ISSN: 2821-157X | htpps://www.jasd.khadu.ir

Journal of Aerospace Defense
Volume 2, Issue 5
Summer2026
P.P.1- 27

Research Paper;

Fine-Grained Visual Classification Using a Convolutional Sparse Auto-Encoder

Farhad Sadeghi Almalou[1], Farbod Razzazi*[2], Arash Amini[3]

1. Ph.D. Student, Department of Electrical and Computer Engineering, SR.C, Islamic Azad University, Tehran, Iran.
E-mail: farhad.sadeghialmalou@iau.ac.ir
2. Associate Professor, Department of Electrical and Computer Engineering, SR.C, Islamic Azad University, Tehran, Iran.
E-mail: farbod_razzazi@iau.ac.ir
3. Professor, Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran.
E-mail: **aamini@sharif.ir**

Article Information

Accepted:
2026/07/27

Recceived:
2026/03/02

Keywords:

Convolutional sparse network; Multi-layer dictionary learning; Feature extraction; Fine-grained visual classification; Sparse auto encoder

Corresponding Author:

Farbod Razzazi
Email:
farbod_razzazi@iau.ac.ir

Abstract

Introduction: In this research, a novel multi-layer architecture is introduced based on Dictionary Learning and Convolutional Sparse Coding, termed the Convolutional Sparse Network (CSN) to enhance fine-grained classification accuracy. Fine-grained classification is considered one of the most challenging tasks in this domain due to significant intra-class variance and subtle inter-class similarities, necessitating the extraction of highly discriminative features.

Methods: To achieve optimal results in this type of classification, deep neural network-based models are designed to localize informative regions at both the image and region levels, enabling the extraction and utilization of discriminative features. The proposed model maintains the high-level feature extraction capabilities of convolutional neural networks while simultaneously mitigating redundancy across the layer depth. These two factors lead to a form of implicit localization, which facilitates the extraction of discriminative features that are effective and efficient for fine-grained tasks.

Findings: Within this framework, we introduce a Sparse Autoencoder (SAE) which, when integrated with a standard pre-trained deep neural network, improves classification accuracy on fine-grained datasets.

Conclusion: The proposed hybrid model increases classification accuracy on the CUB-200 dataset by 3.5% compared to the base model and achieves a 28% reduction in the misclassification error rate.

Sadeghi Almalou, Farhad; Razzazi, Farbod and Amini, Arash.(2026). Fine-Grained Visual Classification Using a Convolutional Sparse Auto-Encoder. *Journal of Aerospace Defense*, Vol2(Issue5), Page1-27

ISSN: 2821-157X | htpps://www.jasd.khadu.ir

فصلنامه علمی دفاع هوافضایی
دورهٔ ۵، شمارهٔ ۲
تابستان ۱۴۰۵
صفحات ۲۷-۱

مقاله پژوهشی؛

طبقه‌بندی ریزدانه دیداری مبتنی بر یک اتوانکدر کانولوشنال تنک

فرهاد صادقی آلمالو[۱]، فربد رزازی*[۲]، آرش امینی[۳]

۱. دانشجوی دکتری، دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات، تهران، ایران
رایانامه: farhad.sadeghialmalou@iau.ac.ir

۲. دانشیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، واحدعلوم و تحقیقات، تهران، ایران.
رایانامه: farbod_razzazi@iau.ac.ir

۳. استاد، دانشکده مهندسی برق، دانشگاه صنعتی شریف، تهران، ایران.
رایانامه: aamini@sharif.ir

اطلاعات مقاله

تاریخ دریافت:
۱۴۰۴/۱۲/۱۱

تاریخ پذیرش:
۱۴۰۵/۰۵/۰۵

کلیدواژه‌ها:
شبکه کانولوشنال تنک
؛آموزش لغت‌نامه‌های چند لایه؛
استخراج ویژگی؛
طبقه‌بندی دیداری ریزدانه؛
اتوانکدر تنک

نویسندهٔ مسئول:
فربد رزازی
ایمیل:
farbod_razzazi@iau.ac.ir

چکیده

مقدمه: در این تحقیق یک معماری متفاوت چند لایه بر مبنای یادگیری لغت-نامه و کدگذاری کانولوشن بنام شبکهٔ کدگذاری تنک جهت ارتقاء دقت طبقه-بندی ریزدانه معرفی است. طبقه‌بندی ریزدانه به دلیل تنوع درون‌کلاسی و شباهت‌های ظریف بین‌کلاسی یکی از چالش‌برانگیزترین مأموریت‌های این حوزه محسوب شده و نیازمند ویژگی‌های تفکیک‌کننده می‌باشد.

روش: برای رسیدن به نتایج مطلوب در این نوع طبقه‌بندی، مدل‌های مبتنی بر شبکه‌های عصبی عمیق به گونه‌ای طراحی می‌شوند که بتوانند با مکان‌یابی بخش‌های حاوی اطلاعات مهم در سطح تصویر و در سطح ناحیه، ویژگی‌های تفکیک‌کننده را استخراج و در طبقه‌بندی استفاده نمایند. مدل پیشنهادی قابلیت استخراج ویژگی‌های سطح بالا را همچون شبکه‌های عصبی عمیق کانولوشنال دارد و از طرفی همزمان از افزونگی اطلاعات در عمق لایه‌ها جلوگیری می‌کند. این دو عامل منجر به نوعی مکانیابی ضمنی می‌شود که استخراج ویژگی‌های تفکیک‌کننده از این نواحی می‌تواند در مأموریت ریزدانه موثر و کارآمد باشد.

یافته‌ها: ما در این پژوهش بر مبنای معماری یاد شده، یک اتوانکدر تنک معرفی کرده‌ایم که ترکیب آن با یک شبکهٔ عمیق عصبی از پیش آموزش داده شده استاندارد، دقت طبقه‌بندی در مجموعه داده‌های ریزدانه را بهبود بخشید.

نتیجه‌گیری: مدل ترکیبی پیشنهادی دقت طبقه‌بندی را به میزان ۳/۵ درصد در مجموعه دادهٔ CUB-200 نسبت به مدل پایهٔ مورد استفاده در این معماری افزایش می‌دهد و نرخ خطای تشخیص کلاس‌های نادرست را در این مجموعه داده به میزان ۲۸ درصد کاهش می‌دهد.

استناد: صادقی آلمالو، فرهاد؛ رزازی، فربد؛ امینی، آرش. (۱۴۰۶). طبقه‌بندی ریزدانه دیداری مبتنی بر یک اتوانکدر کانولوشنال تنک. فصل نامه دفاع هوافضایی، دوره۵ (شماره۲)، صفحات ۲۷-۱

۱–مقدمه

طبقه بندی دیداری ریز دانه (FGVC)[۱] به عنوان یکی از مهم‌ترین و چالش‌برانگیزتـرین زمینـه‌هـای پژوهشی در حوزه یادگیری ماشین و بینایی ماشـین شـناخته مـی‌شـود کـه توجـه زیـادی را در بـین پژوهشگران و محقیق به خود جلب کرده است. این نوع طبقـه‌بنـدی بـه فرآینـد شناسـایی اشـیاء بـا زیرگروه‌های بسیار مشابه اشاره دارد. به عنوان مثال، تشخیص انواع مختلف پرندگان [1]، هواپیمـاهـا [2]، حیوانات [3] یا خودروها [4] که ممکن است ظاهری تقریباً یکسان داشته باشند، نمونـه‌هـایی از این نوع طبقه‌بندی هستند. از آنجا که طبقه‌بندی ریزدانه در کاربردهـای متنـوعی از جملـه شناسـایی گونه‌های گیاهی و جانوری (زیست‌شناسی)، نظارت بر ترافیک شهری، بهبود سـامانه‌هـای جسـتجو و مراقبت تصویری در کاربردهای نظامی– امنیتی و غیر نظامی، کنترل ترافیـک هـوائی، احـراز هویـت کارکنان سازمان‌ها و نظایر آن، نقش مهمی ایفاء می‌نماید، افزایش کیفیت و دقت این نوع طبقه‌بنـدی می‌تواند تأثیر عمیقی بر نتایج حاصل از این سامانه‌ها داشته باشـد. امـا اسـتفاده از روش‌هـای سـنتی یادگیری ماشین در این عرصه ممکن است به دلیل حساسیت بالای ویژگی‌ها و پیچیدگی‌های موجود در داده‌ها به نتایج مطلوب نرسد. چالش طبقه‌بندی ریزدانه از سـه عامـل اصـلی ناشـی مـی‌شـود: ۱) شباهت‌های بین کلاسی خیلی زیاد دسته‌ها، ۲) تفاوت و تنوع زیـاد درون‌کلاسـی ماننـد نمـای دیـد، چرخش، مقیاس، و ۳) تعداد نمونه‌های آموزشی کم بـه دلیـل پیچیـدگی و نیـاز بـه مهـارت زیـاد در برچسپ زدن به نمونه‌ها. بنابراین دست یافتن به دقت بالای طبقه‌بندی حتی با استفاده از مـدل‌هـای استخراج ویژگی پیچیدهٔ عمیق CNN [۲] استاندارد مانند VGG [5] و ResNet [6] نیز امکـان پـذیر نیست و تشخیص این تفاوت‌ها به مدل‌ها و معماری‌های دقیق و کارامدی نیاز دارد.

به دلایل گفته شده موفقیت در FGVC نیازمند داشتن ویژگی‌های محلی و مکان‌یـابی بخـش‌هـای تمایز بخش و استخراج ویژگی از آنهاست و یا نیازمند سازوکارهای توجـه[۳] و تلفیـق[۴] ویژگـی‌هاسـت. ادبیات حاکم بر این موضوع از شبکه‌های کانولوشنال پایه و تنظیم نهائی شده[۵] آغاز شد و با مدل‌هـای CNN دوجمله‌ای[۶] ویژگی‌ها جهت گرفتن برهم کنش دو به دو محلی تقویت گردید [7]. این رویکرد امروزه با بهره‌گیری از شبکه‌های از پیش آموزش داده شده و ترنسفورمرها که قـادر بـه مـدل کـردن همسایگی دور هستند بهبود یافته است [8]. نمونه‌های اندک و برچسب‌گذاری پر هزینه نیز طراحی و پیشنهاد معماری‌های نظارت ضعیف و کم‌نمونه را پر رنگ‌تر نموده‌اند.

---

[۱] Fine-Grained Visual Classification

[۲] Convolutional Neural Network

[۳] Attention

[۴] Fusion

[۵] Fine-tune

[۶] Bilinear-CNN

بسیاری از رویکردهای موفق FGVC یا بصورت آشکار و صریح از ناحیه‌های تفکیک‌کننده استفاده می‌کنند و یا به طور ضمنی از نقشه‌های توجه و فعال‌سازی بمنظور تمرکز روی اجزای کلیدی اطلاعات بهره می‌برند. به عنوان مثال در مدل SEF [1]ماژول گروه‌بندی معنائی[2] زیرگروه‌های ویژگی را در کانال‌های مختلف تولید می‌کند تا اطلاعات تفکیک‌کننده از ناحیه‌ها[3] همراه اطلاعات جامع استخراج شود. در این مدل دو نوع ویژگی تولید شده با وزن‌دهی مناسب با یکدیگر تجمیع می‌شوند تا دقت طبقه‌بندی را افزایش دهند [9]. در روش AKEN، بلوک‌های کدگذاری طولی و عرضی[4] بکار رفته‌اند تا بتوانند اطلاعات تفکیک‌پذیر ناحیه‌ها را استخراج نمایند [10]. در مدل API بلوک تعامل دو به دوی همراه با توجه[5] بکار رفته است که می‌تواند یک جفت تصویر ریزدانه را از هم تشخیص دهد [11]. مدل DB-GBL دارای دو بلوک تنوع‌بخشی[6] و افزایش شیب[7] هست که اولی نواحی برجسته را مکان‌یابی می‌کند و دومی خطای پیشبینی کلاس‌ها را کاهش می‌دهد [12]. در مدل FDL جهت افزایش دقت طبقه‌بندی با استفاده از شبکهٔ پیشنهاد نواحی (RPN)[8] مبتنی بر یادگیری فیلترها، نواحی حاوی اطلاعات تولید و انتخاب می‌شوند [13]. مدل PGM به جای مکان‌یابی نواحی تصویر، یک فرآیند یادگیری تدریجی[9] را معرفی نموده که ویژگی‌های ناحیه‌های دانه‌بندی شدهٔ[10] تصویر را استخراج می‌نماید [14]. در مدل SFFF، ویژگی‌ها هم در حوزه مکانی و هم در حوزه فرکانسی استخراج شده و به دو صورت محلی و جامع با هم تجمیع می‌شوند [15]. در مدل LSDA افزونگی[11] در سطح ویژگی اجرا شده است و آموزش در سطح ویژگی‌ها در جهت‌های معنائی[12] انجام شده است که باعث کاهش مشکل از دست رفتن نواحی متمایز می‌شود. این فرآیند توسط یک شبکه پیش‌بینی ماتریس کوواریانس[13] انجام می‌شود [16]. سرانجام در مدل CAMF که ستون فقرات[14] آن ترنسفورمر است جهت استخراج اطلاعات نواحی حاوی اطلاعات مهم از ماژول توجه مکمل (CAM)[15] استفاده شده است. در این مدل یک بلوک تلفیق چند منظوره ویژگی[16] نیز برای ترکیب ویژگی های چند شاخه مختلف پیشبینی شده است [17].

---

[1] Semantically Enhanced Feature
[2] semantic group
[3] part
[4] longitudinal and transverse encoding module
[5] Attentive Pairwise Interaction
[6] diversification
[7] gradient-boosting
[8] Region Proposal Network
[9] progressive training strategy
[10] granularities
[11] Augmentation
[12] semantically directions
[13] Covariance Matrix Prediction Network
[14] backbone
[15] Complemental Attention Module
[16] Multi-Feature Fusion Module

همانگونه که در نمونه‌های ذکر شدهٔ بالا مشاهده می‌شود، بلوک‌های مکان‌یابی صریح و ضمنی در مدل‌های طبقه‌بندی ریزدانهٔ مبتنی بر CNN جهت استخراج ویژگی‌های تفکیک‌کننده، نقش موثری ایفاء می‌کنند. هر چند شبکه‌های عصبی عمیق بستر مناسبی در حوزهٔ بینائی ماشین جهت استخراج ویژگی و طراحی بلوک‌های اجرائی وظیفه‌های گوناگون نظیر بلوک مکان‌یابی و بلوک توجه محسوب می‌شوند، لیکن چالش‌های مهمی نیز دارند. موفقیت عملی یادگیری عمیق گاهی بیش از حد به کمیت و کیفیت داده‌های آموزشی موجود بستگی دارد [18]. در سناریوهایی که نمونه‌های آموزشی فراوان با کیفیت بالا در دسترس نیستند، مانند تصویربرداری پزشکی [19]، [20] و بازسازی سه بعدی تصاویر [21]، و یا زمانی که استخراج ویژگی‌های تفکیک‌کننده از نمونه‌های آموزشی مورد نیاز است مانند طبقه‌بندی ریزدانه [22]، عملکرد شبکه‌های عصبی عمیق ممکن است به‌طور قابل‌توجهی کاهش یابد. این شبکه‌ها با تعداد زیادی پارامتر مرتبط هستند و از تطبیق دامنه[۱] بهره نمی‌برند و حتی تحت افزونگی داده‌ها[۲] نیز دچار بیش برازش[۳] می‌شوند. از طرفی بطور کلی تفسیر و تحلیل نظری مدل های عمومی شبکه های عصبی عمیق دشوار است. فقدان تفسیر پذیری شبکه‌های عصبی عمیق در برخی کاربردها مانند پزشکی و هدایت خورکار [23]، محدودیت جدی در مقایسه با مدل-های تکرارپذیر الگوریتم‌محور مورد استفاده در پردازش سیگنال ایجاد می نماید.

به منظور ایجاد مدل‌های تفسیر پذیر و تعمیم پذیر عمیق یک راه حل منطقی این است که بتوان شبکه‌های عصبی عمیق را با معماری‌های مبتنی بر الگوریتم‌های تکرارپذیر ترکیب نمود [24]. یکی از رویکردهای عمده که در سالهای اخیر از آن استفاده می‌شود ترکیب کدگذاری تنک (SC)[۴] با شبکه‌های عصبی عمیق می‌باشد. تنک‌سازی[۵] از اواخر دههٔ ۸۰ میلادی در حوزه‌های یادگیری ماشین و شبکه‌های عصبی مطرح گردید [25]؛ لیکن در چند سال گذشته مبانی نظری مدل‌های توسعه یافته به ویژه مدل‌های مبتنی بر کدگذاری تنک کانولوشنال (CSC)[۶] در حوزه یادگیری ماشین معرفی شده‌اند [16]، [18]،[19] ،[23]، [24]، [25]، [26]، [27]، [28]، [29]، [36] و [37]. استفاده از هرس[۷] پارامترهای یادگیرنده و تعبیر به تنک‌سازی در این عمل [38]، [39] و [40]، بکارگیری اتوانکدر تنک در مدل‌های یادگیری [27] ،[41]، [28]، بکارگیری ساختارهای بازشدهٔ[۸] الگوریتم‌های تنک‌سازی و متناظر کردن اجزای آن با شبکه‌های عصبی عمیق[29]،[42] ، [43] و [44] و نظایر اینگونه کارهای تحقیقاتی بخشی از روش‌های ترکیب مدل‌های مبتنی بر شبکه‌های عصبی و مدل‌های یادگیری الگوریتم‌محور تنک به ویژه کدگذاری تنک محسوب می‌شوند.

---

۱ Domain adaptation
۲ Augmentations
۳ Overfit
۴ Sparse Coding
۵ Sparsity
۶ Convolutional Sparse Coding
۷ Prune
۸ Unrolled /Unfolded

در مثال‌های ذکرشدهٔ بالا نیز شاکلهٔ اصلی مدل و یادگیری پارامترها را معماری‌های مبتنی بر شبکه-های عصبی عهده‌دار هستند و الگوریتم‌های تنک‌سازی بسته به مأموریت مدل، جایگزین و یا اضافه می‌گردند.

در معماری‌های یادگیری ویژگی، هدف اصلی استخراج نمایش‌هایی است که علاوه بر حفظ اطلاعات کاربردی، افزونگی داده‌ها را حذف نمایند. در این راستا، مدل‌های CSC با بهره‌گیری از بهینه‌سازی لاسو، توانایی بالایی در استخراج ویژگی‌های محلی و جامع داده‌ها دارند [32], [43], [49] ,[48] ,[47] ,[46] ,[45] و[50]. با تکامل این رویکرد، ایده چینش سلسله‌مراتبی مدل‌های CSC مطرح گردید که منجر به معرفی معماری کدگذاری کانولوشنال چند لایهٔ تنک (ML-CSC)[1] شد [51], [52], [53], [54], [55]. این رویکرد تحلیلی، امکان توصیف ساختارهای ژرف را در قالب نمایش‌های تنک فراهم می‌آورد و به دلیل ماهیت ریاضی خود، مانع از افزونگی اطلاعات شده و ویژگی‌های تفکیک‌کننده را در خروجی مدل حفظ می‌نماید. بررسی‌های نظری در حوزه ML-CSC نشان می‌دهد که در صورت در نظر گرفتن نمایش‌های تنک لایه‌ها به‌عنوان متغیرهای یک مسئله بهینه‌سازی واحد، می‌توان پایداری و سازگاری نمایش‌ها را در کل شبکه تضمین نمود [51]. اگرچه این رویکرد جامع، تحلیل دقیق‌تری از خواص مدل ارائه می‌دهد، اما پیچیدگی محاسباتی بالای آن، پیاده‌سازی در مقیاس‌های بزرگ را با چالش مواجه می‌کند. در همین راستا، صورت‌بندی‌های دقیق‌تری برای بازیابی نمایش تنک در مدل‌های چندلایه معرفی شدند که ارتباط میان نمایش تنک و رفتار شبکه‌های عصبی ژرف را تبیین می‌کنند [52]. کاربردهای عملی این معماری در مدل‌های پیشرفته‌ای نظیر U-Net مشاهده شده است که در آن‌ها از ضرایب تنک به عنوان نقشه‌های ویژگی برای بهبود کارایی خوشه‌بندی تصاویر استفاده شده است [56] و [26]. با این حال، اکثر پژوهش‌های پیشین در حوزه ML-CSC یا بر تفسیر تک‌لایه متمرکز بوده‌اند یا در مراحل آموزش لغت‌نامه با محدودیت‌های معماری مواجه بوده‌اند. در این راستا، نویسندگان در پژوهشی پیشین یک چارچوب جامع برای آموزش لغت‌نامه‌های چندلایه و پیاده‌سازی کلاس مدل‌های ML-CSC ارائه دادند که مبانی تحلیلی و ریاضی آن در آن مرجع به‌طور کامل تبیین شده است [57].

در پژوهش حاضر، هدف ما بهره‌گیری از این زیرساخت اثبات‌شده و ترکیب آن با شبکه‌های عصبی عمیق کانولوشنال CNN است. ما در این تحقیق جهت اجرای طبقه‌بندی ریزدانه ابتدا یک معماری مبتنی بر کدگذاری کانولوشنال چند لایهٔ تنک (ML-CSC)[2] و آموزش لغت‌نامه را با هویت مستقل و خالص معرفی کرده‌ایم که از لحاظ تحلیلی و نظری بطور کامل با شبکه‌های عصبی چند لایه متفاوت بوده و در عین حال همهٔ قابلیت‌های شبکه‌های عصبی را دارا می‌باشد. در گام بعدی این مدل مستقل

---

[1] Multi-Layer Convolutional Sparse Coding

[2] Multi-Layer Convolutional Sparse Coding

که آن را شبکه کدگذاری تنک (CSN)[۱] نام‌گذاری کرده‌ایم، در کنار یک مدل دیگر از نوع شبکه‌های عصبی CNN یک معماری دوشاخه‌ای را جهت طبقه‌بندی ریزدانه تشکیل خواهند داد.
مشارکت های ما به شرح زیر خلاصه می‌شوند:

- ما با استفاده از لغت‌نامه، یک شبکهٔ کدینگ کانولوشنال تنک چند لایه معرفی کرده ایم که در مسیر پیش سو دقیقاً مشابه یک شبکه CNN عمل می‌کند و آموزش فیلترهای آن (اتم-های) نیز در مسیر پیش‌سو انجام می‌شود.
- تفاوت مدل پیشنهادی ما با یک شبکه CNN در این است که مدل ما در طول مسیر پیش‌سو افزونگی داده را مهار نموده و اطلاعات جامع و محلی را توأمان استخراج می‌نماید. به همین دلیل قادر به استخراج ویژگی‌های تفکیک‌پذیرتری نسبت به مدل‌های CNN می‌باشد.
- مدل پیشنهادی ما از الگوریتم های تکرارپذیر استفاده مینماید و دارای تفسیرپذیری بالاتری نسبت به مدل‌های شبکه‌های عصبی عمیق می‌باشد. با این حال الگوریتم‌های استفاده شده فقط با چند تکرار در هر لایه اجرا شده و هیچ ابر[۲] پارامتر اضافه‌ای را به شبکه تحمیل نمی‌کنند.
- کارائی مدل پیشنهادی ما در طبقه‌بندی ریزدانه که یک وظیفه چالشی در حوزه بینائی ماشین محسوب می‌شود، نشان داده خواهد شد؛ بطوریکه با بکارگیری این مدل در یک شبکه عصبی عمیق دقت طبقه‌بندی افزایش می‌یابد.

مابقی سازماندهی مقاله به این صورت است که در بخش دوم روابط تحلیلی مربوطه بررسی و بر مبنای آن مدل پیشنهادی ارائه خواهد شد. بخش سوم به آزمایشات شبیه‌سازی شه و نتایج آن اختصاص دارد. در بخش چهارم نتایج تحلیل شده و در خصوص جزئیات نتایج بحث خواهد شد و در بخش پنجم جمع‌بندی و نتیجه‌گیری این تحقیق ارائه خواهد شد.

۲– شبکهٔ کانولوشنال تنک (CSN)

سیگنال $\boldsymbol{s} \in \mathbb{R}^N$ و لغت‌نامهٔ تماماً کامل $(N \ll M)$ $\boldsymbol{D} \in \mathbb{R}^{N\times M}$ را در نظر بگیرید. در این صورت بازنمائی تنک سیگنال $\boldsymbol{s}$ به صورت $\boldsymbol{s} \simeq \boldsymbol{D}\boldsymbol{x}$ قابل بیان است با این شرط که بردار $\boldsymbol{x}$ دارای کمترین تعداد عناصر غیر صفر باشد. بیان فوق با استفاده از جملهٔ جریمهٔ ℓ_1 می‌تواند بصورت مسئله بهینه سازی BP یا رگراسیون لاسو نوشته شود [58] ,[50]:

$$\hat{x} = \underset{x}{argmin}\frac{1}{2}\|s - Dx\|_2 + \lambda\|x\|_1 \qquad (۱)$$

---

[۱] Convolutional Sparse Network

[۲] Hyper

ضریب رگولارایزر $\lambda > 0$ میزان تنک‌سازی رابطه بالا را کنترل می‌کند. شکل کانولوشنال رابطهٔ رگراسیون لاسو یا همان CBS [1] را می‌توان به صورت زیر نوشت :

$$\hat{\boldsymbol{x}} = \underset{\{\boldsymbol{x}_m\}}{argmin} \frac{1}{2} \sum_m \|\boldsymbol{d}_m * \boldsymbol{x}_m - \boldsymbol{s}\|_2^2 + \lambda \sum_m \|\boldsymbol{x}_m\|_1 \qquad (۲)$$

در رابطهٔ بالا $\{\boldsymbol{d}_m\}$ مجموعه ای از M فیلتر (اتم) لغت‌نامهٔ آموزش یافته، نماد * کانولوشن $\{\boldsymbol{x}_m\}$، مجموعه ضرایب تنک یا ویژگی‌ها و $\boldsymbol{s}$ سیگنال (تصویر) ورودی می باشد. برای حل رابطه بالا روش‌های مختلفی در مراجع مختلف ارائه شده است که از جمله میتوان به [59] ,[49] ,[34] اشاره نمود. با فرض اینکه لغت‌نامهٔ $\boldsymbol{D}$ یک ماتریس کانولوشنال به شکل $\boldsymbol{D} = (\boldsymbol{d}_0 \ \boldsymbol{d}_1 \ \dots \boldsymbol{d}_m)$ بوده و با تعریف $\boldsymbol{X} = (\boldsymbol{x}_0 \ \boldsymbol{x}_1 \ \dots \boldsymbol{x}_m)^T$ ، رابطه (۲) را می‌توان به شکل استاندار رابطه BP نوشت:

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{X}}{argmin} \frac{1}{2} \|\boldsymbol{D}\boldsymbol{X} - \boldsymbol{S}\|_2^2 + \lambda \|\boldsymbol{X}\|_1$$

همانگونه که اشاره گردید با سری نمودن چندین رابطه CBS، یک مدل چند لایه ML-CSC ایجاد می‌شود که می‌تواند مسیر پیشسو شبکه عصبی عمیق کانولوشنال CNN را بازسازی نماید. این شبکه با تخمین بازنمائی های تنک هر لایه بصورت $\hat{\boldsymbol{s}} = \boldsymbol{D}_1\boldsymbol{X}_1$ و $\hat{\boldsymbol{X}}_1 = \boldsymbol{D}_2\boldsymbol{X}_2$ و . . . $\hat{\boldsymbol{X}}_{L-1} = \boldsymbol{D}_L\boldsymbol{X}_L$ و با اعمال تنک‌سازی لایه‌های میانی (جمله $\|.\|_1$) قابل تحقق است. ما با الهام گرفتن از این ظرفیت ساختارهای کانولوشن یادگیری لغت‌نامه و مفاهیم مطرح شده در منابعی همچون [55] ,[54] ,[53] ,[51]، مدل CSN پیشنهادی را معرفی نموده‌ایم [57].

۲-۱ ساختار پیشنهادی CSN

اصلی‌ترین بخش مدل CSN، معماری لایهٔ کدگذاری کانولوشنال تنک (CSC) و یادگیری لغت‌نامه (DL) [2] است که نقش کلیدی شبکه پیشنهادی را بر عهده دارد و ما آنرا لایه CSC-DL نامیده‌ایم. ویژگی مهم این لایه که آن را از سایر معماری‌های ML-CSC متمایز می‌سازد این است که در مدل پیشنهادی بهینه‌سازی تابع هدف هم نسبت به نقشه‌های ویژگی و هم نسبت به لغت‌نامه‌ها بصورت توأمان اجرا می شود. **شکل ۱** دو لایهٔ متوالی از مدل پیشنهادی را نمایش می‌دهد. در این شکل عملیات کانولوشن فیلترهای لغت‌نامه (در شکل یک فیلتر به عنوان نماینده در نظر گرفته شده است) با ویژگی‌های (کدهای تنک) همان لایه انجام می‌شود و در هر تکرار نتیجهٔ کانولوشن این دو باید با ویژگی‌های لایهٔ قبل مقایسه و کمینه شود.

---

[1] Convolutional Basis Pursuit

[2] Dictionary Learning

تابع هدفی که در لایه L ام CSC-DL بایستی بهینه‌سازی شود، رابطه CBS است. فرض کنید که تعداد n تصویر به صورت $[\boldsymbol{s_1}, \boldsymbol{s_2}, \ldots, \boldsymbol{s_n}]$ وجود داشته باشد بطوریکه $\boldsymbol{s}_n \epsilon \mathbb{R}^{N\times M\times C}$ تصویر n ام ورودی به لایه L ام با C کانال باشد. همچنین $[\boldsymbol{d_1}, \boldsymbol{d_2}, \ldots, \boldsymbol{d_m}]$ و $[\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}]$ نیز به ترتیب اتم های لغتنامه و ضرایب تنک این لایه می‌باشند که در آنها $\boldsymbol{d}_m \epsilon \mathbb{R}^{p\times p\times C}$ اتم m ام لغت‌نامهٔ لایه و $\boldsymbol{x}_n \epsilon \mathbb{R}^{N\times M\times m}$ ویژگی تنک n ام باشد. N و M طول و عرض تصویر یا ویژگی های تنک بوده و p نیز طول و عرض اتم‌های لغت‌نامه (سایز لغت‌نامه) می‌باشد. در این صورت تابع هدف CBS لایه برای تصویر n ام بصورت زیر قابل بیان است:

$$\underset{\{\boldsymbol{x_{m.n}}\},\{\boldsymbol{d_m}\}}{arg\,min} \frac{1}{2}\sum_m \left\|\boldsymbol{d}_m * \boldsymbol{x}_{m,n} - \boldsymbol{s}_n\right\|_2^2 + \lambda \sum_m \|\boldsymbol{x}_{m.n}\|_1 \qquad (۳)$$

ویژگی های تنک تولید شده به عنوان ورودی در لایه $L+1$ عمل خواهند کرد و تعداد کانال‌های تصویر ورودی n ام در لایه $L+1$ برابر تعداد اتم‌های لغت‌نامه در لایه L ام خواهند بود. رابطهٔ بالا همزمان نسبت به $\{\boldsymbol{d}_m\}$ و $\{\boldsymbol{x}_{m.n}\}$ محدب و هموار نیست. لذا در روش پیشنهادی از الگوریتم ADMM [59] [۱] استفاده شده تا تابع هدف بطور متناوب یک بار با محاسبه $\{\boldsymbol{d}_m\}$ در رابطهٔ DL و استفادهٔ آن در مسئلهٔ CSC نسبت به $\{\boldsymbol{x}_{m.n}\}$ کمینه شود و بار دیگر با محاسبهٔ $\{\boldsymbol{x}_{m.n}\}$ در رابطهٔ CSC و استفادهٔ آن در مسئلهٔ DL نسبت به $\{\boldsymbol{d}_m\}$ کمینه گردد. بدین ترتیب در هر بار به‌هنگام-سازی، سه زیر معادله برای مرحلهٔ DL و سه زیر معادله برای مرحلهٔ CSC حل می‌شود. بنابراین در هر تکرار الگوریتم ADMM، شش زیرمعادله به ترتیب حل می‌شوند تا هم فیلترهای لغت‌نامه‌ها آموزش داده شوند و هم کدهای تنک به عنوان نقشه‌های ویژگی تولید و به لایه بعدی ارسال گردند [57]. این فرآیند تحت عنوان الگوریتم ADMM-CSC-DL در الگوریتم ۱ نشان داده شده است.

[۱] Alternating Direction Method of Multipliers

الگوریتم ۱: ADMM-CSC-DL

Require: Signals $[s_1, s_2, \ldots, s_n]$, initial layer dictionary atoms $[d_{0,1}, d_{0,2}, \ldots, d_{0,m}]$ in C_{PN}
, layer regularization parameter λ

1: ***for*** $l = 1, 2, \ldots, L$ ***do***
2: Initialize $\{d_m\} \leftarrow \{d_{0,m}\}$;
3: ***for*** $i = 1, 2, \ldots, n$ ***do***
4: ***for*** $j = 1, 2, \ldots, m$ ***do***
5: Update sparse codes using current dictionary$\{x_{m,n}\}$:

$$\{x_{m,n}\}^{new} \leftarrow \underset{\{x_{m,n}\}}{arg\ min} \frac{1}{2}\sum_m \|d_m * x_{m,n} - s_n\|_2^2 + \lambda \sum_m \|x_{m,n}\|_1 \quad (4)$$

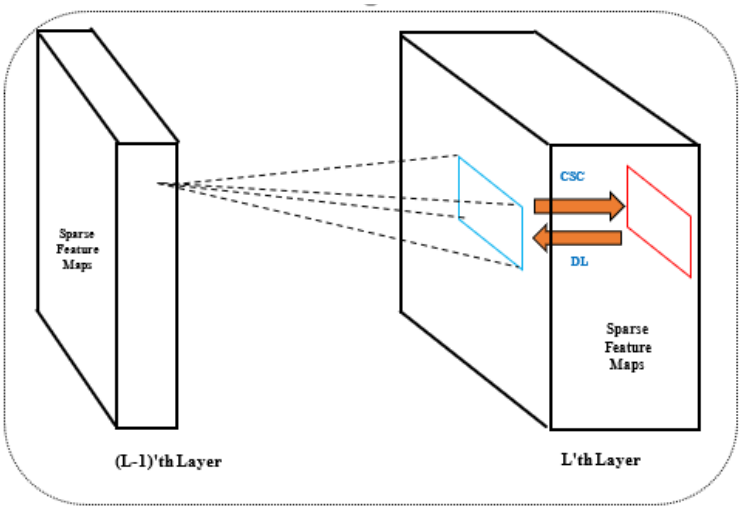
6: Update dictionary atoms using new sparse codes$\{d_m\}$:

$$\{d_m\}^{new} \leftarrow argmin_{\{d_m\} \in C_{PN}} \frac{1}{2} \|\sum_m d_m * x_{m,n} - s_n\|_2^2 \quad (5)$$

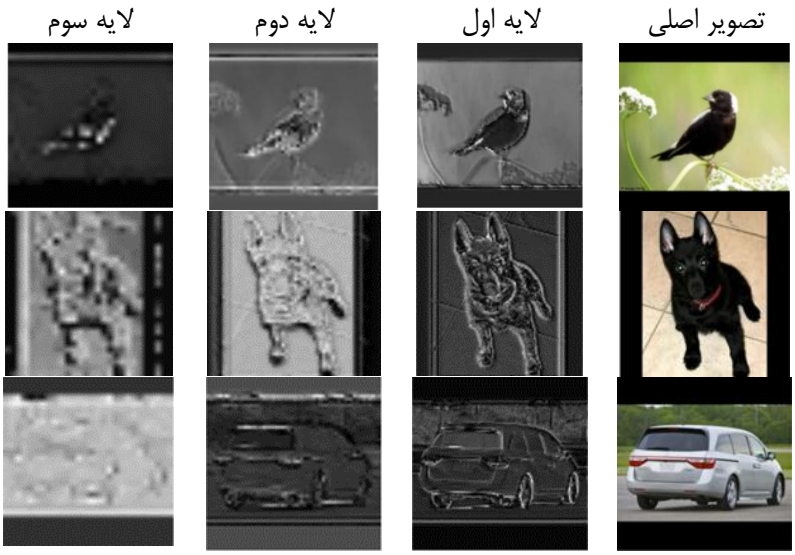
7: ***end for***
8: ***end for***
9: ***end for***
10: ***return*** the dictionary atoms $\{d_m\}$ in C_{PN} and sparse codes $\{x_{m,n}\}$

برای درک بهتر عملکرد شبکه CSN، یک مدل سه لایه پیاده سازی شده و بازنمائی لایهٔ اول، دوم و سوم شبکه یاد شده در شکل ۲ نمایش داده شده است. همانگونه که دیده می شود علی‌رغم اینکه بازنمائی لایه دوم تنک‌تر از لایه اول است ولی ناحیه برجستهٔ[۱] تصویر در لایه سوم باقی مانده است. همچنین به وضوح دیده می‌شود در لایهٔ اول ویژگی‌های پایه مانند لبه‌ها کاملاً برجسته گردیده و با عبور ویژگی‌ها به سمت لایهٔ سوم، علی‌رغم انتزاعی‌تر شدن ویژگی‌ها، نقاط حاوی اطلاعات مهم نمایان شده است (مرز نقاط روشن به تیره و بلعکس). این قابلیت‌ها از نقاط قوت CSN محسوب می‌شود و در بخش بعدی ما از این قابلیت‌ها استفاده کرده‌ایم تا ویژگی‌های تفکیک‌کننده را بدست آورده و آنها را برای افزایش دقت طبقه بندی FGVC به کار بگیریم.

[۱] salient



شکل ۱: دولایهٔ متوالی از شبکهٔ CSN



شکل ۲ بازنمائی تنک تصاویر در سه لایهٔ متوالی شبکهCSN

۲–۲ اتوانکدر مبتنی بر شبکهٔ کانولوشنال تنک (CSN-AE)

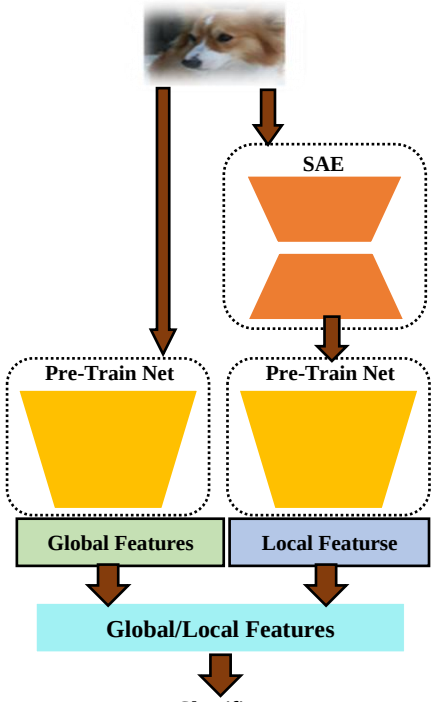
در این بخش معماری پیشنهادی CSN جهت افزایش دقت طبقه‌بندی دیداری ریزدانه (FGVC) مورد استفاده قرار گرفته است. به منظور تحقق عملی یک معماری ترکیبی ریزدانه، ما با استفاده از بلوک‌های ADMM-CSC-DL نشان داده شده در شکل ۱ یک اتوانکدر (AE)[۱] تنک (SAE)[۲]

---

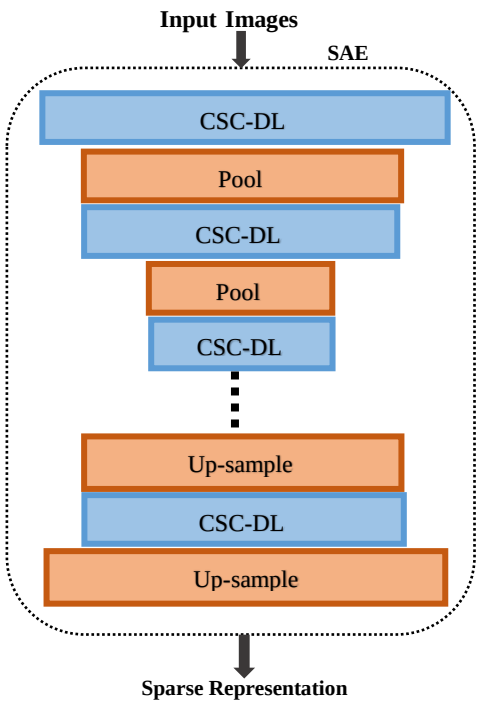
[۱] Auto-Encoder

[۲] Sparse-AE

پیاده‌سازی کردیم که نمای کلی آن در شکل ۳ نشان داده شده است. ویژگی مهم SAE این است که در حالت بدون نظارت در لایه‌های CSC-DL انکدر، بازنمائی تنکی از تصاویر تولید شده و افزونگی اطلاعات از بین می‌رود تا دیکدر بتواند تصاویر را بصورت تنک بازسازی نماید. در نتیجه اطلاعات مهم تصاویر ورودی باقی مانده و در مرحله بعدی ویژگی‌های تفکیک‌پذیر صرفاً از همان اطلاعات مهم باقی مانده توسط شبکه از پیش‌آموزش داده شده (ستون فقرات) استخراج می‌شود. هر چند می‌توان SAE یا هر مدل مبتنی بر CSN را با سایر معماری‌های کارامد رایج مانند ترانسفورمرها ترکیب نموده و مدل‌های ترکیبی موثری معرفی نمود، لیکن در این تحقیق سعی گردیده است با اجتناب از پیچیدگی‌های اضافی صرفاً کارائی مدل پیشنهادی در مأموریت دسته‌بندی ریزدانه نشان داده شود. بنابراین ما در کنار SAE از یک شبکهٔ استاندارد مبتنی بر CNN که از قبل بر روی ImageNet [60] آموزش داده شده است، استفاده کرده و یک مدل ترکیبی دسته‌بندی ریزدانه را معرفی نموده‌ایم که در شکل ٤ نشان داده شده است.



شکل ٤ مدل ترکیبی SAE مبتنی بر معماری پیشنهادی CSN



شکل ۳ یک نمونه از اتوانکدر تنک مبتنی بر معماری CSN

در این مدل ترکیبی، تصاویر ورودی وارد دو شاخهٔ جداگانه می‌شوند و در انتهای هر شاخه ویژگی‌های استخراج شده با یکدیگر تجمیع گردیده و سپس وارد طبقه‌بند می‌شوند. در شاخهٔ حاوی SAE، فرآیند

استخراج ویژگی در دو مرحله اجرا می‌شود. درمرحله نخست بوسیله SAE بازنمائی تنکی از تصاویر ورودی بازسازی می‌شود. به منظور درک شهودی بهتر، نمونه‌ای از تصاویر اصلی، تصاویر بازسازی شدهٔ تنک در خروجی SAE و فعال‌سازهای[1] این خروجی در **شکل ۵** نشان داده شده است. همانگونه در ستون فعال‌سازها دیده می‌شود، عبور تصاویر اصلی از SAE موجب شده است که نواحی برجسته و حاوی اطلاعات در حد زیادی تفکیک نمایان شود که این موضوع کمک شایانی به مرحلهٔ استخراج ویژگی‌های تفکیک‌کننده توسط شبکه ستون فقرات خواهد نمود.

شکل ۵: بازنمائی تصاویر تنک در مدل SAE. به ترتیب از چپ به راست: تصاویر طبیعی، فعال‌ساز ویژگی‌های تنک در خروجی انکدر، تصویر بازسازی شده در دیکدر و فعال‌ساز خروجی دیکدر

در مرحله دوم این تصاویر بازسازی شدهٔ تنک به شبکه استخراج ویژگی ستون فقرات محول می‌شود تا ویژگی‌های محلی و تفکیک‌کننده استخراج گردد. در شاخه دوم صرفاً توسط شبکه ستون فقرات، ویژگی‌های عمیق و جامع استخراج می‌شوند. سرانجام ویژگی‌های استخراج شده از دو شاخه با یکدیگر تجمیع شده و بردار ویژگی نهائی را تشکیل می‌دهند. این بردار بطور توأمان شامل اطلاعات جامع و محلی بوده و جهت طبقه‌بندی ریزدانه مورد استفاده قرار می‌گیرد.

---

[1] Activations

۳-آزمایشات

در این بخش ابتدا بطور مختصر مجموعه‌داده‌هائی که در این تحقیق بمنظور دسته‌بندی (طبقه‌بندی) ریزدانهٔ تصاویر استفاده شده است، معرفی می‌کنیم. ما از سه مجموعه داده ریزدانه استفاده نموده‌ایم که در سال‌های اخیر مقالات متعددی نیز بر روی آنها دسته‌بندی ریزدانه را اجرا کرده‌اند. این مجموعه‌ها عبارتند از: CUB-200-2011 [1]، Stanford Dogs [3] و Stanford Cars [4] که مشخصات آنها در **جدول ۱** درج گردیده است. باید یادآوری کنیم مجموعه داده‌های مذکور دارای برچسب و حاشیه‌نویسی در سطح بخش نیز هستند که در آزمایشات انجام گرفته از آنها استفاده نشده است. ما صرفاً برچسب‌های سطح تصویر را در آزمایشات بکار گرفته‌ایم.

جدول ۱ مشخصات مجموعه داده‌های ریزدانه

Dataset	Target	# Class	# Train	# Test
CUB-200-2011	Bird	200	5994	5794
Stanford Dogs	Dog	120	12000	8580
Stanford-Cars	Cars	196	8144	8041

در آزمایشات این بخش شبکهٔ عصبی CNN از قبل آموزش داده شده که به عنوان ستون فقرات استخراج ویژگی از آن استفاده شده است، شبکهٔ EfficientNet می‌باشد که در سال ۲۰۱۹ در هشت نسخه معرفی شده است(نسخه‌های B0 تا B7) [61]. از نظر ساختاری، معماری EfficientNet-B0 بر پایهٔ بلوک‌های کانولوشن گلوگاه معکوس[۱] متحرک بنا شده است که با بهره‌گیری از کانولوشن‌های تفکیکی عمقی، هزینه‌های محاسباتی را به‌شدت کاهش می‌دهد. همچنین این مدل از استراتژی مقیاس‌بندی ترکیبی[۲] پیروی می‌کند که در آن عمق، عرض و رزولوشن شبکه به‌صورت یکنواخت با استفاده از یک ضریب مشخص مقیاس‌بندی می‌شوند تا تعادلی بهینه میان پیچیدگی مدل و کارایی استخراج ویژگی ایجاد گردد. در آزمایشات این مقاله از EfficientNet-B0 استفاده کرده‌ایم که مهم‌ترین ویژگی آن تعداد کم پارامترهای این شبکه است که تأثیر عمده‌ای در کاهش زمان اجرای محاسبات دارد. همچنین این شبکه دارای دقت بالاتری نسبت به شبکه‌های هم‌گروه خود یعنی ResNet-50 و DenseNet-169 در دسته‌بندی تصاویر ImageNet دارد (۷۷/۱**%** در مقابل ۷۶**%** و ۷۶/۲**%**). ویژگی‌های استخراج شده در لایهٔ آخر کانولوشنال و قبل از لایه‌های تماماً متصل دارای ابعاد ۷×۷×۱۲۸۰ می‌باشند که پس از عبور از G-AvP[۳] بصورت بردار ویژگی با ابعاد ۱×۱۲۸۰ به دسته‌بند سافت‌مکس[۴] ارسال می‌شوند.

---

[۱] MBConv

[۲] Compound Scaling

[۳] Global Average Pooling

[۴] SoftMax

همانگونه که گفته شد و در **شکل ٤** نشان داده شده است، نمونه‌های آموزشی و تست جهت استخراج ویژگی به دو صورت از شبکهٔ ستون فقرات عبور می‌کنند. این نمونه‌ها به دو نسخهٔ یکسان تکثیر شده و یکی از آنها ابتدا وارد مدل SAE شده و به صورت تنک بازسازی می‌شود و سپس از شبکهٔ ستون فقرات EfficientNet-B0 عبور کرده و ویژگی‌های آن استخراج می‌گردد. نسخهٔ دوم بطور مستقیم وارد شبکهٔ EfficientNet-B0 شده و از این مسیر نیز ویژگی‌های نمونه‌های ورودی استخراج می‌شود. در نهایت این دو گروه ویژگی به یکدیگر مزید شده و یک بردار ویژگی با ابعاد ۱×۲۵۶۰ تشکیل می‌دهند که جهت دسته‌بندی مورد استفاده قرار می‌گیرد.

شبکهٔ EfficientNet-B0 علاوه بر اینکه نقش ستون فقرات مدل ترکیبی استخراج ویژگی را بر عهده دارد، به عنوان یک مدل پایه، مبنای مقایسه با مدل ترکیبی نیز محسوب می‌شود. یعنی دقت دسته‌بندی شبکهٔ EfficientNet-B0 به تنهائی و بدون ترکیب با مدل SAE بر روی مجموعه‌داده‌های ریزدانه با دقت دسته‌بندی مدل ترکیبی این دو مقایسه خواهد شد.

۳-۱ نتایج آزمایشات

همانگونه که در **جدول ۱** مشاهده می‌شود، تعداد نمونه‌های مجموعه‌داده‌های ریزدانه اغلب کمتر از نمونه‌های مجموعه‌های رایج کلاسیک می‌باشد. به همین دلیل برای استخراج ویژگی از این تصاویر ابتدا از روش‌های بسط نمونه‌ها[۱] در مرحلهٔ پیش‌پردازش استفاده می‌نمایند. همچنین هنگام بکار گیری شبکه‌های از قبل آموزش داده شده از فرآیند انتقال یادگیری[۲] و تنظیم نهائی(FT)[۳] مدل نیز استفاده می‌شود. بمنظور مقایسهٔ عملکرد معماری پیشنهادی علاوه بر مدل EfficientNet-B0 که به عنوان ستون فقرات مدل پیشنهادی بکار رفته است، از این شبکه و شبکهٔ ResNet-50 نیز به عنوان مدل پایه برای مقایسه استفاده شده است. ما مدل‌هائی را که دارای تنظیمات نهائی هستند با FT مشخص نموده‌ایم. نتایج نهائی آزمایشات دسته‌بندی ریزدانه در **جدول ۲** نشان داده شده است.

در **جدول ۲** علاوه بر شبکه‌های از قبل آموزش داده شده، نتایج چندین مدل دیگر نیز که در سال‌های اخیر برای دسته‌بندی ریزدانه مورد استفاده قرار گرفته است درج شده است (که در بخش‌های اول بطور مختصر دربارهٔ آنها توضیح داده شده است). مدل SFFF برای کسب ویژگی‌های تفکیک‌کننده، عملیات استخراج ویژگی را هم در حوزهٔ مکان و هم در حوزهٔ فرکانس انجام می‌دهد و سپس بردار ویژگی نهائی را از تجمیع این دو فرآیند تشکیل می‌دهد [15]. در این مدل از شبکهٔ ResNet-50 بمنظور ستون فقرات استخراج ویژگی استفاده شده است. مدل LSDA نیز به جای افزونگی نمونه‌ها در سطح تصویر، نوعی افزونگی نمونه‌ها در سطح ویژگی را معرفی نموده است [16]. این مدل هر دو شبکهٔ ResNet-**50** و EfficientNet-B0 را به عنوان ستون فقرات مورد ارزیابی قرار داده است. با دقت در **جدول ۲** می‌توان عملکرد مطلوب مدل پیشنهادی مبتنی بر معماری CSN را به وضوح مشاهده نمود. مدل

---

[۱] augmentation

[۲] transfer learning

[۳] fine-tune

پیشنهادی SAE علاوه بر اینکه عملکرد مطلوب‌تری از مدل‌های پایهٔ شبکه‌های عصبی از قبل آموزش داده شده با تعداد پارامترهای زیاد دارد، از برخی مدل‌های ترکیبی پیچیده‌تر نیز بر روی بعضی مجموعه‌داده‌های ریزدانه نیز بهتر عمل کرده است. این موضوع به ویژه در مجموعه دادهٔ CUB-**200-2011** کاملاً مشهود است. برخی از مدل‌های جدول مذکور دارای پارامترهای خیلی بیشتری از مدل پیشنهادی هستند، با این وجود نتایج مدل پیشنهادی بهتر از آنهاست.

جدول ۲ مقایسه دقت دسته‌بندی مدل پیشنهادی با سایر مدل‌های ریزدانه

مدل‌ها	سال	ستون فقرات	پارامترها (M)	CUB (%)	Dogs (%)	Cars (%)
ResNet50 [62]	۲۰۱۸	ResNet50	۲۵/۶	۷۵/۱۵	۶۹/۹۲	۹۱/۵۲
FT- ResNet50 [9]	۲۰۲۰	ResNet50	۲۵/۶	۸۴/۵	۸۸/۱	۹۱/۷
RSB-ResNet [63]	۲۰۲۱	ResNet50	۲۵/۶	-	-	۹۲/۷
DFL-CNN [64]	۲۰۱۸	ResNet50	۲۵/۶	۸۷/۴	-	۹۳/۱
WSDL[65]	۲۰۱۹	VGGNET	۱۳۸/۴	۸۵/۷۱	-	۹۲/۳
SEF[9]	۲۰۲۰	ResNet50	۲۵/۶	۸۷/۳	۸۸/۸	۹۴/۰
AKEN[10]	۲۰۲۱	VGGNET	۱۴۳/۷	۸۷/۱	-	۹۳/۹
API-NET [11]	۲۰۲۰	ResNet50	۲۵/۶	۸۷/۷	۸۸/۳	۹۴/۸
DB-GBL [12]	۲۰۲۰	ResNet50	۲۳/۹	۸۷/۷	۸۷/۱	۹۴/۳
FDL [13]	۲۰۲۰	DenseNet161	۱۴/۳	۸۹/۰۹	۸۴/۸۶	۹۴/۰۲
PMG [14]	۲۰۲۲	ResNet50	۴۷/۴	۸۹/۹	۸۹/۱	۹۵/۴
SFFF [15]	۲۰۲۳	ResNet50	۴۷/۴	۸۵/۴	۸۸/۰	۹۴/۴
CSQA-Net [66]	۲۰۲۴	ResNet50	۴۷/۴	۹۰/۵	-	۹۵/۶
LSDA [16]	۲۰۲۴	ResNet50	۲۵/۶	۸۶/۷	-	۹۴/۳
LSDA [16]	۲۰۲۴	EfficientNetB0	۵	۸۶/۳	-	-
CAMF [17]	۲۰۲۱	Swim-Base	۹۳	۸۹/۷	۹۱/۸	۹۴/۸
FT-EfficientNetB0	-	EfficientNetB0	۴/۷	۸۳/۰	۸۴/۵	۹۰/۴۸
CSN-AE (Ours)	-	EfficientNetB0	۹/۲	۹۱/۰	۸۶/۷	۹۱/۷۱

هر چند مدل‌های ترکیبی پیچیده‌ای در جدول بالا برای دسته‌بندی ریز دانه وجود دارد که علاوه بر سطح تصویر در سطح بخش نیز مکان‌یابی کرده و در مقیاس‌های مختلف و با ایده‌های پیچیده، قابلیت رسیدن به دقت‌های بالاتری را دارند، لیکن هدف ما در این بخش نشان‌دادن عملکرد مطلوب معماری پیشنهادی بوده و قصد مقایسه مدل ساده SAE با مدل‌های پیچیدهٔ معرفی شده در مقالات را نداریم. آنچه که اهمیت ویژه‌ای دارد این است که معماری پیشنهادی مبتنی بر شبکهٔ CSN یک ساختار پایه کانولوشنال ممزوج با تنک‌سازی بوده و قابلیت ترکیب و بکارگیری در هر مدل مبتنی بر CNN را جهت افزایش دقت دسته‌بندی ریزدانه دارد.

۳-۱ مطالعهٔ فرسایشی (مرحله‌ای)

در این بخش آزمایش مرحله‌ای را بر روی شاخه ستون فقرات و شاخه شامل CSN-AE در نظر می‌گیریم که در آن اثر شاخهٔ حاوی CSN-AE را در دقت کل مدل بررسی کنیم. در محاسبه دقت ها علاوه بر دقت پیشبینی برتر[۱] (رتبهٔ اول)، دقت پنج پیشبینی برتر[۲] (پنج رتبهٔ اول) نیز مورد نظر هستند. نتایج ارزیابی‌ها روی هر سه مجموعه داده در جدول ۳ تا جدول ۵ نشان داده شده است. در همه این آزمایشات افزونگی[۳] استاندارد روی شبکه EffecientNetB0 انجام شده است و به جز مجموعه داده Dogs شبکه ستون فقرات تنظیم نهائی[۴] نیز شده است.

جدول ۳مقایسه دقت های طبقه بندی روی مجموعه داده CUB200

مدل‌ها	افزونگی	دقت (%)	
		رتبه اول	۵رتبه اول
FT_EfficientNet_B0	-	۸۶/۵۰±۰/۱	۹۵/۴۶±۰/۱۸
FT_EfficientNet_B0	√	۸۷/۵۰±۰/۱۳	۹۵/۸۷±۰/۰۶
FT_EfficientNet_B0 + CSN_AE(Ours)	-	۸۹/۶۰±۰/۴۵	۹۵/۶۳±۰/۶۱
FT_EfficientNet_B0 + CSN_AE(Ours)	√	۹۱/۰۰±۰/۳۴	۹۷/۵۰±۰/۱

جدول ٤مقایسه دقت های طبقه بندی روی مجموعه داده Stanford Dogs

مدل‌ها	افزونگی	دقت (%)	
		رتبه اول	۵رتبه اول
FT_EfficientNet_B0	-	۸۳/۰۳±۰/۱۳	۹۷/۱۰±۰/۲۷
FT_EfficientNet_B0	√	۸۴/۲۰±۰/۲۵	۹۷/۵۷±۰/۴۶
FT_EfficientNet_B0 + CSN_AE(Ours)	-	۸۶/۶۰±۰/۴۵	۹۸/۶۹±۰/۲۲
FT_EfficientNet_B0 + CSN_AE(Ours)	√	۸۶/۷۰±۰/۶۶	۹۸/۸۷±۰/۱۲

جدول ۵ مقایسه دقت های طبقه بندی روی مجموعه داده Stanford Cars

مدل‌ها	افزونگی	دقت (%)	
		رتبه اول	۵رتبه اول
FT_EfficientNet_B0	-	۸۵/۵۷±۰/۲۱	۹۳/۷۸±۰/۳۷
FT_EfficientNet_B0	√	۹۰/۴۸±۰/۲۲	۹۸/۳۸±۰/۱۳
FT_EfficientNet_B0 + CSN_AE(Ours)	-	۸۷/۳۷±۰/۵۳	۹۴/۶۴±۰/۵۹
FT_EfficientNet_B0 + CSN_AE(Ours)	√	۹۱/۷۱±۰/۸۱	۹۸/۷۲±۰/۰۷

همانگونه که دیده می شود افزونگی استاندارد شبکه ستون فقرات (EffecientNetB0) دقت رتبهٔ اول طبقه‌بندی را به نحو چشمگیری در مجموعه داده Cars افزایش می دهد(به میزان **۴/۹۱%**).

---

[۱] Top-1 accuracy
[۲] Top-5 accuracy
[۳] Augmentation
[۴] Fine-tune

این تأثیر در مجموعه داده‌های CUB و Dogs به ترتیب ۱/۵% و ۱% می‌باشد. از طرفی حتی بدون افزونگی نیز مزید کردن شاخه CSN-AE باعث افزایش دقت طبقه‌بندی می‌شود به طوریکه دقت رتبهٔ اول طبقه‌بندی در مجموعه داده‌های CUB200 ، Dogs و Cars به ترتیب ۳/۱% ، ۳/۵۷% و ۱/۸% بهبود یافته است. در حالتی که افزونگی هم در شبکه ستون فقرات و هم در CSN-AE انجام می‌شود، نتیجه مجدداً بهبود پیدا می‌کند. به عنوان مثال روی مجموعه دادهٔ Cars دقت ۴/۳۴% بهبود پیدا میکند (۹۱/۷۱ %در مقابل ۸۷/۳۷%). نکته مهم دیگر این است که در همه مجموعه داده‌ها ترکیب شاخه CSN-AE همواره موجب افزایش دقت‌ها گردیده است. لذا می‌توان نتیجه گرفت که با اضافه کردن شاخه CSN-AE به سایر مدل های ریزدانه می‌توان دقت طبقه‌بندی را افزایش داد.

۳-۲ بهبود نرخ خطا

در جدول ٦ نرخ بهبود دقت و خطای رتبهٔ اول و پنج رتبهٔ اول مدل پیشنهادی نسبت به شبکه ستون فقرات در حالتیکه هر دو شاخه افزونگی استاندارد شده‌اند، نشان داده شده است. دیده می‌شود که دقت روی مجموعه داده CUB-200 بیشتر از مجموعه داده‌های دیگر بهبود یافته است (۳/۵%) و نرخ خطای رتبهٔ اول مدل پیشنهادی روی مجموعه دادهٔ CUB-200 نیز بیشترین کاهش را نشان می دهد (۲۸%). در صورتیکه نرخ خطای پنج رتبهٔ اول در مجموعه دادهٔ Dogs بیشترین کاهش را داشته است (۵۳/۴۹%) . بطور کلی خطای پنج رتبهٔ اول در همه مجموعه داده‌ها بطور قابل توجهی کاهش یافته است.

جدول ٦ بهبود نرخ دقت و خطا در هر سه مجموعه داده

مجموعه داده	دقت(%) رتبهٔ اول	خطا(%) رتبهٔ اول	دقت(%) ۵رتبهٔ اول	خطا(%) ۵رتبهٔ اول
CUB	۳/۵۰	۲۸/۰۰	۱/۶۳	۳۷/۶۴
Dogs	۲/۵۰	۱۵/۸۲	۱/۳۰	۵۳/۴۹
Cars	۱/۲۳	۱۲/۹۲	۰/۳۴	۲۰/۹۸

۴-بحث

هر چند در این بخش آزمایشات مختلفی را با مدل پیشنهادی ارائه داده و نتایج دقت و صحت عملکرد آن را با سایر مدل‌های به‌روز و پیشرفته سال‌های اخیر مقایسه نمودیم، لیکن هدف اصلی ما در این مقاله نشان دادن کارائی مدل‌های مبتنی بر معماری CSN بود. برای این منظور مدل CSN-AE پیشنهادی به عنوان یکی از مدل‌های ساده قابل تحقق با حداقل پارامتر ارائه گردید تا نشان دهد با چنین مدل سادهٔ مبتنی بر CSN نیز می‌توان به نتایج مقایسه‌پذیری در طبقه‌بندی ریزدانه دست پیدا کرد؛ اگرچه مدل‌های طبقه‌بندی ریزدانه برای رسیدن به دقت‌های بالا از الگوریتم‌های مکان‌یابی یا تلفیق متعددی استفاده می‌کنند که در مدل پیشنهادی ما وجود ندارد. بطور کلی اکثر مدل های اخیر

ابتدا بصورت صریح یا ضمنی مکان‌یابی نواحی را در تصویر انجام می‌دهند که این فرآیند با الگوریتم-های پیچیده‌ای همراه است. بیشتر مقالات اشاره شده در **جدول ٤** در این گروه هستند مانند [9], [17] ,[13] ,[12] ,[10]. برخی دیگر نیزعلاوه بر مکان یابی از الگوریتم های تلفیق نیز استفاده می‌کنند [9] و [15]. برخی از مدل‌ها الگوریتم‌های خاص افزونگی را به سایر مدل‌ها اضافه می‌کنند تا دقت طبقه‌بندی را افزایش دهند [16]. هر چند معماری پیشنهادی ما می‌تواند با بیشتر مدل‌های گفته شده ترکیب شده و بازده خوبی را رقم بزند اما در این مقاله ما خواسته‌ایم بر روی معرفی یک معماری تحلیلی و تفسیرپذیر و نشان دادن کارائی و اثر بخشی آن تأکید نماییم. به ویژه اینکه ماهیت تنک-سازی الگوریتم‌های کانولوشن بکار رفته در معماری پیشنهادی نوعی مکان یابی ضمنی را بوجود آورده و مانند مکانیزم توجه در معماری‌های مبتنی بر ترانسفورمرها عمل می‌کند. این موضوع در بازنمائی نشان داده شده در **شکل ۲** مشهود است. به نظر می‌رسد مدل CSN-AE بطور ذاتی دارای خاصیت محلی‌سازی اشیاء[1] می‌باشد و هنگام بازسازی تصویر در مرحله دیکدر فقط اطلاعات مهم و تفکیک‌کننده تصویر بازسازی می‌شوند. این موضوع در **شکل ۵** به وضوح دیده می‌شود. در این شکل هر چند به نظر می‌رسد تصویر بازسازی شده در خروجی SAE در مقایسه با تصویر طبیعی، جلوهٔ بصری کمتری دارد، لیکن نقشهٔ گرمائی تولید شده از خروجی اتوانکدر تنک نواحی حاوی اطلاعات مهم را نشان‌می‌دهد. از طرف دیگر با وجود اینکه ابعاد بردار ویژگی روش پیشنهادی در طبقه بند دو برابر مدل پایه است (۲۵۶۰ در مقابل ۱۲۸۰) لیکن در آزمایشات هیچگونه بیش‌برازش در طبقه‌بند مدل پیشنهادی دیده نشد که دلیل دیگری برای استخراج ویژگی‌های تفکیک‌پذیر توسط مدل پیشنهادی ما می‌باشد. به همین دلیل مزید شدن ساختار CSN به شاخه‌های حاوی شبکه‌های از پیش آموزش داده شده همواره دقت طبقه‌بندی را افزایش داده و نرخ خطا را کاهش داد. البته در کارهای بعدی می‌توان دقیق‌تر به این موضوع پرداخت و با ترکیب معماری CSN با سایر مدل‌های مرز دانش ابعاد دقیق‌تر این ساختار را بررسی نمود. موضوع دیگر استفاده ما از شبکه از پیش آموزش داده شده EfficientNet است. همانگونه که در **جدول ۲** دیده می شود این شبکه تقریبا کمترین میزان ابرپارامتر را دارا بوده و لذا بار محاسباتی و پیچیدگی مدل ما را به شدت کاهش داده است. از طرفی نتایج ازمایشات نشان دادند که ترکیب این با مدل پیشنهادی CSN-AE به ویژه بر روی مجموعه داده CUB-200 نتایج خوبی داشته و علی رغم سادگی مدل پیشنهادی دقت‌های مرز دانش را رقم زده است.

بطور کلی نتایج بدست آمده از پژوهش انجام شده در این مقاله و کارهای تحقیقاتی پیشین به ویژه آزمایش‌های انجام شده با مدل CSN، نشان‌دهندهٔ این مهم است که معماری‌های مستقل مبتنی شبکهٔ CSN که در این مقاله نیز ارائه گردید، تا این مرحله هنوز قابلیت دستیابی به نتایج مرز دانش را در مأموریت‌های چالشی همانند طبقه‌بندی ریزدانه ندارند که دلایلی برای آن متصور است. همانگونه

---

[1] object

که قبلاً نیز اشاره گردید مدل‌های مبتنی بر معماریCSN پیشنهادی در مراحل اولیه توسعه و کاربرد قرار دارد؛ کما اینکه شبکه‌های عصبی CNN نیز در بدو معرفی، شامل مدل‌های ساده و چند لایه مانند LeNet [67] بوده و پس از چندین سال شبکه‌های بسیار عمیق مبتنی بر CNN با چند صد لایه مانند شبکه‌های ResNet [68]، DenseNet [69] و EfficientNet معرفی گردیدند که از قابلیت‌های استخراج ویژگی بسیار مطلوبی برخوردار بودند. از طرفی به نظر می‌رسد میزان تنک‌سازی شبکه‌های CSN در هر مأموریت یک امر تجربی است و در آزمایشات گوناگون قابل تعیین باشد؛ همانطور که در شبکه‌های CNN، لزوماً با افزایش تعداد لایه‌ها نتایج مطلوبی حاصل نمی‌شود. در هر صورت این قبیل از چالش‌ها می‌تواند زمینه‌های پژوهشی آینده را در این حوزه به خود اختصاص دهد، همچنانکه تأثیر ترکیب انواع شبکه‌های عصبی عمیق با شبکه‌های مبتنی بر CSN و یا تأثیر سایر شبکه‌های عصبی به عنوان ستون فقرات مدل CSN پیشنهادی می‌تواند در کارهای آینده مورد توجه قرار گیرد.

۵–نتیجه‌گیری

در این مقاله ما یک معماری مبتنی بر CSN را معرفی کردیم که بر مبنای CSC-DL چند لایه طراحی و پیاده سازی شده است. این ساختار مشابه مدل‌های CNN در شبکه های عصبی بوده و در عین حال از افزونگی اطلاعات جلوگیری می‌کند. برای نشان دادن کارائی معماری پیشنهادی، یک مدل CSN-AE معرفی کردیم که می‌تواند با مدل‌های پایهٔ ریزدانه ترکیب شده و کارائی آنها را افزایش دهد. در این مقاله مدل پایه، شبکه EfficientNet-B0 انتخاب گردید که به تنهائی دارای دقت مناسبی در طبقه‌بندی ریزدانه بوده و کمترین میزان پارامترها را دارد. ترکیب CSN-AE با مدل EfficientNet-B0 توانست دقت‌های طبقه‌بندی را در سه مجموعه داده ریزدانه بهبود ببخشد که نشان‌دهنده کارامدی مدل CSN-AE می‌باشد. در آینده مایل هستیم اثربخشی معماری CSN را در مدل‌های طبقه‌بندی پایه بیشتری آزمایش کنیم.

۶–تشکر و قدردانی

نویسندگان از کارکنان مرکز اطلاع‌رسانی دانشگاه پدافند هوائی خاتم الانبیاء(ص) که منابع سخت-افزاری لازم را جهت اجرای شبیه‌سازی‌های این مقاله در اختیار گذشتند تشکر و قدردانی می‌نمایند.

۷–تعارض منافع

نویسنده(گان) اعلام می‌دارند که در مورد انتشار این مقاله تضاد منافع وجود ندارد. علاوه بر این، موضوعات اخلاقی شامل سرقت ادبی، رضایت آگاهانه، سوء رفتار، جعل داده‌ها، انتشار و ارسال مجدد و مکرر توسط نویسندگان رعایت شده است.

۸–دسترسی آزاد

(بازترکیب، تغییر شکل و بازسازی بر اساس محتوا) را می‌دهد.

۹–منابع

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset,” 2011.

[2] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, “Fine-Grained Visual Classification of Aircraft,” Jun. 2013, [Online]. Available: http://arxiv.org/abs/1306.5151

[3] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, “Novel dataset for fine-grained image categorization,” *CVPR Work. Fine-Grained Vis. Categ.*, pp. 806–813, 2011.

[4] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D object representations for fine-grained categorization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 554–561. doi: 10.1109/ICCVW.2013.77.

[5] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv:1409.1556v6 [cs.CV]*, pp. 1–14, Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[6] K. Zhang, M. Sun, S. Member, and T. X. Han, “Residual Networks of Residual Networks : Multilevel Residual Networks,” vol. 14, no. 8, pp. 1–12, 2016.

[7] T. Y. Lin, A. Roychowdhury, and S. Maji, “Bilinear Convolutional Neural Networks for Fine-Grained Visual Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1309–1322, 2018, doi: 10.1109/TPAMI.2017.2723400.

[8] Y. Xie, Q. Gong, X. Luan, J. Yan, and J. Zhang, “A Survey of Fine-Grained Visual Categorization Based on Deep Learning,” *J. Syst. Eng. Electron.*, vol. 35, no. 6, pp. 1337–1356, 2023, doi: 10.23919/jsee.2022.000155.

[9] W. Luo, H. Zhang, J. Li, and X. S. Wei, “Learning semantically enhanced feature for fine-grained image classification,” *IEEE Signal Process. Lett.*, vol. 27, pp. 1545–1549, 2020, doi: 10.1109/LSP.2020.3020227.

[10] F. V. Categorization *et al.*, “Attentional Kernel Encoding Networks for Fine-Grained Visual Categorization,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 301–314, 2021, doi: 10.1109/TCSVT.2020.2978115.

[11] P. Zhuang, Y. Wang, and Y. Qiao, “Learning attentive pairwise interaction for fine-grained classification,” *AAAI 2020 - 34th AAAI Conf. Artif. Intell.*, no. Bruner, pp. 13130–13137, 2020, doi: 10.1609/aaai.v34i07.7016.

[12] G. Sun, H. Cholakkal, S. Khan, F. S. Khan, and L. Shao, “Fine-grained recognition: Accounting for subtle differences between similar classes,” *AAAI 2020 - 34th AAAI*

Conf. Artif. Intell., pp. 12047–12054, 2020, doi: 10.1609/aaai.v34i07.6882.

[13] C. Liu, H. Xie, Z. Zha, L. Ma, L. Yu, and Y. Zhang, "Filtration and Distillation : Enhancing Region Attention for Fine-Grained Visual Categorization," *Proc. AAAI Conf. Artif. Intell. Vol. 34. No. 07. 2020.*, 2020, [Online]. Available: https://scholar.google.com/scholar?cluster=11663869378030505297&hl=en&as_sdt=0,5

[14] R. Du, J. Xie, Z. Ma, D. Chang, Y.-Z. Song, and J. Guo, "Progressive Learning of Category-Consistent Multi-Granularity Features for Fine-Grained Visual Classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9521–9535, Dec. 2022, doi: 10.1109/TPAMI.2021.3126668.

[15] M. Wang, P. Zhao, X. Lu, F. Min, and X. Wang, "Fine-Grained Visual Categorization: A Spatial-Frequency Feature Fusion Perspective," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2798–2812, 2023, doi: 10.1109/TCSVT.2022.3227737.

[16] Y. Pu, Y. Han, Y. Wang, J. Feng, C. Deng, and G. Huang, "Fine-Grained Recognition with Learnable Semantic Data Augmentation," *IEEE Trans. Image Process.*, vol. 33, pp. 3130–3144, 2024, doi: 10.1109/TIP.2024.3364500.

[17] Z. Miao, X. Zhao, J. Wang, Y. Li, and H. Li, "Complemental Attention Multi-Feature Fusion Network for Fine-Grained Classification," *IEEE Signal Process. Lett.*, vol. 28, pp. 1983–1987, 2021, doi: 10.1109/LSP.2021.3114622.

[18] K. L. and L. F.-F. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, "ImageNet: A Large-Scale Hierarchical Image Database," *IEEE Conf. Comput. Vis. pattern recognition. Ieee, 2009*, no. IEEE, pp. 248–255, 2009.

[19] A. Lou, S. Guan, and M. Loew, "CFPNet-M: A light-weight encoder-decoder based network for multimodal biomedical image real-time segmentation," *Comput. Biol. Med.*, vol. 154, 2023, doi: 10.1016/j.compbiomed.2023.106579.

[20] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020, doi: 10.1016/j.neunet.2019.08.025.

[21] H. Xie, Z. Chen, F. Hong, and Z. Liu, "CityDreamer: Compositional Generative Model of Unbounded 3D Cities," *2024 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 9666–9675, Sep. 2023, doi: 10.1109/CVPR52733.2024.00923.

[22] P.-Y. Chou, Y.-Y. Kao, and C.-H. Lin, "Fine-grained Visual Classification with High-temperature Refinement and Background Suppression," no. March, pp. 1–9, 2023, [Online]. Available: http://arxiv.org/abs/2303.06442

[23] L. Liu *et al.*, "Computing Systems for Autonomous Driving: State of the Art and Challenges," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6469–6486, 2021, doi: 10.1109/JIOT.2020.3043716.

[24] G. Sen Xie, X. Y. Zhang, S. Yan, and C. L. Liu, "Hybrid CNN and Dictionary-Based Models for Scene Recognition and Domain Adaptation," *IEEE Trans.*

Circuits Syst. Video Technol., vol. 27, no. 6, pp. 1263–1274, 2017, doi: 10.1109/TCSVT.2015.2511543.

[25] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal Brain Damage (Pruning)," *Adv. Neural Inf. Process. Syst.*, pp. 598–605, 1990.

[26] H. Tang *et al.*, "CSC-Unet: A Novel Convolutional Sparse Coding Strategy Based Neural Network for Semantic Segmentation," *IEEE Access*, vol. 12, no. January, pp. 35844–35854, 2024, doi: 10.1109/ACCESS.2024.3373619.

[27] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, "Sparse Autoencoders Find Highly Interpretable Features in Language Models," *12th Int. Conf. Learn. Represent. ICLR 2024*, pp. 1–20, 2024.

[28] A. Makelov, G. Lange, and N. Nanda, "Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control," no. 1, 2024, [Online]. Available: http://arxiv.org/abs/2405.08366

[29] Z. Zhao *et al.*, "Deep Convolutional Sparse Coding Networks for Interpretable Image Fusion," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2023-June, pp. 2369–2377, 2023, doi: 10.1109/CVPRW59228.2023.00234.

[30] X. Sun, N. M. Nasrabadi, and T. D. Tran, "Supervised deep sparse coding networks," *Proc. - Int. Conf. Image Process. ICIP*, pp. 346–350, 2018, doi: 10.1109/ICIP.2018.8451701.

[31] V. Papyan, J. Sulam, and M. Elad, "Working Locally Thinking Globally - Part II: Stability and Algorithms for Convolutional Sparse Coding," pp. 1–13, 2016, [Online]. Available: http://arxiv.org/abs/1607.02009

[32] V. Papyan, J. Sulam, and M. Elad, "Working locally thinking globally: Theoretical guarantees for convolutional sparse coding," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5687–5701, 2017, doi: 10.1109/TSP.2017.2733447.

[33] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. Trends Comput. Graph. Vis.*, vol. 8, no. 2–3, pp. 85–283, 2014, doi: 10.1561/0600000058.

[34] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. 2, pp. 391–398, 2013, doi: 10.1109/CVPR.2013.57.

[35] M. Elad, *Sparse and Redundant Representations*. New York, NY: Springer New York, 2010. doi: 10.1007/978-1-4419-7011-4.

[36] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," *Adv. Neural Inf. Process. Syst.*, no. Nips, pp. 2082–2090, 2016.

[37] Y. Zhang *et al.*, "Advancing Model Pruning via Bi-level Optimization," *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, pp. 1–18, 2022.

[38] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” *7th Int. Conf. Learn. Represent. ICLR 2019*, pp. 1–42, 2019.

[39] S. Liu and Z. Wang, “Ten Lessons We Have Learned in the New ‘Sparseland’: A Short Handbook for Sparse Neural Network Researchers,” pp. 1–16, 2023, [Online]. Available: http://arxiv.org/abs/2302.02596

[40] Y. He and L. Xiao, “Structured Pruning for Deep Convolutional Neural Networks: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 2900–2919, 2024, doi: 10.1109/TPAMI.2023.3334614.

[41] S. Rajamanoharan *et al.*, “Improving Dictionary Learning with Gated Sparse Autoencoders,” pp. 1–37, Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.16014

[42] V. Monga, Y. Li, and Y. C. Eldar, “Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing,” *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, 2021, doi: 10.1109/MSP.2020.3016905.

[43] H. Sreter and R. Giryes, “Learned Convolutional Sparse Coding,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 2191–2195, 2018, doi: 10.1109/ICASSP.2018.8462313.

[44] K. Gregor and Y. LeCun, “Learning Fast Approximations of Sparse Coding,” *Proc. 27th Int. Confer- ence Mach. Learn. Haifa, Isr. 2010.*, pp. 399–406, May 2010.

[45] F. Gao, X. Deng, M. Xu, J. Xu, and P. L. Dragotti, “Multi-Modal Convolutional Dictionary Learning,” *IEEE Trans. Image Process.*, vol. 31, pp. 1325–1339, 2022, doi: 10.1109/TIP.2022.3141251.

[46] T. Liu, A. Chaman, D. Belius, and I. Dokmanic, “Learning Multiscale Convolutional Dictionaries for Image Reconstruction,” *IEEE Trans. Comput. Imaging*, vol. 8, pp. 425–437, 2022, doi: 10.1109/TCI.2022.3175309.

[47] C. Garcia-Cardona and B. Wohlberg, “Convolutional Dictionary Learning: A Comparative Review and New Algorithms,” *IEEE Trans. Comput. Imaging*, vol. 4, no. 3, pp. 366–381, Sep. 2017, doi: 10.1109/TCI.2018.2840334.

[48] B. Wohlberg, “Convolutional sparse representation of color images,” in *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, 2016, pp. 57–60. doi: 10.1109/SSIAI.2016.7459174.

[49] B. Wohlberg, “Efficient algorithms for convolutional sparse representations,” *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, 2016, doi: 10.1109/TIP.2015.2495260.

[50] Robert Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996, [Online]. Available: https://www.jstor.org/stable/2346178

[51] A. Aberdam, J. Sulam, and M. Elad, “Multi-Layer Sparse Coding: The Holistic Way,” *SIAM J. Math. Data Sci.*, vol. 1, no. 1, pp. 46–77, Apr. 2019, doi:

10.1137/18m1183352.

[52] J. Sulam, A. Aberdam, A. Beck, and M. Elad, “On Multi-Layer Basis Pursuit, Efficient Algorithms and Convolutional Neural Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 320649, pp. 1–13, 2019, doi: 10.1109/TPAMI.2019.2904255.

[53] V. Papyan, Y. Romano, J. Sulam, and M. Elad, “Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks,” *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 72–89, 2018, doi: 10.1109/MSP.2018.2820224.

[54] J. Sulam, V. Papyan, Y. Romano, and M. Elad, “Multilayer convolutional sparse modeling: Pursuit and dictionary learning,” *IEEE Trans. Signal Process.*, vol. 66, no. 15, pp. 4090–4104, 2018, doi: 10.1109/TSP.2018.2846226.

[55] V. Papyan, Y. Romano, and M. Elad, “Convolutional neural networks analyzed via convolutional sparse coding,” *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 1–52, 2017.

[56] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *IEEE Access*, vol. 9, pp. 16591–16603, May 2015, doi: 10.1109/ACCESS.2021.3053408.

[57] F. S. Almalou, F. Razzazi;, and A. Amini, “A Multi- Layer Convolutional Sparse Network for Pattern Classification Based on Sequential Dictionary Learning,” *IET Comput. Vis.*, vol. 20, no. 1, 2026, doi: https://doi.org/10.1049/cvi2.70055.

[58] S. Chen and D. . Donoho, “Basis Pursuit,” *Proc. 1994 28th Asilomar Conf. Signals, Syst. Comput.*, pp. 41–44, 1994.

[59] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” *Found. Trends® Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010, doi: 10.1561/2200000016.

[60] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[61] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[62] A. Dubey, O. Gupta, R. Raskar, and N. Naik, “Maximum Entropy Fine-Grained Classification,” *32nd Conf. Neural Inf. Process. Syst. (NeurIPS 2018), Montréal, Canada*, no. iii, 2018.

[63] R. Wightman and H. Touvron, “ResNet strikes back : An improved training procedure in timm,” pp. 1–22, 2021.

[64] Y. Wang, V. I. Morariu, and L. S. Davis, “Learning a Discriminative Filter Bank Within a CNN for Fine-Grained Recognition,” *Proc. IEEE Comput. Soc. Conf.*

Comput. Vis. Pattern Recognit., pp. 4148–4157, 2018, doi: 10.1109/CVPR.2018.00436.

[65] X. He, Y. Peng, and J. Zhao, "Fast Fine-Grained Image Classification via Weakly Supervised Discriminative Localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1394–1407, May 2019, doi: 10.1109/TCSVT.2018.2834480.

[66] Q. Xu, S. Li, J. Wang, B. Jiang, and J. Tang, "Context-Semantic Quality Awareness Network for Fine-Grained Visual Categorization," pp. 1–13, Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.10298

[67] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Front. Psychol.*, vol. 4, no. APR, pp. 428–429, Dec. 2015, doi: 10.3389/fpsyg.2013.00124.

[69] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 2261–2269, 2017, doi: 10.1109/CVPR.2017.243.