**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 دسته‌ی 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 شی‌ای 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 شی‌ای 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 آزمایشی*

**شکل ۱,۱:** *نمونه‌ای از تصاویر آموزشی و آزمایشی سامانه دسته‌بند تصاویر برای دسته‌ی «صفحه کلید». انتخاب شده از پایگاه داده‌ی 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 عادی (غیر متخصص) بسیار سخت می‌کند.*

روش‌های دسته‌بندی تصویر معمولی در مسایل دسته‌بندی ریزدانه‌ای اکثراً موفق نیستند. 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 ۳,۲ 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 ریزدانه‌ای تصاویر و پایگاه داده‌های مرتبط با آن خواهیم پرداخت. در انتها در بخش ۴ چند آزمایش و ایده برای ادامه کار به همراه 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 شیء‌ای 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 یک دسته‌ی خاص مشترک هستند. این مشاهده باعث شد که روش‌های مبتنی بر لغت‌نامه برای دسته‌بندی متن مورد استفاده قرار بگیرند.

**شکل ۵,۲:** *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. این نمایش برداری توسط مرحله‌ی کدینگ [۱۳] همانطور که در شکل ۶,۲ نشان داده شده است، بدست می‌آید. در ادامه به راه حل‌های مختلف برای ساخت لغت‌نامه و کدینگ می‌پردازیم.

### ۱.۱,۲ ساخت لغت‌نامه

در روش‌های مبتنی بر لغت‌نامه، ساخت لغت‌نامه یکی از اساسی‌ترین گام‌هاست و از آنجایی که فقط کافی است یک بار ساخته شود و در طول اجرای برنامه نیاز به بازسازی ندارد اگر زمان زیادی هم صرف این کار شود ایرادی ندارد. در شکل ۵,۲ یکی از ابتدایی‌ترین و پر کاربردترین روش‌های ساخت لغت‌نامه را مشاهده می‌کنید [۱۰]. 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] تشخیص می‌دهیم.
- 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.

مراکز خوشه‌ها همان لغت‌نامه خواهند بود. این روش یکی قدیمی‌ترین روش‌های ساخت لغت‌نامه است. بعدها روش‌های پیچیده‌تری مثل [۱۳، ۱۴] برای این کار معرفی شدند.

(آ) روش تطبیق هرم مکانی

(ب) روش کیسه‌ای از لغات (ج) روش‌های برپایه‌ی تطبیق هرم مکانی

**شکل ۶,۲:** *نمای کلی روش‌های برپایه لغت‌نامه*

### ۲.۱,۲ روش کیسه‌ای از لغات

روش «کیسه‌ای از لغات» [۱۸] از قدیمی‌ترین روش‌ها برای مسئله دسته‌بندی است [۱۰]. در این روش «کیسه‌ای از لغات» به هیستوگرام تعداد تکرارهای لغات درون لغت‌نامه در تصویر گفته می‌شود. لغت‌نامه همانطور که در قسمت ۱.۱,۲ بیان شد ساخته می‌شود. شمای کلی این روش را می‌توان در تصویر ۶,۲ب مشاهده کرد.

ابتدا در گام استخراج ویژگی، از تصویر ورودی نقاط کلیدی را استخراج و توصیف می‌کنیم، این توصیف‌ها توسط دایره‌های آبی رنگ نشان داده شده‌اند. سپس در گام کدینگ این توصیف‌ها را به نزدیک‌ترین لغت درون لغت‌نامه نگاشت می‌کنیم که حاصل کدها (دایره‌های قرمز) خواهند بود. نهایتاً در مرحله‌ی ادغام [۱۹] تمام کدها را توسط میانگین گیری به نمایش نهایی تبدیل می‌کنیم. این ادغام را «ادغام میانگین [۲۰]» می‌نامند.

به صورت ریاضی این روش را می‌توان بدین صورت مطرح کرد. فرض کنید که $X$ مجموعه‌ی توصیف‌های نقاط کلیدی استخراج شده از تصویر باشد، یعنی: $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{D\times N}$. با فرض داشتن لغت‌نامه‌ای با $M$ لغت، $B = [b_1, b_2, ..., b_M] \in \mathbb{R}^{D\times M}$ حال باید $x_i$ها را به برداری $M$ بعدی تبدیل کنیم. سپس از این بردارهای $M$ بعدی می‌توان نمایش نهایی را ساخت. در روش کیسه‌ای از لغات از تدریج برداری استفاده

می‌شود که مسئله بهینه‌سازی زیر را حل می‌کند.

$$\begin{aligned} &\arg\min_{C} \sum_{i=1}^{N} ||x_i - Bc_i||^2 \\ &s.t. ||c_i||_{\ell^0} = 1, ||c_i||_{\ell^1} = 1, c_i \succeq 0, \forall i \end{aligned} \tag{1}$$

با فرض اینکه کدینگ $C = [c_1, c_2, ..., c_N] \in \mathbb{R}^{M \times N}$ مجموعه کدهای مربوط به $X$ هستند. شرط $||c_i||_{\ell^0} = 1$ (نرم صفر $c_i$ها) در بهینه‌سازی ۱ بدین معناست که فقط یک عنصر غیر صفر در $c_i$ وجود داشته باشد. همچنین شروط $||c_i||_{\ell^1} = 1, c_i \succeq 0$ بدین معناست که آن عنصر غیر صفر باید برابر با مثبت یک باشد. در عمل این مسئله بهینه‌سازی همانطور که گفته شد با جستجوی نزدیک‌ترین همسایه در فضای توصیف حل می‌شود.

این روش ساده در پایگاه داده‌های ابتدایی نتایج قابل قبولی داشت. مهم‌ترین نکته قوت این روش مقاوم بودن در مقابل تبدیل‌های انتقالی است. بدین معنا که برای این روش فرقی نمی‌کند که شیء در کدام قسمت تصویر وجود داشته باشد. از طرفی همین موضوع می‌توان باعث نقطه ضعف این روش شود، زیرا این نکته باعث می‌شود موقعیت مکانی ویژگی‌ها را حفظ نکند، که در نهایت باعث می‌شود توانایی مدل کردن شکل را نداشته باشد.

از بین توسعه‌هایی که به روش کیسه‌ای از لغت داده شد [۱۴، ۱۵، ۱۶] روش «تطبیق هرم مکانی [۲۱]» [۱۷] از دیگران نتایج بهتری داشت. در ادامه به بررسی این روش می‌پردازیم.

### ۳.۱,۲ روش تطبیق هرم مکانی

این روش [۱۷] سعی داشت که مشکل از دست دادن اطلاعات مکانی را در روش کیسه‌ای از لغات حل کند. کلیت این روش در شکل ۶,۲آ نشان داده شده است. در این روش تصویر به ناحیه‌های هرچه ریزتر تقسیم می‌شود و هیستوگرام کدها در هر کدام از زیر ناحیه‌ها محاسبه می‌شوند. و در انتها این هیستوگرام‌ها به عنوان نمایش نهایی کنار هم قرار داده می‌شود.

در این روش استخراج ویژگی و کدینگ دقیقاً مثل روش کیسه‌ای از لغات است. تنها تفاوت در گام ادغام رخ می‌دهد. (به شکل ۶,۲ب و ۶,۲ج توجه نمایید) در این گام ابتدا ناحیه‌بندی صورت می‌گیرید. زیر ناحیه‌های معمولاً به صورت $2^l \times 2^l, l = 0, 1, 2$ هستند. در شکل ۶,۲آ زیر ناحیه‌های $2 \times 2$ و $1 \times 1$ نشان داده شده است. سپس برای هر کدام از ناحیه‌های ادغام میانگین انجام می‌گیرد و هیستوگرام‌هایی تولید می‌شود. در نهایت گام چسباندن این هیستوگرام‌ها را کنار هم قرار داده و به نمایش نهایی تبدیل می‌کند.

یکی از نقطه ضعف‌های این روش این است که برای کسب نتایج خوب باید از دسته‌بندهای غیر خطی مثل SVM با هسته‌های غیر خطی مثل Chi-square استفاده شود. این موضوع باعث می‌شود که هزینه‌ی محاسباتی در گام آموزش $\mathcal{O}(n^3)$ و در هنگام آزمایش $\mathcal{O}(n)$ باشد که $n$ تعداد بردارهای پشتیبان [۲۲] است. دیگر نقطه ضعف مهم این روش که آن را از روش‌های کیسه‌ای از لغات به ارث برده است، خطای تدریج در مرحله‌ی کدینگ است. این خطا همانطور که در [۱۵] به تفصیل به آن پرداخته شده است، مهم‌ترین نقطه ضعف این روش است. برای بهبود مقیاس‌پذیری و از بین بردن خطای تدریج این روش محققان بر روی روش‌های کدینگ غیر خطی کار کردند که با دسته‌بندهای خطی نیز خوب کار کنند. در ادامه به برخی از این روش‌ها خواهیم پرداخت.

### ۴.۱,۲ روش‌های کدینگ تنک

برای از بین بردن خطای تدریج، شرط قوی $||c_i||_{\ell^0} = 1$ در معادله‌ی ۱ را می‌توان با اضافه کردن جمله‌ی منظم‌سازی [۲۳]، نرم کرد. در روش «کدینگ تنک در تطبیق هرم مکانی» [۱۸] مرحله‌ی کدینگ با همین ایده به صورت زیر در آورده شده است:

$$\arg\min_{C} \sum_{i=1}^{N} ||x_i - Bc_i||^2 + \lambda ||c_i||_{\ell^1} \tag{2}$$

**شکل ۷,۲:** *ماتریس ابهام برای پایگاه داده‌ی ImageNet و ساختار درون آن*

جمله‌ی مرتب‌سازی با نرم ۱ باعث تنک شدن $c_i$ها خواهد شد [۱۹]. بدین ترتیب می‌توان خطای تدریج را به میزان قابل ملاحظه‌ای کاهش داد. آزمایشات نشان داد که این روش کدینگ با دسته‌بندخطی از روش تطبیق هرم مکانی بسیار بهتر عمل می‌کند [۱۸]. لازم به ذکر است که ساختار کلی این روش‌ها مانند شکل ۶,۲ج است و فقط در مرحله‌ی کدینگ با یکدیگر تفاوت دارند.

همچنین روش‌های دیگری مثل [۲۰، ۲۱، ۲۲] برای کدینگ مطرح شدند که نتایج دسته‌بندی را بهبود می‌دادند، که بدلیل کمبود فضا از توضیح آن‌ها خودداری می‌کنیم. در [۲۳] مقایسه‌ی کاملی بین روش‌های کدینگ انجام شده است، در نهایت می‌توان گفت بهترین نتیجه برای دسته‌بندی توسط کدینگ هسته‌ی فیشر [۲۲] بدست می‌آید.

## ۲,۲ روش‌های مبتنی بر یادگیری عمیق

روش‌های مبتی برا یادگیری عمیق [۲۴] از سال‌های قبل در بینایی کامپیوتری استفاده می‌شدند [۲۴]. اما تا سال ۲۰۱۲ موفقیت چندانی در دسته‌بندی تصاویر واقعی نداشتند. در سال ۲۰۱۲ در مقاله‌ی مهمی از دانشگاه تورنتو [۲۵] شبکه‌ی عصبی عمیقی معرفی شد که توجه همه را به خود جمع کرد. این شبکه‌ی عصبی موسوم به AlexNet در مسابقات دسته‌بندی تصویر معتبر ILSVRC[۲۶] در سال ۲۰۱۲ توانست با اختلاف قابل ملاحظه‌ی از نفر دوم، اول شود. از این به بعد بود که محققان و صنعتگران از تمام جهان متوجه روش‌های یادگیری عمیق و قابلیت‌های آن در عرصه‌های مختلف شدند.

کمی قبل‌تر از معرفی AlexNet محققان متوجه رشد کم در بهبود روش‌های بازشناسی تصویر شده بودند و در پی این بودند که بدانند چرا انسان‌ها قدرت بسیار بالایی در بازشناسی تصویر دارند ولی هوش مصنوعی هنوز در این زمینه ضعف زیادی دارد. تعدادی از محققین با انجام آزمایشات گسترده بر روی انسان و روش‌های یادگیری ماشین متوجه شدند [۲۷] که نقطه ضعف روش‌های بینایی ماشین در ویژگی‌هایی (ویژگی‌هایی مثل SIFT [۱۲] و HOG [۲۸]) است که استفاده می‌کند. به عبارت دیگر توانایی بالای انسان نه به خاطر آموزش با داده‌های آموزشی بیشتر و نه به خاطر مدل دسته‌بندی پیشرفته‌تر، بلکه بیشتر به خاطر ویژگی‌های بهتری است که استفاده می‌کند.

یکی از محرک‌های اصلی روش‌های یادگیری عمیق همین مسئله بود، به عبارت دیگر این روش‌ها نه تنها دسته‌بند، بلکه خود ویژگی‌ها را نیز با داده‌های آموزشی یاد می‌گیرند.

## ۳,۲ نتیجه‌گیری

در این بخش مرور مختصری بر روش‌های پایه‌ای و مهم در دسته‌بندی تصویر داشتیم. در بین این روش‌ها آن‌هایی که برپایه‌ی لغت‌نامه هستند، بهترین نتیجه را با کدینگ هسته‌ی فیشر کسب می‌کنند. البته در مجموع روش‌های بر پایه‌ی یادگیری عمیق به خاطر ویژگی‌هایی که یاد می‌گیرند بهترین عملکرد را دارند. در ادامه دلیل اصلی ورود به دسته‌بندی ریزدانه‌ای تصویر را مطرح می‌کنیم.

یکی از بهترین و بزرگترین پایگاه داده‌ها برای مقایسه‌ی روش‌های دسته‌بندی تصویر پایگاه داده‌ی ImageNet [۲] است. بر روی این پایگاه داده و زیر مجموعه‌ی آن یعنی ILSVRC [۲۶] روش‌های یادگیری عمیق مثل [۲۵، ۲۹، ۳۰] بهترین نتایج را کسب کرده‌اند. از نتایج این روش‌ها می‌توان

(آ) *پایگاه داده‌ی ماشین‌های استنفورد*

(ب) *پایگاه داده‌ی هواپیماهای آکسفورد*

(ج) *پایگاه داده‌ی پرندگان کلتک*

**شکل ۸,۳:** *برخی از پایگاه داده‌های دسته‌بندی تصاویر ریزدانه‌ای*

نکات جالبی را استخراج کرد. برای نمونه تصویر ۷,۲ ماتریس ابهام [۲۵] را برای روش [۳۱] بر روی پایگاه داده ImageNet نشان می‌دهد. نکته‌ای که باید به آن توجه شود ساختار موجود در ماتریس ابهام است. همانطور که در تصویر نیز مشخص است، مدل دسته‌بند بیشترین ابهام را در دسته‌هایی دارد که به یکدیگر از نظر ظاهری و معنایی نزدیک هستند. برای نمونه در بین پرنده‌ها ابهام زیاد وجود دارد و از پرنده‌ها در بین پرنده‌های آبی ابهام زیادتر است. به عبارت دیگر ابهام در بین دسته‌های ریزدانه‌ای است. این موضوع باعث می‌شود که نتیجه بگیریم برای افزایش دقت روش‌های دسته‌بندی تمرکز بر روی دسته‌های ریزدانه‌ای بسیار مهم است. در ادامه به بررسی پایگاه داده‌ها و برخی از این روش‌ها می‌پردازیم.

## ۳ دسته‌بندی ریزدانه‌ای تصویر

همانطور که گفته شد، اگر در دسته‌بندی تصویر، دسته‌های مورد بررسی زیر دسته‌ی [۲۶] یک دسته‌ی کلی‌تر باشند(مانند گونه‌های مختلف پرندگان، مدل‌های مختلف خودروهای سواری و انواع مختلف هواپیماها)، آنگاه مسئله را «دسته‌بندی ریزدانه‌ای تصویر» می‌نامند. در قسمت ۳,۲ دلیل اصلی برای مطالعه بر روی دسته‌بندی ریزدانه‌ای را بررسی کردیم.

محققان در حوزه‌ی روانشناسی از سال‌ها قبل بر روی توانایی‌های انسان در دسته‌بندی مطالعه می‌کرده‌اند. برای نمونه دانشمندان نشان دادند [۳۲] که توانایی انسان برای انجام دسته‌بندی سطح اصلی [۲۷] یا معمولی (تشخیص خودروی سواری از گربه یا تشخیص خیابان از خانه) خیلی قبل‌تر از توانایی انسان برای دسته‌بندی سطح فرعی [۲۸] یا ریزدانه‌ای توسعه می‌یابد. خیلی جالب است که در زمینه‌ی بینایی کامپیوتری هم همین منوال حفظ شده است، یعنی محققین ابتدا بر روی روش‌های دسته‌بندی سطح اصلی یا معمولی کار زیادی انجام داده‌اند و پیشرفت‌های خوبی حاصل شده است، امام اخیراً کار بر روی دسته‌بندی ریزدانه‌ای آغاز شده است. در ادامه به بررسی برخی از پایگاه داده‌های دسته‌بندی ریزدانه‌ای خواهیم پرداخت.

## ۱,۳ پایگاه داده‌های دسته‌بندی ریزدانه‌ای

از بین پایگاه داده‌های مربوط به دسته‌بندی ریزدانه‌ای سه پایگاه داده مهم را انتخاب کرده‌ایم و مطالعات را بر روی آن‌ها انجام خواهیم داد. در ادامه ابعاد و نوع داده های موجود در این پایگاه داده‌ها را بررسی خواهیم کرد.

### ۱.۱,۳ ماشین‌های استنفورد

این پایگاه داده در سال ۲۰۱۳ معرفی شد [۳۳] و شامل ۱۶۱۸۵ تصویر از ۱۹۶ مدل مختلف خودروهای سواری است. برای هر خودرو جعبه‌ی دور [۲۹] خودرو هم داده شده است. که برای حذف پس زمینه کاربرد زیادی دارد. نمونه‌ای از تصاویر این پایگاه داده را در شکل ۸,۳آ مشاهده می‌کنید. برای این پایگاه داده چون می‌دانیم که خودروهای سواری را مطالعه می‌کنیم، بسیاری از روش‌ها [۳۳، ۳۴، ۳۵] سعی می‌کنند که از دانش سه‌بعدی که در مورد خودروها داریم برای حل کردن بهتر مسئله استفاده کنند. این دانش سه‌بعدی شامل مدل سه‌بعدی و شکل خودروی سواری است که می‌توان از طریق اینترنت به آن‌ها دست پیدا کرد.

### ۲.۱,۳ هواپیماهای آکسفورد

این پایگاه‌داده جدید در سال ۲۰۱۴ معرفی شد [۳۶] و شامل ۷۵۰۰ تصویر از ۷۵ نوع مختلف هواپیما است. برای هر هواپیما جعبه دور هواپیما به همراه تعداد زیادی از ویژگی‌های اجزای آن (مثل تعداد موتور، تعداد چرخ، شکل بال‌ها و ...) به همراه ناحیه‌بندی [۳۰] آن جزء هواپیما موجود است. این پایگاه داده از نظر کامل بودن بسیار کامل است، ولی به خاطر جدید بودن هنوز مقاله‌ای به جز مقاله‌ی معرفی خود پایگاه داده از آن استفاده نکرده‌است. ضمناً اندازه داده‌های آن به نسبت دیگر پایگاه داده‌ها کم است. نمونه‌ای از تصاویر این پایگاه داده را در شکل ۸,۳ب مشاهده می‌کنید.

### ۳.۱,۳ پرندگان کلتک

این پایگاه داده به نسبت قدیمی است و در سال ۲۰۱۰ و ۲۰۱۱ معرفی شده است [۳، ۳۷]. این پایگاه داده شامل ۱۱۷۸۸ تصویر از ۲۰۰ گونه مختلف از پرندگان است. برای هر پرنده ۱۵ جزء بدن (برای مثال بال راست، سر، پای چپ و ...) به همراه جعبه دور آن مشخص شده است. نمونه‌ای از تصاویر این پایگاه داده را در شکل ۸,۳ج مشاهده می‌کنید. بیشتر مقالات کنونی بر روی این پایگاه‌داده نتایج خود را گزارش می‌کنند، زیرا هم قدمتش بیشتر است و هم تعداد داده‌ها و نوع داده‌های همراه آن (۱۵ جزء) بسیار مفید است. در مرور کارهای پیشین بیشتر مقالاتی که بر روی این پایگاه داده نتایج خود را گزارش کرده‌اند بررسی می‌کنیم. همچنین در بخش ۱,۴ نتایج ابتدایی خود را بر روی این پایگاه داده گزارش می‌کنیم.

## ۲,۳ روش‌های همراه با انسان

در این روش هدف این است که دسته‌بندی نه به صورت کاملاً خودکار بلکه با کمک انسان و به صورت نیمه خودکار صورت بگیرد. سامانه بدین صورت کار می‌کند که پس از دریافت تصویر ورودی، سامانه از کاربر تعدادی سؤال می‌پرسد (برای مثال آیا رنگ بال پرنده آبی است یا خیر؟) پس از پاسخ به این سؤال سامانه تخمین خود از دسته‌ای که تصویر مربوط به آن است را به روز می‌کند (شکل ۹,۳آ). این ایده از آنجا آمده است که دسته‌بندی معمولی برای انسان کار راحتی است ولی دسته‌بندی ریزدانه‌ای کار سختی است و افراد غیر متخصص توانایی انجام دقیق آن را ندارند. ولی تشخیص صفت‌های یک دسته‌ی ریزدانه‌ای مثل رنگ یک جزء کار آسانی برای انسان است (شکل ۹,۳ب). در [۳۸] سامانه به گونه‌ای طراحی شده است که با کمترین تعداد پرسش بهترین نتیجه دسته‌بندی حاصل شود.

فرض کنید که دسته‌های ما $c \in \{۱, ..., C\}$ باشند. در هر گام هدف استفاده از اطلاعات درون تصویر و سابقه‌ی پرسش‌ها و پاسخ‌های انجام شده تا به حال برای انتخاب سؤال بعدی است. فرض کنید که $Q = \{q_۱, ..., q_n\}$ مجموعه کل پرسش‌ها باشد، $A_i$ پاسخ‌های ممکن به سؤال $q_i$ است. $a_i \in A_i$ متغییر تصادفی پاسخ کاربر به سؤال است. کاربر به غیر از پاسخ می‌توان اطمینان خود را از پاسخی که داده است را هم مشخص کند، که با $r_i$

(آ) نحوه به روزرسانی تخمین با پرسش از کاربر

(ب) انسان به چه سؤالاتی می‌تواند پاسخ دهد؟

**شکل ۹,۳:** نمای کلی روش همراه با انسان

نمایش داده می‌شود و $r_i \in \mathcal{V}$. برای مثال می‌توانیم قرار دهیم $\mathcal{V} = \{$حدس، تقریبا، مطمئن$\}$ پس کاربر زوج مرتب متغییرهای تصادفی $u_i = (a_i, r_i)$ را به سامانه خواهد داد. در هر زمان $t$ باید سؤال $q_{j(t)}$ را انتخاب کرد $j(t) \in \{۱, ..., n\}$. تعریف می‌کنیم $U^{t-۱} = \{u_{j(۱)}, ..., u_{j(t-۱)}\}$ مجموعه‌ی پاسخ‌هایی است که در گام‌های قبلی توسط کاربر پاسخ داده شده است. همچنین $I(c; u_i|x, U^{t-۱})$ را امید ریاضی بهره اطلاعات در صورت بیان کردن سؤال $q_i$ تعریف می‌کنیم. حال سؤالی را انتخاب می‌کنیم که امید ریاضی ما از بهره اطلاعات را بیشینه کند. امید ریاضی بهره اطالعات پس از ساده سازی به صورت زیر خواهد بود:

$$I(c; u_i|x, U^{t-۱}) = \sum_{u_i \in A_i \times \mathcal{V}} p(u_i|x, U^{t-۱}) \Big( H(c|x, u_i \cup U^{t-۱}) - H(c|x, U^{t-۱}) \Big) \tag{۳}$$

و همچنین داریم:

$$H(c|x, U^{t-۱}) = -\sum_{c=۱}^{C} p(c|x, U^{t-۱}) log p(c|x, U^{t-۱}) \tag{۴}$$

در معادله‌ی ۳ احتمال $p(u_i|x, U^{t-۱})$ را مدل کاربر و در معادله‌ی ۴ احتمال $p(c|x, U^{t-۱})$ مدل بینایی ماشین تعریف می‌کنیم. که بستگی به مدلی که انتخاب می‌کنیم می‌تواند متفاوت باشد.

نتایج بدست آمده در این [۳۸] مقاله نیز جالب است. اگر هیچ سؤالی پرسیده نشود، دقت دسته‌بندی ۱۹ درصد است و با بیست سؤال دقت به بالای ۵۰ درصد ارتقا پیدا می‌کند.

این روش از نظر تئوری بسیار زیباست ولی هدف ما دسته‌بندی کاملا خودکار است، به همین خاطر از این روش‌ها می‌گذریم.

## ۳,۳ روش‌های جدید دیگر

در سال‌های اخیر روش‌های مختلفی برای حل مسئله دسته‌بندی ریزدانه‌ای تصاویر مطرح شده است که دسته‌بندی آن‌ها سخت است. در ادامه به چند محور مشترک در این روش‌ها خواهیم پرداخت.

(آ) *Blue headed vireo*

(ب) *Red eyed vireo*

**شکل ۱۰,۳:** *دوگونه‌ی پرنده که در رنگ اطراف چشم تفاوت دارند*

### ۱.۳,۳ پیدا کردن اجزای مهم به صورت خودکار

بسیاری از روش‌ها [۳۹، ۴۰، ۴۱، ۴۲] سعی می‌کنند به صورت خودکار بخش‌هایی از تصویر را پیدا کنند که قابلیت تمایز بالایی دارند. برای درک بهتر به تصویر دو گونه‌ی پرنده شکل ۱۰,۳ نگاه کنید. پیدا کردن تفاوت بین این دو گونه تنها در صورتی که بتوانیم چشم‌های پرنده را در تصویر پیدا کنیم امکان پذیر است. این روش‌ها اکثراً بر پایه‌ی محاسبات سنگینی هستند که در گام‌های پیاپی سامانه‌های تشخیص شیء [۳۱] مخصوص پیدا کردن این اجزاء تمایز دهنده را آموزش می‌دهند که کار بسیار سنگینی از نظر محاسباتی است. (ما توانایی محاسباتی آن را نداریم.)

بهترین نتیجه‌ای که این روش‌ها می‌گیرند [۴۲] بر روی پایگاه داده‌ی [۳] برابر ۵۶,۸ درصد است. نکته جالب این است که اگر از اجزای داده شده اصلی استفاده کنند دقت این روش به ۷۳,۳ درصد افزایش می‌یابد که اهمیت مکان اعضا را نشان می‌دهد. دیگر کارهای در این ارتباط پیدا کردن اجزا توسط جمع سپاری [۳۲] است [۴۳] که به خاطر خودکار نبودن مورد بررسی قرار نمی‌گیرد.

### ۲.۳,۳ انتقال اطلاعات از داده‌های آموزشی به داده‌های آزمایشی

اهمیت مکان اجزا در قسمت قبل بیان شد. حال سؤال این است که اگر اطلاعات اجزا را در داده‌های آموزشی داشته باشیم، آیا می‌توانیم در داده‌های آزمایشی این اطلاعات را بدست آوریم. واضح است که هرچه در زمان آزمایش نیاز به اطلاعات کمتری داشته باشیم بهتر است. مقالات جدید [۴۴، ۴۵] بر اساس مقاله مهم [۴۶] سعی بر این کار دارند و نتایج بسیار خوبی نیز کسب می‌کنند. فرض این روش‌ها این است که در صورتی که برای داده‌ی آزمایشی، از بین داده‌های آموزشی موردی را پیدا کنیم که از نظر شکل [۳۳] به داده‌ی آزمایشی شبیه باشند، می‌توان اطلاعات اجزا را از آن داده‌های آموزشی به داده‌ی آزمایشی انتقال داد و تخمینی از مکان اجزا بدست آورد. با این تخمین ویژگی‌های هسته‌ی فیشر یا کیسه‌ای از لغات را برای هر کدام از اجزا بدست می‌آورد، سپس دسته‌بندی را با این ویژگی‌ها انجام می‌دهد. با این ایده روش [۴۴] به دقت میانگین ۶۲,۷ درصد و روش [۴۵] به دقت میانگین ۵۷,۸۴ درصد دست می‌یابد.

### ۳.۳,۳ یادگیری عمیق

همانطور که در بخش ۲,۲ گفته شد، روش‌های یادگیری عمیق به خاطر قابلیت‌های بالایشان در یادگیری ویژگی سریعاً برای دسته‌بندی ریزدانه‌ای نیز مورد استفاده قرار گرفتند. برای نمونه در این رابطه می‌توان به [۴۷، ۳۴، ۴۸] اشاره کرد. این روش‌ها به قدری قوی هستند که اگر حتی ویژگی‌ها بر روی پایگاه داده‌ی دیگری به جز آنچه بر روی آن آزمایش انجام می دهیم نیز آموزش داده شوند، باز نتایج بسیار خوبی کسب می‌کنند [۴۷]. به علت قدرت این روش‌ها ما بر روی آن‌ها آزمایشاتی انجام دادیم که در بخش ۱,۴ به آن اشاره می‌کنیم. همچنین دقت روش‌های مختلف در بخش ۱,۴ با یکدیگر مقایسه شده‌اند.

**شکل ۱۱,۴:** *ساختار شبکه عصبی موسوم به AlexNet[۲۵] که در آزمایشات مورد استفاده قرار گرفته است.*

# ۴ روش‌پیشنهادی و جمع‌بندی

آنچه بیان شد، خلاصه‌ای از روش‌های مهم در دسته‌بندی تصویر معمولی و دسته‌بندی ریزدانه‌ای تصویر بود. دیدیم که در روش‌هایی که بر پایه‌ی کیسه‌ای از لغات هستند با کدینگ هسته‌ی فیشر بهترین نتیجه را کسب می‌کنند. از این روش‌ها بهتر روش‌های مبتنی بر یادگیری عمیق بودند که مطالعه تئوری آن‌ها فراتر از مباحث این گزارش می‌باشد. لذا در ادامه به این روش‌ها به عنوان یک روش استخراج ویژگی نگاه می‌کنیم و به جزئیات آن‌ها نمی‌پردازیم. در ادامه به بررسی روش پیشنهادی می‌پردازیم.

## ۱,۴ روش پیشنهادی و نتایج اولیه

مطابق بهترین روش‌ها پیشنهاد می‌شود که برای استخراج ویژگی از ویژگی‌ها مدل‌های یادگیری عمیق استفاده شود. همانطور که در [۴۷] نیز اشاره شده است، لازم نیست که شبکه عصبی مصنوعی عمیق را برای هر پایگاه‌داده‌ای آموزش دهیم (البته در برخی موارد آموزش شبکه عصبی مصنوعی عمیق با توجه به تعداد کم داده‌های ممکن نیست)، اگر یک بار شبکه بر روی پایگاه‌داده‌ای مثل ILSVRC [۲۶] آموزش داده شود می‌توان از همان ویژگی‌های یادگرفته شده برای پایگاه داده‌های دیگر نیز استفاده کرد [۴۷]. برای پیاده‌سازی ما از ابزار متن باز Caffe [۴۹] استفاده کردیم. به عنوان مدل هم از شبکه عصبی معروف AlexNet [۲۵] که در شکل ۱۱,۴ قابل مشاهده است، استفاده شده. این شبکه ابتدا با داده‌های ILSVRC [۲۶] آموزش داده شده است و برای استخراج ویژگی توسط راه حل پیشنهادی ما مورد استفاده قرار می‌گیرید. به عنوان مدل یادگیرنده از SVM [۹] با هسته‌ی خطی و مقدار $C = ۳$ استفاده شد. با این ابزارها سه آزمایش مختلف انجام گرفت. که تفاوت آن‌ها در داده‌های آموزشی است. پایگاه داده مورد استفاده نیز CUB-200-2011 [۳] انتخاب شد که به راحتی بتوان با روش‌های مختلف مقایسه انجام داد.

**آزمایش اول:** در آزمایش اول از هیچ اطلاعات کمکی مثل جعبه دور شیء و اجزا استفاده نشد. عکس‌های آموزشی و آزمایشی برای استخراج ویژگی بعد از تغییر اندازه به ۲۵۶ × ۲۵۶ به شبکه عصبی به عنوان ورودی داده می‌شد و اطلاعات لایه‌ی یکی مانده به آخر به عنوان ویژگی استخراج می‌شد. این بردار ویژگی دارای ۴۰۹۶ بعد بود. سپس با این ویژگی‌ها یک SVM خطی آموزش داده می‌شد. داده‌های آموزش و آزمایش نیز طبق تقسیم‌بندی پایگاه داده استفاده شد. دقت در این روش برابر ۴۵,۴۴ درصد است.

**آزمایش دوم:** فرق آزمایش اول با آزمایش دوم در این بود که ابتدا با استفاده از جعبه دور شیء، تصویر بریده می‌شود، سپس همان مراحل آزمایش اول تکرار می‌شود. دقت در این آزمایش به میزان قابل توجهی افزایش پیدا کرد و به ۵۵,۶۲ درصد رسید. این موضوع نشان دهنده‌ی گمراه کننده بودن اطلاعات موجود در پس زمینه‌ی اشیاء در این پایگاه داده است. پس هرچه بتوان این اطلاعات را حذف کرد موفقیت بیشتری حاصل می‌شود.

**آزمایش سوم:** در این آزمایش ویژگی‌های استخراج شده در آزمایش اول و دوم به هم چسبانده شد و یک بردار ویژگی ۸۱۹۲ بعدی برای هر تصویر بدست آمد. با این کار نتایج باز هم افزایش یافت و به ۵۷,۸۳ درصد رسید که قابل قبول است.

**مقایسه با روش‌های مهم:** در جدول زیر آزمایش‌های خود را با بهترین نتایج مقایسه می‌کنیم.

**جدول ۱,۴:** *مقایسه میانگین دقت روش ارائه شده با روش‌های دیگر*

| روش | سال ارائه | استفاده از جعبه دور | استفاده از اطلاعات اجزا | دقت ٪ |
|---|---|---|---|---|
| CUB [۳] | ۲۰۱۱ | ✓ | | ۱۰,۳ |
| CUB [۳] | ۲۰۱۱ | ✓ | ✓ | ۱۷,۳ |
| PPK [۵۰] | ۲۰۱۲ | ✓ | | ۲۸,۱۸ |
| style [۴۱] | ۲۰۱۳ | ✓ | | ۴۱,۰۱ |
| POOF [۴۲] | ۲۰۱۳ | ✓ | | ۵۶,۸ |
| Alignments [۴۴] | ۲۰۱۳ | ✓ | | ۶۲,۷ |
| DPD [۵۱] | ۲۰۱۳ | ✓ | ✓ | ۵۱,۰ |
| Decafe [۵۲] | ۲۰۱۴ | ✓ | | ۶۵,۰ |
| PNDCN [۴۸] | ۲۰۱۴ | ✓ | | ۷۵,۷ |
| CNN [۴۷] | ۲۰۱۴ | ✓ | | ۵۳,۳ |
| CNN-aug [۴۷] | ۲۰۱۴ | ✓ | | ۶۱,۰۸ |
| Ours-۱ | - | | | ۴۵,۴۴ |
| Ours-۲ | - | ✓ | | ۵۵,۶۲ |
| Ours-۳ | - | ✓ | | ۵۷,۸۳ |

## ۲,۴ کارهای آینده

همانطور که در جدول ۱,۴ می‌توان دید روش پیشنهادی نتایج قابل مقایسه و خوبی نسبت به روش‌های دیگر دارد. البته فعلاً روش ما از اطلاعات سه‌بعدی، اطلاعات اجزا در داده‌های آموزشی و از افزودن داده‌های آموزشی [۳۴] استفاده نمی‌کند. همچنین روش‌ها معمولاً بر روی تمام دسته‌ها خوب جواب نمی‌دهد و اینکه روش‌های مختلف بر روی دسته‌های مختلفی خوب عمل می‌کنند، تحلیل دلیل ضعف و قوت هر کدام از روش‌ها می‌تواند به ما در طراحی مدل‌های بهتر کمک کند.

**جدول ۲,۴:** *زمان‌بندی انجام پروژه*

| عنوان فعالیت | مدت زمان لازم | درصد پیشرفت | زمان اتمام |
|---|---|---|---|
| مطالعه روش‌های پیشین برای دسته‌بندی معمولی و دسته‌بندی ریزدانه‌ای | ۰,۵ ماه | ۱۰۰ | شهریور ۹۳ |
| پیاده‌سازی و اجرای روش‌های پایه | ۱ ماه | ۵۰ | آبان ۹۳ |
| پیاده‌سازی روش پیشنهادی جدید | ۳ ماه | ۲۰ | بهمن ۹۳ |
| نگارش پایان‌نامه | ۱ ماه | ۱۰ | بهمن ۹۳ |
| جمع‌بندی نتایج و نوشتن مقاله | ۰,۵ ماه | ۰ | اسفند ۹۳ |

# 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.

[13] F. Perronnin, C. Dance, G. Csurka, and M. Bressan, "Adapted vocabularies for generic visual categorization," in *ECCV*, 2006.

[14] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *ICCV*, 2005.

[15] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *CVPR*, 2008.

[16] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *CVPR*, 2005.

[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[18] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009.

[19] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in neural information processing systems*, 2006.

[20] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[21] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007.

[22] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.

[23] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *BMVC*, 2011.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, 1998.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2014.

[27] D. Parikh and C. L. Zitnick, "The role of features, algorithms and data in visual recognition," in *CVPR*, 2010.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[29] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[30] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional neural networks," in *ECCV*, 2014.

[31] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in *ECCV*, 2010.

[32] K. E. Johnson and A. T. Eilers, "Effects of knowledge and development on subordinate level categorization," *Cognitive Development*, vol. 13, no. 4, pp. 515–545, 1998.

[33] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[34] J. Krause, T. Gebru, J. Deng, L.-J. Li, and L. Fei-Fei, "Learning features and parts for fine-grained recognition," in *International Conference on Pattern Recognition*, Stockholm, Sweden, August 2014.

[35] K. Ramnath, S. N. Sinha, R. Szeliski, and E. Hsiao, "Car make and model recognition using 3d curve alignment," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[36] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. B. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss *et al.*, "Understanding objects in detail with fine-grained attributes," in *CVPR*, 2014.

[37] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-ucsd birds 200," Caltech, Tech. Rep. CNS-TR-201, 2010.

[38] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," in *ECCV*, 2010.

[39] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *CVPR*, 2011.

[40] B. Yao, G. Bradski, and L. Fei-Fei, "A codebook-free and annotation-free approach for fine-grained image categorization," in *CVPR*, 2012.

[41] Y. J. Lee, A. A. Efros, and M. Hebert, "Style-aware mid-level representation for discovering visual connections in space and time," in *ICCV*, 2013.

[42] T. Berg and P. N. Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *CVPR*, 2013.

[43] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *CVPR*, 2013.

[44] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in *ICCV*, 2013.

[45] C. Goring, E. Rodner, A. Freytag, and J. Denzler, "Nonparametric part transfer for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[46] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[47] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," *arXiv preprint arXiv:1403.6382*, 2014.

[48] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014.

[49] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding," http://caffe.berkeleyvision.org/, 2013.

[50] N. Zhang, R. Farrell, and T. Darrell, "Pose pooling kernels for sub-category recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[51] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[52] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

## واژه‌نامه

Image classification[۱]
Computer - Keyboard[۲]
Subclass[۳]
Fine-grained image classification[۴]
Tern[۵]
Discriminative[۶]
Content based image retrieval[۷]
Surveillance systems[۸]
Object classification[۹]
Scene classification[۱۰]
Attribute recognition[۱۱]

Action recognition[۱۲]
Coding[۱۳]
Keypoints[۱۴]
Dense[۱۵]
Vector quantization[۱۶]
Clusters[۱۷]
Bag of keypoints[۱۸]
Pooling[۱۹]
Average pooling[۲۰]
Spatial pyramid matching[۲۱]
Support Vector[۲۲]
Regularization[۲۳]

Deep learning[۲۴]
Confusion matrix[۲۵]
Subclass[۲۶]
Basic level[۲۷]
Sub-ordinate level[۲۸]
Bounding box[۲۹]
Segmentation[۳۰]
Object detection[۳۱]
Crowdsourcing[۳۲]
Shape[۳۳]
Data augmentation[۳۴]