احتمال تعمیم‌پذیر نبودن روش‌های ارزیابی کیفیت مبتنی بر بردار پشتیبان

آرمین احمدزاده *و پوریا چراغی †

مرکز پردازش سریع، پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی، تهران، ایران

چکیده

یک مدل محاسباتی که بتواند نظرِ انسان در مورد کیفیتِ تصاویر را پیش‌بینی نماید، کاربردهای فراوانی در مسائل مربوطَ به پردازش تصویر دارد. به دلیلِ پیچیدگی دستگاه بینایی بشر، مدل‌های یادگیری ماشین برای شبیه‌سازیِ نحوه‌ی قضاوتِ انسان‌ها استفاده می‌شوند. یکی از این مدل‌ها، وایازش بردار پشتیبان است. در این مقاله، نشان می‌دهیم که روش رایج برای بکارگیریِ بردارِ پشتیبان در ارزیابی کیفیت تصویر، لزوماً به مدل‌های قابلِ تعمیم نمی‌انجامد.

واژه‌های کلیدی: هوش مصنوعی، پردازش تصویر، دستگاه بینایی انسان، یادگیری ماشین، ارزیابی کیفیت تصویر، نامفاد، ماشین بردار پشتیبان، وایازش

۱ مقدمه

تصاویر دیجیتال بخش قابل توجهی از مصرف رسانه‌ای بشر را تشکیل می‌دهند [۱]. بدیهی است که کیفیّتِ کَمِ این رسانه، منجر به نارضایتی کاربران و ناکامی ما در دریافت اطّلاعات مدّنظرمان خواهد شد. لذا، مطلوب است که وضعیت تصاویر ارسالی، از منظر کیفیت‌شان، تحت نظارت باشد.

مطمئن‌ترین راه برای ارزیابی کیفیت تصویر (که به اختصار «اَکْتْ» می‌خوانیمش)، پرسش از انسان‌هاست [۲]. به این ترتیب که تصویری را به جمعی از ناظرین[۱] نشان داده و نظر آن‌ها در مورد کیفیت تصویر را در قالب نمره‌ای از یک بازه‌ی مشخّص (، مثلاً $[0, 100]$،) جویا می‌شویم. میانگین نظرات[۲]، یک کمّیّت قابل اطمینان از کیفیّت تصویر است [۳]. چنین آزمایشی نیز اَکْتِ «انسانی[۳]» نام دارد.

بدیهی است که اَکْتِ انسانی برای کاربردهای برـخط[۴] و حجیم امکان‌پذیر نیست. در این موارد، نرم‌افزاری مطلوب است که بتواند قضاوت دستگاه بینایی بشر در مورد کیفیت یک تصویر را پیش‌بینی نماید. دقت و سرعت این پیش‌بینی، معیارهای عملکرد یک مدل محاسباتیِ اَکْتْ[۵] هستند. ساخت چنین مدلی، موضوع مورد مطالعه‌ی بخش قابل توجهی از تحقیقاتِ پردازش تصویر است[۴]. برای ارزیابی و آموزش روش‌های محاسباتی، از نتایج اَکْتِ انسانی استفاده می‌شود.

ساده‌ترین راه برای اَکْتِ محاسباتی، محاسبه‌ی اختلاف دو تصویر است. موارد زیادی وجود دارند که یک تصویر سالم (با کیفیت احتمالاً مطلوب،) دچار تخریب شده و یک نسخه‌ی تخریب‌شده از آن بوجود می‌آید. وقتی هردوی این دو تصویر در دسترس باشند، می‌توان برای ارزیابی کیفیتِ تصویرِ تخریب‌شده، به اطلاعات موجود در تصویر سالم رجوع کرد. به این کار اَکْتِ «مرجع کامل» گفته می‌شود [۵]. یک مثال از کاربرد اَکْتِ مرجع کامل، فشرده‌سازی تصاویر است. الگوریتم فشرده‌کننده با کاهشِ اطلاعاتِ نشانک[۶]، حجم آن را کاهش داده و از طرفی کیفیّت آن را نیز خدشه‌دار می‌نماید. الگوریتم اَکْتِ مرجع کامل می‌تواند در هر لحظه تصویر فشرده‌شده را با تصویر اصلی مقایسه کرده و در صورت تخریبِ بیش از حدِّ کیفیت، این مورد را به الگوریتم فشرده‌کننده بازخورد دهد [۶].

اگر $x(i,j)$ نشانک مرجع و $y(i,j)$، نشانک تخریب‌شده باشد، میانگین مربّعات خطا، «MSE[۷]»، اختلاف مقادیر آن دو را خواهد سنجید:

$$MSE(x,y) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (x(i,j) - y(i,j))^2 \quad (۱)$$

M و N در (۱)، ابعاد تصاویر هستند. نشان داده می‌شود که «MSE» یا بیشینه‌ی نسبت نشانک به نوفه[۸]، «PSNR[۹]»، نمی‌توانند درک انسان از کیفیت تصویر را شبیه‌سازی نمایند [۷]. وانگ[۱۰] و همکارانش نشان دادند که تخریب ساختارهای تصویر در کاهش کیفیت آن مؤثر است[۷]. در سال ۲۰۰۴، روشی به نام «SSIM[۱۱]» برای کمّی‌سازی شباهت ساختاری دو تصویر ارائه کردند [۸] که با اختلاف از روش‌های مبتنی برـ«MSE» بهتر بود [۹].

با موفقیت SSIM، محققین سعی کردند ساختارهای تصویر را با ویژگی‌های دیگری، مثل لبه‌ها، مدل کنند. آماره‌های تصاویر طبیعی[۱۲] [۱۰] و پایداری اطلاعاتی[۱۳] [۱۱]، دیگر معیارهای پیشنهاد شده برای کیفیت تصویر هستند. روش‌های یادگیری ماشین سرـتاـسری[۱۴] نیز، برای اَکْتِ محاسباتی استفاده شده‌اند [۱۲، ۱۳، ۱۴].

یک ساز و کار رایج برای اَکْتْ، استفاده از ماشین بردار پشتیبان، «SVR[۱۵]» [۱۵] است [۱۶]. کلّیات بکارگیری «SVR» در شکل ۱ دیده می‌شود. در روش‌های مبتنی بر بردار پشتیبان، نوآوری اصلی، در طراحی ویژگی‌ها صورت می‌گیرد. چنین روشی، ابتدا از تصویر ورودی یک بردار ویژگی استخراج می‌کند. اگر روش مرجع کامل باشد، می‌تواند برای استخراج ویژگی به نسخه‌ی سالم نیز رجوع کند. بردار ویژگی، $\vec{f}$،

*a.ahmadzadeh@ipm.ir

†cheraaqee@ipm.ir

[۱] همان «subject»ها در آزمایش‌های علمی

[۲] که به اختصار «MOS» نامیده می‌شود و مخفّف عبارت «Mean Opinion Score» است. لازم به توضیح است که «MOS» نسبت مستقیمی با کیفیت ادراکی و «Difference-MOS» یا «DMOS» نسبتی عکس با کیفیت تصویر دارد.

[3] subjective image quality assessment

[۴] معادل فارسی «on-line»

[5] objective image quality assessment

[۶] معادل فارسی «signal»

[7] Mean Squared Error

[۸] معادل فارسی «noise»

[9] Peak Signal-to-Noise Ratio

[10] Zhou Wang

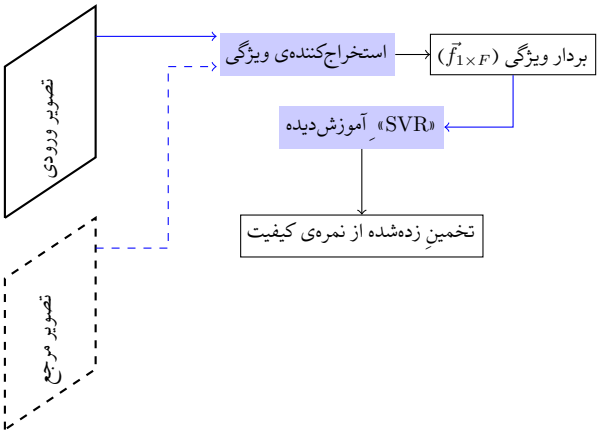
[11] Structural Similarity

[12] Natural Scene Statistics- NSS

[13] Information Fidelity

[14] end-to-end

[15] Support Vector Regression



شکل ۱: بکارگیری «SVR» برای اَکْتْ

آرایه‌ای از اعداد است. اگر F ویژگی استخراج شوند، ابعاد این بردار، $1 \times F$ خواهد بود. یک مدل مبتنی بر وایازش[۱۶] بردار پشتیبان، آموخته است که این بردار ویژگی را به نمره‌ی کیفیت نگاشت کند. برای آموزش چنین مدلی، از تصاویر و نمرات ارزیابی‌های انسانی استفاده می‌شود. در این مقاله نشان می‌دهیم که مدل حاصل از روش رایج برای آموزش بردار پشتیبان که در اَکْتْ‌های محاسباتی استفاده می‌شود، لزوماً قابل تعمیم[۱۷] به تصاویر مختلف نیست.

مبانی مورد نیاز و برخی از کارهای مرتبط در قسمت۲ مرور می‌شوند. قسمت ۳، آزمایش‌های انجام شده را تشریح خواهد کرد و مقاله در قسمت ۴ جمع‌بندی می‌شود.

۲ مبانی و مرور ادبیات

در این قسمت برخی مفاهیم و قراردادهای رایج در تحقیقات اَکْتِ محاسباتی تشریح می‌شوند. همچنین، یک دسته‌بندی از حوزه‌های ارزیابی کیفیت ارائه می‌گردد.

۱.۲ مجموعه‌داده‌ها

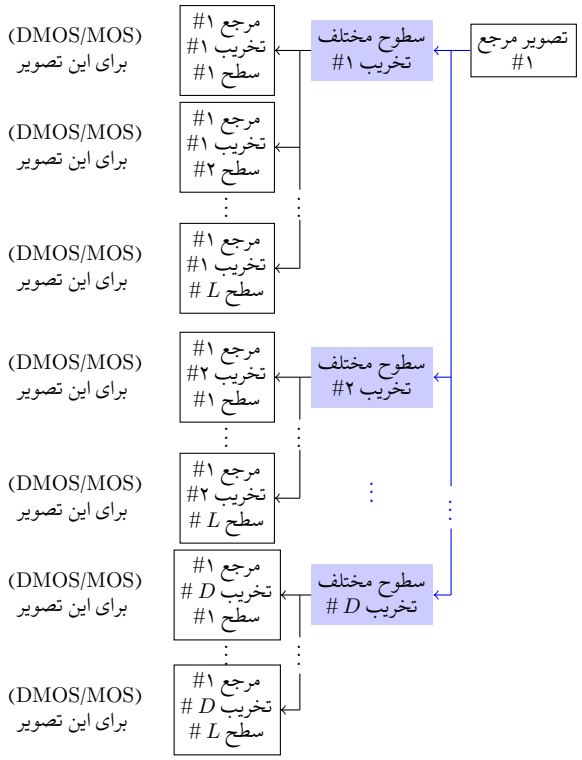
در اکثر کاربردها، انسان ناظرِ نهایی تصاویر است. لذا، نظر انسان بهترین معیارِ اَکْتْ برای این کاربردها خواهد بود. روش‌های محاسباتی سعی می‌کنند که نمراتی مثل نتایج اَکْتِ انسانی تولید کنند. مجموعه‌داده‌هایی وجود دارد که نظرات انسان به همراه تصاویر مربوطه را در اختیار محققین قرار می‌دهند، تا روش‌های اَکْتِ محاسباتی خود را محک بزنند [۳]. به این مجموعه‌داده‌ها، مجموعه‌داده‌های اَکْتْ گفته می‌شود.

یک مجموعه‌داده‌ی اَکْتْ، شامل یک سری تصویر مرجع، نسخ تخریب‌شده و نمرات انسانی است (شکل ۲). برخی اتفاقات برای تصاویر رایج هستند؛ مثلاً تار شدن به علت فاصله‌ی کانونی نامناسب، یا تَصَنُّعاتی[۱۸] که به دلیل فشرده‌سازی ایجاد می‌شود. به دلیل رایج بودن این تخریب‌ها، مهم است که توانایی الگوریتم‌های اَکْتْ در سنجش آن‌ها بررسی شود. به همین دلیل، یک سری تصویر سالم، با کیفیت قابل قبول، انتخاب کرده و این تخریب‌ها را به صورت مصنوعی روی آن‌ها اعمال می‌کنند. هر تخریب هم می‌تواند با شدّت متفاوتی اعمال شود. یک تصویر می‌تواند مقداری تار، یا خیلی تار باشد. بدیهی است که نمره‌ی

[۱۶] معادل فارسی «regression»

[17] generalization

[18] artifacts



شکل ۲: قسمتی از آنچه که در یک مجموعه‌داده‌ی اَکْتْ وجود دارد. فرض می‌کنیم که R تصویر مرجع و D نوع تخریب وجود دارد که هر تخریب با L سطح از شدّت اعمال می‌شود. شکل، موارد را برای مرجع #۱ نشان می‌دهد. همین موارد برای مراجع #۲ تا #R نیز تکرار می‌شوند.

انسانی برای شدّت‌های مختلف، متفاوت خواهد بود. انتظار می‌رود که نمره‌ی حاصل‌شده از روش محاسباتی نیز با نمره‌ی انسانی مطابق باشد [۱۷].

بنابراین، هر مجموعه‌داده‌ی اَکْتْ، شاملِ مجموعه‌ای از «نمونه»ها است. هر نمونه، یک زوج مرتب به شکل (نمره‌ی انسانی, تصویر) است. نمره‌ی انسانی می‌تواند به صورت MOS یا DMOS ذخیره شده باشد. به این ترتیب که، «تصویر»، ورودیِ یک روش محاسباتی، و «نمره‌ی انسانی»، پاسخِ صحیحِ موردِ انتظار[۱۹] است. حال اگر روش مرجع کامل باشد، تصویر مرجعِ متناظر نیز به عنوان ورودی به الگوریتم داده می‌شود. برای ارزیابی یک الگوریتم، نمرات آن برای تصاویر موجود در یک مجموعه‌داده محاسبه شده و سپس همبستگیِ نمرات الگوریتم با نمرات انسانی اندازه‌گیری می‌شود. معیار همبستگی، همان شاخص‌های آماری (پیرسُنْ[۲۰]، اِسْپیرْمَنْ[۲۱] و غیره) هستند. هرچه این همبستگی بیشتر باشد، الگوریتم در پیش‌بینی نظر انسان، دقیق‌تر عمل کرده است [۹].

۲.۲ آموزش و آزمون روش‌های مبتنی بر SVR

طبق روش رایج برای بکارگیری SVR در اَکْتْ [۱۸]، ابتدا باید صحنه‌های یک مجموعه‌داده را افراز[۲۲] کنیم. منظور از «صحنه[۲۳]» در یک

[۱۹] همان «ground truth» در ادبیات یادگیری ماشین

[20] Pearson Correlation Coefficient- PLCC

[21] Spearman Rank Order Correlation Coefficient- SROCC

[۲۲] افراز یک مجموعه به معنی تقسیم آن به زیرمجموعه‌هایی است که با هم اشتراک ندارند. یعنی اگر یک مجموعه‌داده را به زیرمجموعه‌هایی برای آموزش و آزمون افراز کنیم، آن زیرمجموعه‌ها عضو مشترکی ندارند.

[23] scene

مجموعه‌داده، مجموعه‌ی تمامیِ تصاویرِ تخریب‌شده‌ای است، که متعلق به **یک** مرجع هستند. (یعنی یک منظره‌ی یکسان را نشان می‌دهند، منتها با تخریب‌های متفاوت و شدّت‌های متفاوت.) اگر ٪۸۰ **صحنه‌ها** را، به طور تصادفی، برای آموزش، و ٪۲۰ را برای آزمون کنار گذاشته باشیم، یک «**افرازِ ۸۰–۲۰**» انجام داده‌ایم.

روشِ مبتنی بر SVR، ابتدا از تصویر یک بردار ویژگی استخراج می‌کند. اگر در مجموعه‌داده، S نمونه، به شکل DS_{image} در (۲) داشته باشیم، استخراج‌کننده‌ی ویژگی، S بردار ویژگی محاسبه می‌کند که مجموعه‌ی $DS_{\text{feature vector}}$ در (۳) را تشکیل می‌دهند.

$$DS_{image} = \{(\text{نمره‌ی انسانی}_S, \text{تصویر}_S), \ldots, (\text{نمره‌ی انسانی}_1, \text{تصویر}_1)\} \quad (۲)$$

$$DS_{\text{vector feature}} = \{(\text{نمره‌ی انسانی}_S, \vec{f}_S), \ldots, (\text{نمره‌ی انسانی}_1, \vec{f}_1)\} \quad (۳)$$

وقتی افرازِ تصادفیِ ۸۰–۲۰ را انجام دهیم، برخی اعضای $DS_{\text{feature vector}}$ مربوط به صحنه‌هایی هستند که در مجموعه‌ی آموزش[۲۴] قرار گرفته‌اند و سایر اعضای آن مربوط به صحنه‌هایی هستند که در مجموعه‌ی آزمون[۲۵] قرار گرفته‌اند. می‌توانیم عامل‌های[۲۶] یک بردار پشتیبان را با استفاده از بردارهای ویژگی و نمرات انسانیِ مجموعه‌ی آموزش بهینه کرده و عملکرد آن را روی مجموعه‌ی آزمون ارزیابی کنیم.

از آنجایی که انتخاب صحنه‌ها برای افراز به مجموعه‌های آموزش و آزمون به صورت تصادفی صورت می‌گیرد، می‌توانیم افرازِ تصادفیِ ۸۰–۲۰ را چند بار انجام دهیم و میانه‌ی[۲۷] عملکرد وایازنده[۲۸] در این چند بار را به عنوان دقت الگوریتم در نظر بگیریم. اگر این کار را n دفعه انجام دهیم، اصطلاحاً می‌گویند که «**اعتبار سنجیِ متقابلِ nـلایه**[۲۹]» انجام داده‌ایم. مقدار پیشنهاد شده برای n، در اَکْتْ، ۱۰۰۰ است [۱۸]. یعنی هزار بار، افرازِ ۸۰–۲۰ را به صورت تصادفی انجام دهیم، آموزش و آزمون را اجرا نمائیم و میانه‌ی عملکرد مدل در این هزار دفعه را به عنوان دقت الگوریتم گزارش کنیم.

البته باید این نکته را هم در نظر بگیریم که وایازنده‌ی مبتنی بر بردار پشتیبان، دو فراعامل[۳۰]، به نام‌های c(ost) و γ، دارد. می‌توانیم نحوه‌ی کار وایازنده را به صورت رابطه‌ی (۴) بیان نماییم.

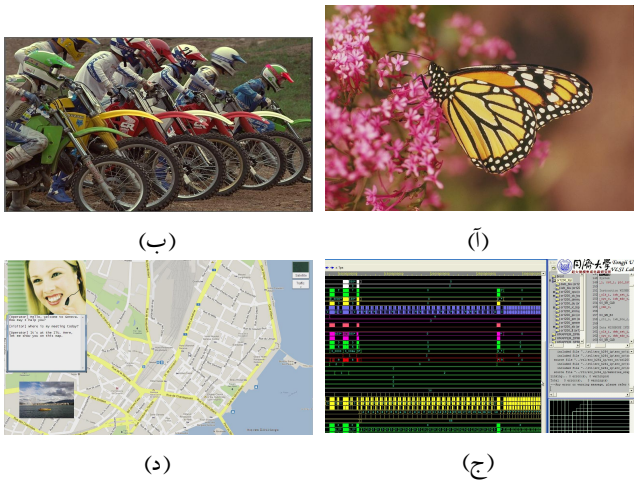
$$\text{پیش‌بینی} = \text{SVR}_{c,\gamma}(\vec{W} \cdot \vec{f}^T + \vec{b}) \quad (۴)$$

که $\vec{W}$ و $\vec{b}$ وزن‌های بهینه شده با استفاده از داده‌های آموزشی و $\vec{f}$ بردار ویژگیِ ورودی است. مقادیر c و γ قابل یادگیری نبوده و به صورت دستی تعیین می‌شوند. می‌توان بازه‌ای از مقادیر برای این دو فراعامل در نظر گرفت و اعتبار سنجی متقابل ۱۰۰ـلایه را روی یک مجموعه‌داده برای تمامی این مقادیر انجام داد. زوج (c, γ)ای که بهترین عملکرد را داشته باشند، به عنوان فراعامل‌های انتخابی در نظر گرفته می‌شوند [۱۸].

۱.۲.۲ ارزیابی جداگانه‌ی عملکرد الگوریتم روی هر یک از تخریب‌ها

همانطور که در قسمت ۱.۲ گفتیم، برخی تخریب‌ها رایج هستند و مجموعه‌داده‌ها نمونه‌هایی با این تخریب‌ها دارند. برای اندازه‌گیری دقت

[24] training set
[25] test set
[26] معادل فارسی «parameter»
[27] median
[28] regressor
[29] n-fold cross validation
[30] hyper-parameter



شکل ۳: دو نمونه تصویر طبیعی (در (آ) و (ب)) و دو نمونه نمامفاد (در (ج) و (د))

یک الگوریتم در ارزیابی این تخریب‌ها، همان اعتبار سنجیِ ۱۰۰۰ـلایه را انجام می‌دهیم. تفاوت این است که هنگام سنجش الگوریتم روی مجموعه‌ی آزمون، تنها نمونه‌هایی را در نظر می‌گیریم که دارای تخریب مورد نظر هستند [۱۸].

۳.۲ حوزه‌های مختلف اَکْتْ

مشخص است که دقت الگوریتمِ مبتنی بر بردار پشتیبان، به طور اصلی، به ویژگی‌هایی که استخراج می‌شوند بستگی دارد. هر چه این ویژگی‌ها گویاتر[۳۱] باشند، SVR هم کار راحت‌تری برای یادگیریِ نحوه‌ی نگاشت ویژگی‌ها به نمره‌ی کیفیت خواهد داشت [۱۷]. ویژگی‌های گویا برای هر تصویری متفاوت خواهند بود و طراحی روشی که بتواند برای هر تصویری ویژگی‌های مربوط به کیفیت را محاسبه کند، کار دشواری است [۱۹].

غالب تحقیقات اَکْتْ محاسباتی، برای تصاویر طبیعی انجام گرفته است [۲۰]. تصویرِ طبیعی، تصویری است که با دوربین‌های حساس به طیفِ مرئیِ موج الکترومغناطیس و از جهان فیزیکیِ پیرامون گرفته شده باشد [۲۱].

با شیوع تصویربرداری دیجیتال، تصاویر زیادی از اسناد و نوشته‌ها بوجود آمده است. ارزیابی کیفیت این تصاویر به روش‌هایی نیاز دارد که با روش‌های مربوط به تصاویر طبیعی متفاوت هستند [۲۲]. دلیل این تفاوت، وجود مقادیر زیاد نویسه‌ها و نواحی سیاه‌ـسفید است.

ترکیبِ تصاویر طبیعی و تصاویر اسناد، «**نَمامُفادْها**[۳۲]» را می‌سازد (شکل ۳). محاسبات از راه دور[۳۳] و آموزشِ مجازی، شرایطی را بوجود می‌آورند که نیاز است مُفادّ صفحه‌نمایشِ یک رایانه به عنوان تصویر مخابره شود. در [۲۰] نشان داده می‌شود که نمامفادها هم به روش‌های مختص به خود نیازمند هستند. مجموعه‌داده‌های کیفیت متخص نمامفادها نیز برای تحقیقات اَکْتْ ارائه شده است [۲۰، ۲۳].

در این قسمت، نحوه‌ی آموزش و آزمون روش‌های اَکْتِ مبتنی بر بردار پشتیبان را مرور کردیم. همچنین، دیدیم که اُکْتِ محاسباتی، کاربردهای مختلفی برای انواع تصاویر دیجیتال دارد. در قسمت بعد، برخی جنبه‌های مربوط به یادگیری ماشین در آموزش SVRها را تحلیل می‌نماییم.

[31] expressive
[32] معادل فارسی «screen content image». دقت شود که «نماگرفت»ها، «screen-shots»، یک حالت خاصِّ نمامفادها هستند.
[33] remote computing

۳ آزمایش‌های پیشنهادی و نتایج

در این قسمت می‌بینیم که SVRهای آموزش‌دیده برای اَکْتْ، ممکن است مشکلاتی از نظر تعمیم‌پذیری داشته باشند. ابتدا یک آزمایش تعریف کرده و سپس به تحلیلِ عملکردِ SVR می‌پردازیم.

۱.۳ آزمایش پیشنهادی

گفتیم که نمامفادها از نوشته‌ها، تصاویر طبیعی و نگاشتارها[۳۴] تشکیل شده‌اند (قسمت ۳.۲). یک راه برای ارزیابی کیفیت نمامفادها، تحلیل جداگانه‌ی این نواحی است [۲۴]. اگر یک روش، مختصِّ ارزیابی کیفیت تصاویر اسناد داشته باشیم، به نظر می‌رسد که ترکیبِ آن با روشی مختصِّ تصاویر طبیعی، برای ارزیابی کیفیت نمامفاد، شانس داشته باشد.

$LBPSI(.,.)$ [۲۵]، یک روش مرجع کامل، برای اَکْتِ اسناد و نوشته‌هاست. به این ترتیب که اگر x و y مثل (۱) تعریف شده باشند، خواهیم داشت:

$$\text{LBPSI} = LBPSI(x, y) \quad \text{(نمره‌ی کیفیت تخمینی به وسیله‌ی LBPSI)} \quad (۵)$$

همینطور، $SQMS(.,.)$ [۲۶]، یک روش مرجع کامل برای نمامفاد است. برای ارزیابی نواحی طبیعی، از $HaarPSI(.,.)$ [۲۷] و $GMSD(.,.)$ [۲۸] استفاده می‌کنیم که جنبه‌های مختلف ساختاری را می‌سنجند. از آنجایی که ارزیابی تباین[۳۵] کار دشواری است [۲۹]، می‌توانیم از $VIF(.,.)$ [۱۱] هم استفاده کنیم که در این زمینه دقیق است.

با استفاده از این روش‌ها، می‌توانیم یک روش ترکیبیِ مرجع کامل [۳۰]، برای ارزیابی نمامفادها بسازیم. SQMS، LBPSI، HaarPSI، GMSD و VIF را برای x و y محاسبه می‌کنیم. در این صورت، ۵ عدد خواهیم داشت که می‌توانیم با آن‌ها یک بردار ویژگی تشکیل دهیم:

$$\vec{f} = [\text{SQMS}, \text{LBPSI}, \text{HaarPSI}, \text{GMSD}, \text{VIF}] \quad (۶)$$

با داشتن این بردار ویژگی، می‌توانیم یک SVR را برای نگاشت آن به نمره‌ی کیفیت آموزش دهیم. برای داده‌ی آموزشی، از مجموعه‌داده‌های کیفیت نمامفاد، به نام‌های SIQAD [۲۰] و SCID [۲۳] استفاده می‌کنیم.

SIQAD شامل ۲۰ تصویر مرجع است، که هر کدام با ۷ نوع تخریب، تغییر یافته‌اند. هر تخریب در ۷ سطح اعمال شده است. بنابراین، $980 = 20 \times 7$ تصویر تخریب‌شده در SIQAD وجود دارد. علاوه بر تخریب‌های SIQAD، SCID دو تخریب دیگر را نیز به ۴۰ تصویرِ مرجعِ خود اعمال کرده است. هر تخریب در ۵ سطحِ شدّت شبیه‌سازی شده است، لذا در SCID، ۱۸۰۰ تصویر تخریب‌شده موجود است. برای تصاویر تخریب‌شده در این دو مجموعه‌داده، نمرات کیفیت انسانی تهیه شده است.

علاوه بر (۶)، می‌توانیم ترکیب‌های دیگری، مثل $\vec{f_{\prime}} = [SQMS, LBPSI, GMSD]$ را در نظر بگیریم. از آنجایی که $SQMS(.,.)$، خود یک روش مختصِّ نمامفاد است، تأثیر ترکیب سایر روش‌ها با آن را بررسی می‌کنیم. به این ترتیب، ترکیباتی که می‌توان امتحان کرد را در جدول ۱ خلاصه نموده‌ایم. برای مثال، $\vec{f_1}$، حاصلِ ترکیبِ $SQMS(.,.)$ و $HaarPSI(.,.)$ است. یعنی: $\vec{f_1} = [\text{SQMS}, \text{HaarPSI}]$. یا $\vec{f_6} = [\text{SQMS}, \text{HaarPSI}, \text{VIF}]$. بدیهی است که برای هر یک از بردارهای ویژگیِ $\vec{f_1}$ تا $\vec{f_{14}}$، باید SVR مجزّایی آموزش داده شود. در ادامه، عملکرد هر یک از این مدل‌ها بررسی می‌شوند.

[34] graphics
[35] contrast

جدول ۱: ترکیب‌های درنظر گرفته‌شده برای تشکیل بردار ویژگی. وجودِ ✓ به معنی استفاده شدنِ نمره‌ی روشِ آن ستون است.

بردار	SQMS	HaarPSI	GMSD	VIF	LBPSI
$\vec{f_1}$	✓	✓			
$\vec{f_2}$	✓		✓		
$\vec{f_3}$	✓			✓	
$\vec{f_4}$	✓				✓
$\vec{f_5}$	✓	✓	✓		
$\vec{f_6}$	✓	✓		✓	
$\vec{f_7}$	✓	✓			✓
$\vec{f_8}$	✓		✓	✓	
$\vec{f_9}$	✓		✓		✓
$\vec{f_{10}}$	✓			✓	✓
$\vec{f_{11}}$	✓	✓	✓	✓	
$\vec{f_{12}}$	✓	✓	✓		✓
$\vec{f_{13}}$	✓	✓		✓	✓
$\vec{f_{14}}$	✓		✓	✓	✓

۲.۳ موفقیت ترکیبِ روش‌ها، در هر یک از مجموعه‌داده‌ها

بردارهای $\vec{f_1}$ تا $\vec{f_{14}}$ را برای نمونه‌های SIQAD و SCID محاسبه کرده، و اعتبارسنجیِ متقابلِ ۱۰۰ـ لایه را برای یک یک آن‌ها انجام می‌دهیم. شکل ۴، بهره‌وریِ عملکردِ مدل‌های ترکیبی، نسبت به استفاده‌ی مستقیم از SQMS را نشان می‌دهد. منظور از بهره‌وری، محاسبه‌ی مقدار زیر است:

$$\text{بهره‌وری}_{f_i} = |\text{SROCC}_{f_i}| - |\text{SROCC}_{\text{SQMS}}| \quad (۷)$$

که $i \in \{1, \dots, 14\}$ و SROCC_{f_i}، ضریب همبستگی اِسْپیرمَنِ محاسبه‌شده برای SVRی است که روی f_i آموزش دیده باشد. همانطور که در قسمت ۱.۲ گفته شد، این ضریب همبستگی، بین نمرات الگوریتم و نمرات انسانی محاسبه می‌شود. $\text{SROCC}_{\text{SQMS}}$ هم دقتِ $SQMS(.,.)$ نشان می‌دهد. می‌بینیم که تمام مدل‌ها عملکرد بهتری نسبت به $SQMS(.,.)$ داشته‌اند.

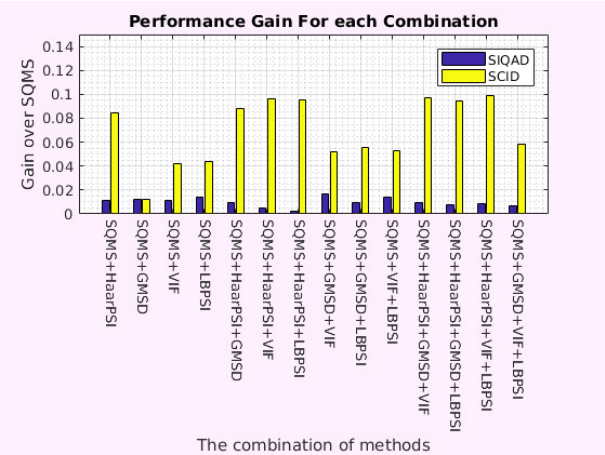
۳.۳ عدم تعمیم‌پذیری به تخریب‌های مختلف

با افزایش دقتِ $SQMS(.,.)$، روی کُلِّ مجموعه‌داده، انتظار می‌رود که ترکیب‌ها بهبودی مشابه را روی تک تکِ تخریب‌ها نیز داشته باشند. میزان دقت مدل‌ها روی تخریب‌ها طبق ۱.۲.۲ محاسبه گردیده و بهره‌وری آن نسبت به $SQMS(.,.)$ در شکل ۵ نشان داده می‌شود. می‌بینیم که نه تنها همه‌ی مدل‌ها عملکرد مشابهی نداشتند، بلکه در بسیاری از موارد، بهره‌وری منفی بوده است. کاملاً منطقی است که یکی از علّت‌ها را ضعف بردارهای ویژگی بدانیم. فارغ از این نظریه، مسئله‌ی دیگری وجود دارد که باید در نظر گرفته شود و آن *فراعامل‌ها* هستند.

همانطور که در انتهای ۲.۲ گفته شد، بهینه‌سازیِ c و γ با اعتبارسنجیِ متقابلِ ۱۰۰ـ لایه صورت می‌گیرد:

$$(c^*, \gamma^*) = \text{opt}(r_c, r_\gamma, \text{dataset}_{80,\text{all}}, \text{dataset}_{20,\text{all}}, 100) \quad (۸)$$

رابطه‌ی ۸ مواردی که باید در نظر گرفته شوند را خلاصه می‌کند. c^* و γ^* فراعامل‌های بهینه هستند. تابع opt نتایج بهینه‌سازی را برمی‌گرداند.



شکل ۴: بهره‌وری هر یک از روش‌های ترکیبی نسبت به $SQMS(.,.)$. برای هر ترکیب، بهره‌وریِ عملکرد بر روی دو مجموعه‌داده، با دو رنگ متفاوت نشان داده شده است.

r_c و r_γ، به ترتیب، مجموعه‌ی مقادیری هستند که برای c و γ امتحان می‌شوند. مثلاً اگر $r_c = \{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\}$ باشد، بهترین مقدار برای c از این مجموعه پیدا می‌شود. مقادیر مورد بررسی به صورت دستی انتخاب شده و تعدادشان به توان پردازشی در دسترس بستگی دارد. در هر بار آموزش و ارزیابی، نیاز به یک مجموعه‌ی آموزش و یک مجموعه‌ی آزمون داریم.

طبق رابطه‌ی۸، مجموعه‌ی آموزش، $\text{dataset}_{80,\text{all}}$ است. این عبارت یک زیرمجموعه از نمونه‌های مجموعه‌داده‌ی dataset را مشخص می‌نماید. این زیرمجموعه، عبارت است از **همه‌ی** (all) تخریب‌های موجود در زیرمجموعه‌ی آموزشِ افرازِ ۸۰-۲۰. برخی نمونه‌ها از این عبارت، می‌توانند $\text{SIQAD}_{80,\text{JPEG}}$ و $\text{SCID}_{20,\text{GB}}$ باشند. که اوّلی، یعنی تمامیِ نمونه‌های SIQAD؛ که در افرازِ ۸۰-۲۰، در مجموعه‌ی آموزش قرار گرفته‌اند؛ و تخریبِ آن‌ها از نوعِ JPEG است. دومی، نمونه‌هایی از SCID را مشخص می‌کند که در مجموعه‌ی آزمون قرار گرفته‌اند و تخریبِ آن‌ها از نوعِ GB[۳۶] است. آخرین ورودیِ تابعِ opt هم، تعداد لایه‌های (دفعاتِ) اعتبارسنجیِ متقابل است.

اگر به جای $\text{opt}(r_c, r_\gamma, \text{SIQAD}_{80,all}, \text{SIQAD}_{20,all}, 100)$، از $\text{opt}(r_c, r_\gamma, \text{SIQAD}_{80,all}, \text{SIQAD}_{20,\text{JPEG}}, 100)$ استفاده کنیم، نتایج متفاوت خواهند بود (شکل ۶). قبل از بررسیِ نتایج، یک رابطه مانند ۸ برای اعتبارسنجیِ متقابل قرارداد می‌کنیم، که نقشِ فراعامل‌ها را نیز در نظر بگیرد:

$$\begin{aligned} &= \text{دقّت الگوریتم} \\ &cv(\text{dataset}_{80,\text{all}}, \\ &\text{opt}(r_c, r_\gamma, \text{dataset}_{80,\text{all}}, \text{dataset}_{20,\text{all}}, 100), \\ &\text{dataset}_{20,\text{all}}, 1000) \end{aligned} \quad (۹)$$

تابعِ $cv(.,.,.,.)$، چهار ورودی دارد که عبارت‌اند از: مجموعه‌ی آموزش، فراعامل‌های بهینه، مجموعه‌ی آزمون و تعداد لایه‌های اعتبارسنجیِ متقابل. این تابع میانه‌ی SROCCهای محاسبه‌شده در ۱۰۰۰ مرتبه را بر می‌گرداند.

[۳۶] نوعی تار شدگی تصویر. برای شرح سایر تخریب‌ها به مقاله‌ی خودِ مجموعه‌داده‌ها رجوع شود.

شکل ۴ نتایج را برای آزمایش زیر نشان می‌دهد:

$$\begin{aligned} &cv(\text{dataset}_{80,\text{all}}, \\ &\text{opt}(r_c, r_\gamma, \text{dataset}_{80,\text{all}}, \text{dataset}_{20,\text{all}}, 100), \\ &\text{dataset}_{20,\text{all}}, 1000) \end{aligned} \quad (۱۰)$$

که dataset در هر بار، یا SIQAD و یا SCID است. برای ارزیابی روی هر یک از تخریب‌ها (شکل ۵)، رابطه (۱۰) به شکل زیر تغییر می‌کند:

$$\begin{aligned} &cv(\text{dataset}_{80,\text{all}}, \\ &\text{opt}(r_c, r_\gamma, \text{dataset}_{80,\text{all}}, \text{dataset}_{20,\text{all}}, 100), \\ &\text{dataset}_{20,\text{dst}}, 1000) \end{aligned} \quad (۱۱)$$

که dataset مطابقِ (۱۰) و $\text{dst} \in \{\text{تخریب‌های موجود در dataset}\}$. اگر در آزمایشی که رابطه (۱۱) نشان می‌دهد، فراعامل‌ها را روی تخریب‌های متناظر بهینه کنیم، نتایجِ شکل ۶ حاصل می‌شوند. به طور رسمی، نتایج شکل ۶، حاصلِ آزمایشِ زیر هستند:

$$\begin{aligned} &cv(\text{dataset}_{80,\text{all}}, \\ &\text{opt}(r_c, r_\gamma, \text{dataset}_{80,\text{all}}, \text{dataset}_{20,\text{dst}}, 100), \\ &\text{dataset}_{20,\text{dst}}, 1000) \end{aligned} \quad (۱۲)$$

پس می‌بینیم که علاوه بر تأثیرِ احتمالیِ بردارهای ویژگی، فراعامل‌ها نیز نقش مهمی در عملکرد SVRها دارند.

۴.۳ عدم تعمیم‌پذیری به سایر مجموعه‌داده‌ها

برای بررسی عدمِ وابستگیِ یک مدل به صحنه‌های یک مجموعه‌داده، آن را روی یک مجموعه‌داده آموزش داده و روی مجموعه‌داده‌ی دیگر می‌آزمایند:

$$\begin{aligned} &cv(\text{SIQAD}_{\text{all,all}}, \\ &\text{opt}(r_c, r_\gamma, \text{SIQAD}_{80,\text{all}}, \text{SIQAD}_{20,\text{all}}, 100), \\ &\text{SCID}_{\text{all,all}}, 1) \end{aligned} \quad (۱۳)$$

$$\begin{aligned} &cv(\text{SCID}_{\text{all,all}}, \\ &\text{opt}(r_c, r_\gamma, \text{SCID}_{80,\text{all}}, \text{SCID}_{20,\text{all}}, 100), \\ &\text{SIQAD}_{\text{all,all}}, 1) \end{aligned} \quad (۱۴)$$

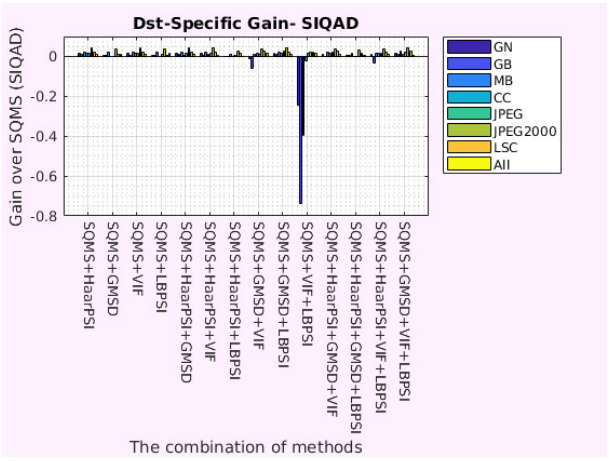
شکل ۷ نتایج را برای آزمایش‌های (۱۴) و (۱۳) نشان می‌دهد.

می‌بینیم که بهبودِ مشاهده شده در آزمایش (۱۰) (با نتایجِ قابلِ مشاهده در شکل ۴) برای SIQAD اتفاق نمی‌افتد. به غیر از این مسئله، اگر مدلِ آموزش‌دیده را برای تک تکِ تخریب‌ها بیازمائیم، باز هم با اُفْتِ عملکردی مشابهِ قسمتِ۳.۳ مشابه می‌شویم. بیان رسمیِ این آزمایش‌ها در روابط (۱۵) و (۱۶) و نتایج آن‌ها در شکل ۸ ارائه شده‌اند.

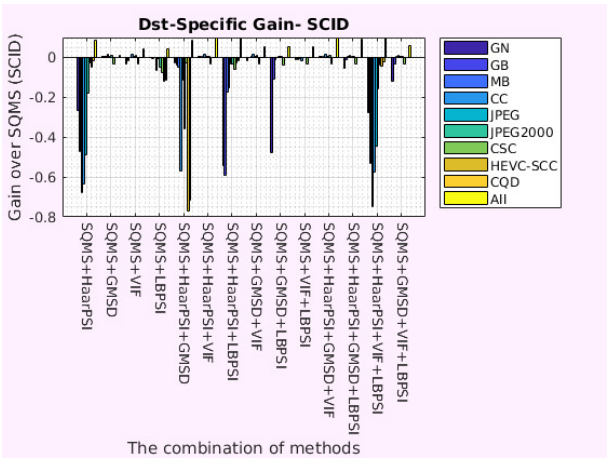
$$\begin{aligned} &cv(\text{SCID}_{\text{all,dst}}, \\ &\text{opt}(r_c, r_\gamma, \text{SCID}_{80,\text{all}}, \text{SCID}_{20,\text{all}}, 100), \\ &\text{SIQAD}_{\text{all,dst}}, 1) \end{aligned} \quad (۱۵)$$

$$\begin{aligned} &cv(\text{SIQAD}_{\text{all,dst}}, \\ &\text{opt}(r_c, r_\gamma, \text{SIQAD}_{80,\text{all}}, \text{SIQAD}_{20,\text{all}}, 100), \\ &\text{SCID}_{\text{all,dst}}, 1) \end{aligned} \quad (۱۶)$$

dst در روابط بالا، عضوِ مجموعه‌ی تخریب‌هایی است که در SCID و SIQAD مشترک هستند. می‌بینیم که نتایج بهترِ آزمون تعمیم‌پذیری برای مجموعه‌ی SCID هم به تک تکِ تخریب‌های آن قابل تعمیم نیست.

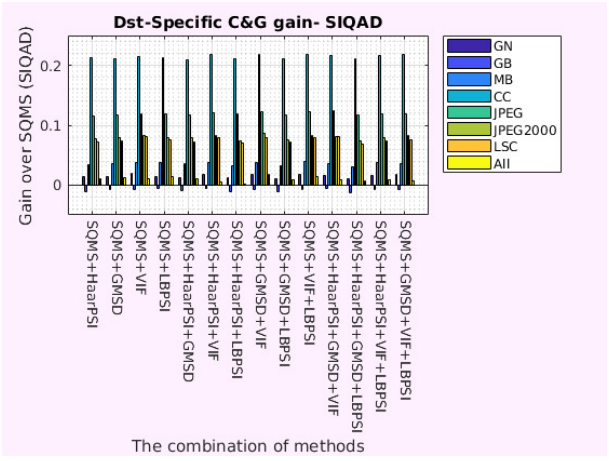


(آ)

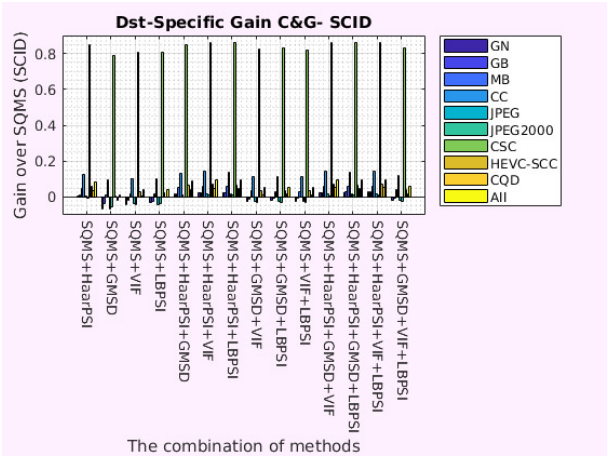


(ب)

شکل ۵: نتیجه‌ی ارزیابی به ازای هر تخریب. نتایج SIQAD در (آ) و SCID در (ب) قابل ملاحظه هستند. بهره‌وری به ازای هر تخریب، با رنگ متفاوتی مشخص شده است.

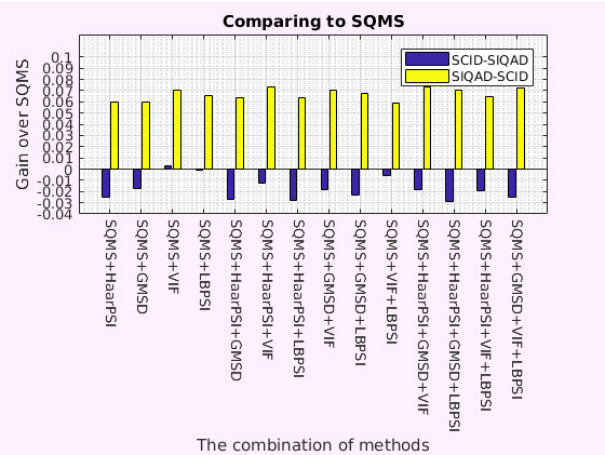


(آ)



(ب)

شکل ۶: نتیجه‌ی ارزیابی به ازای هر تخریب بعد از بهینه‌سازی فراعامل‌ها روی هر یک از تخریب‌ها. نتایج SIQAD در (آ) و SCID در (ب) قابل ملاحظه هستند.



شکل ۷: نتایج آزمونِ تعمیم‌پذیری. ستون‌های زرد مربوط به آموزش روی SIQAD و آزمون روی SCID هستند. ستون‌های سُرمه‌ای مربوط به آموزش و آزمونِ عکس هستند.

۴ جمع‌بندی

در این مقاله بردارهای پشتیبان را، طبق روش رایج، برای اَکْثْ آموزش دادیم. دیدیم که وقتی گزارش عملکرد به روش رایج انجام می‌شود، نقش فراعامل‌ها هم باید بازتاب گردد. یک فرمول‌بندی ارائه شد که این موارد را لحاظ نماید. تأثیرِ بهینه‌سازیِ فراعامل‌ها را مشاهده کردیم و دیدیم که به این مسئله در آزمایش‌های ارزیابی کیفیت پرداخته نمی‌شود. محدود بودنِ تصاویر مجموعه‌داده‌ها، نگرانی در موردِ تعمیم‌پذیری مدل‌ها به تصاویر دنیای واقعی را بیشتر می‌کند. مخصوصاً وقتی که تخریب‌های مجموعه‌داده‌ها مصنوعی بوده و ممکن است با تخریب‌های طبیعی متفاوت باشند. شاید بهتر باشد که نحوه‌ی بهینه‌سازیِ فراعامل‌های مدل‌های یادگیری نیز در گزارش‌های عملکرد مقالات تشریح گردد. آزمایش‌های انجام‌شده با کُدهای موجود در https://github.com/cheraaqee/fusion_iqa قابلِ شبیه‌سازی هستند.

مراجع

[1] CISCO, "Cisco annual internet report," `https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html`, 2022. Accessed: 2022-July-15.

[2] E. Allen and S. Triantaphillidou. *The manual of photography*. CRC Press, 2012.

[3] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *arXiv preprint arXiv:1406.7799*, 2014.

[4] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol.63, no.11, pp.1–52, 2020.

[5] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol.19, no.1, p.011006, 2010.

[6] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal processing*, vol.70, no.3, pp.177–200, 1998.

[7] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol.26, no.1, pp.98–117, 2009.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol.13, no.4, pp.600–612, 2004.

[9] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol.15, no.11, pp.3440–3451, 2006.

[10] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," in *Advances in neural information processing systems*, pp.551–558, 1994.

[11] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol.15, no.2, pp.430–444, 2006.

[12] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal processing magazine*, vol.34, no.6, pp.130–141, 2017.

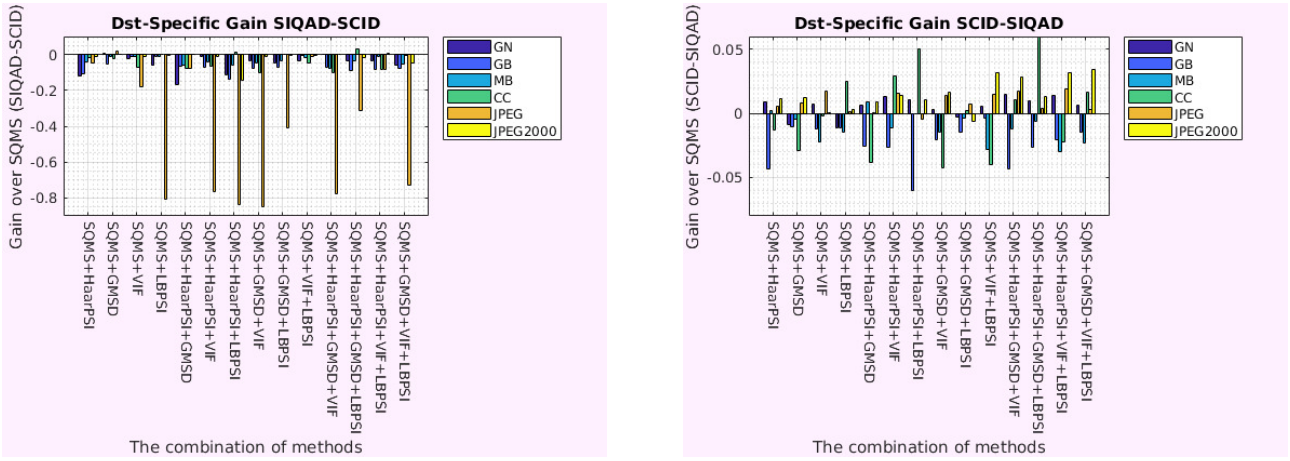
[13] X. Yang, F. Li, and H. Liu, "A survey of dnn methods for blind image quality assessment," *IEEE Access*, vol.7, pp.123788–123806, 2019.

[14] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE conference on computer vision and pattern recognition*, pp.1098–1105, IEEE, 2012.

[15] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[۱۶] پوریا چراغی، "ارائه‌ی یک روش بدون مرجع برای ارزیابی کیفیت تصاویر با تخریب چندگانه،" پایان‌نامه کارشناسی‌ارشد، گروه مهندسی برق و کامپیوتر دانشگاه خوارزمی، شهریور ۱۳۹۸.

[17] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *International Scholarly Research Notices*, vol.2013, 2013.



(ب) (آ)

شکل ۸: نتایجِ ارزیابی هر تخریب، وقتی مجموعه‌داده‌های آموزش و آزمون متفاوت‌اند. (آ): برای SIQAD → SCID و (ب) برای SCID → SIQAD

[25] A. Alaei, D. Conte, M. Blumenstein, and R. Raveaux, "Document image quality assessment based on texture similarity index," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pp.132–137, IEEE, 2016.

[26] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol.18, no.6, pp.1098–1110, 2016.

[27] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A haar wavelet-based perceptual similarity index for image quality assessment," *Signal Processing: Image Communication*, vol.61, pp.33–43, 2018.

[28] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE transactions on image processing*, vol.23, no.2, pp.684–695, 2013.

[29] A. Shokrollahi, B. M.-N. Maybodi, and A. Mahmoudi-Aznaveh, "Histogram modification based enhancement along with contrast-changed image quality assessment," *Multimedia Tools and Applications*, pp.1–22, 2020.

[30] K. Okarma, P. Lech, and V. V. Lukin, "Combined full-reference image quality metrics for objective assessment of multiply distorted images," *Electronics*, vol.10, no.18, p.2256, 2021.

[18] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol.23, no.11, pp.4850–4862, 2014.

[19] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.1733–1740, 2014.

[20] H. Yang, Y. Fang, and W. Lin, "Perceptual Quality Assessment of Screen Content Images," *IEEE Transactions on Image Processing*, vol.24, no.11, pp.4408–4421, 2015.

[21] A. Mittal, *Natural scene statistics-based blind visual quality assessment in the spatial domain*. Phd thesis, The University of Texas at Austin, 2013.

[22] P. Ye and D. Doermann, "Document image quality assessment: A brief survey," in *2013 12th International Conference on Document Analysis and Recognition*, pp.723–727, IEEE, 2013.

[23] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K. K. Ma, "ESIM: Edge Similarity for Screen Content Image Quality Assessment," *IEEE Transactions on Image Processing*, vol.26, no.10, pp.4818–4831, 2017.

[24] Y. Zhang, D. M. Chandler, and X. Mou, "Quality Assessment of Screen Content Images via Convolutional-Neural-Network-Based Synthetic/Natural Segmentation," *IEEE Transactions on Image Processing*, vol.27, no.10, pp.5113–5128, 2018.