باشد. در ادامه به راه حل‌های مختلف برای ساخت این نمایش برداری می‌پردازیم.

۱.۱,۲ روش کیسه‌ای از لغات

روش «کیسه‌ای از لغات» [۱۳] از قدیمی‌ترین روش‌های پیدا کردن نمایش برداری برای مسئله دسته‌بندی است [۱۰]. در این روش کیسه‌ای از لغات به هیستوگرام تعداد تکرارهای الگوهای خاصی در تصویر گفته می‌شود. این الگوهای خاص همان لغات لغت‌نامه هستند. در شکل ۵,۲ روش ساخت لغت‌نامه با پیدا کردن و توصیف نقاط کلیدی نشان داده شده است. گام‌های این روش برای ساخت لغت‌نامه بدین صورت است:

- در مجموعه تصاویر آموزشی نقاط کلیدی [۱۴] را توسط الگوریتمی مثل Harris-Affine [۱۱] یا به صورت چگال [۱۵] پیدا می‌کنیم.
- نقاط کلیدی را توسط توصیفگری مثل SIFT [۱۲] توصیف می‌کنیم.
- با اعمال یکی از روش‌های تدریج برداری [۱۶] مثل KMeans توصیفگر هر قطعه از تصاویر را به یکی از مراکز خوشه‌های [۱۷] لغت‌نامه اختصاص می‌دهیم.

بدین ترتیب روش مناسب برای ساخت لغت‌نامه و پیدا کردن لغات درون تصاویر داریم.

۲٫۱٫۲ روش Spatial Pyramid Matching

۳٫۱٫۲ روش Sparse Coding

۴٫۱٫۲ روش Fisher Kernel

۲٫۲ روش‌های مبتنی بر یادگیری عمیق

۳٫۲ نتیجه‌گیری

۳ دسته‌بندی ریزدانه‌ای تصویر

References

[1] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep., 2007.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[3] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[4] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2004.

[5] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, pp. 1–39, 2014.

[6] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *CVPR*, 2009.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[8] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *ICCV*, 2011.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[10] 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*, 2004.

[11] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *ECCV*, 2002.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

واژه‌نامه

Image classification[۱]	Discriminative[۶]	Action recognition[۱۲]
Computer - Keyboard[۲]	Content based image retrieval[۷]	Bag of keypoints[۱۳]
Subclass[۳]	Surveillance systems[۸]	Keypoints[۱۴]
Fine-grained image classification[۴]	Object classification[۹]	Dense[۱۵]
Tern[۵]	Scene classification[۱۰]	Vector quantization[۱۶]
	Attribute recognition[۱۱]	Clusters[۱۷]