طبقه‌بندی بدون‌نظارت تصاویر طبیعی با استفاده از یادگیری عمیق

گزارش پروژه پایانی درس یادگیری عمیق

تهیه کننده: حمیدرضا نادمی

چکیده‌ـ امروزه با پیشرفت تکنولوژی با افزایش حجم انواع داده‌ها در بسترهای مختلف برچسب‌گذاری نمونه‌ها جهت آموزش یک مدل عمیق کاری زمانبر، پرهزینه و محتمل بروز خطا می‌باشد، از طرفی مدل‌هایی که بصورت نظارتی (همراه با برچسب هر نمونه) ساخته می‌شوند نسبت به مدل‌هایی عمیق که بدون‌نظارت کار می‌کنند در مقیاس کوچکتری قابل استفاده هستند. از این رو رویکردهای آموزش مدل عمیق بصورت بدون‌نظارت در سال های اخیر در بینایی کامپیوتر توجه زیادی را به خود جلب کرده است. در این پژوهش یک رویکرد یادگیری بدون‌نظارت که در سال ۲۰۱۸ در ICLR به چاپ رسیده است را مورد بررسی قرار دادیم و با الگوریتم پیشنهادی در مقاله مجموعه تصاویر از ۱۳ نوع گل مختلف را طبقه‌بندی کردیم و نتایج بدست آمده را با نتایج طبقه‌بندهایی که بصورت نظارتی(همراه با برچسب مجموعه‌داده) مجموعه‌داده را طبقه‌بندی کردند تحلیل و مقایسه کردیم.

کلمات کلیدی- یادگیری بدون‌نظارت، طبقه‌بندی، تصاویر طبیعی، یادگیری عمیق.

۱ . مقدمه

امروزه با تولید حجم انبوه از انواع داده‌ها در بسترهای مختلف مانند شبکه‌های مجازی، دوربین‌های مداربسته در سطح شهر و همچنین تولید پردازنده‌های قدرتمند و افزایش توانایی محاسباتی ماشین‌ها باعث شده تا استفاده از روش‌های مبتنی یادگیری عمیق بیشتر از قبل در پژوهش و صنعت مورد توجه واقع شود. بطور کلی ۴ رویکرد جهت آموزش مدل به منظور انجام کارهای رایج در بینایی کامپیوتر مانند طبقه‌بندی تصاویر، خوشه‌بندی، تقسیم‌بندی معنایی تصویر و... وجود دارد [1] که عبارتند از: یادگیری نظارتی، یادگیری نیمه نظارتی، یادگیری با نظارت ضعیف[۸] و یادگیری بدون‌نظارت. در سه رویکرد اول جهت آموزش مدل نیاز به برچسب داده‌ها داریم (بستگی به رویکرد انتخابی همه یا بخشی از برچسب مجموعه داده در آموزش مدل استفاده می‌شود). مدل‌های عمیق جهت دست‌یابی به عملکرد مناسب، نیاز به حجم انبوه از داده‌های برچسب گذاری شده دارند و از طرفی جمع‌آوری و برچسب گذاری داده‌ها کاری زمان‌بر و پرهزینه است [1] ، از این رو انگیزه جهت بکار بردن رویکردهای یادگیری بدون‌نظارت و زیر مجموعه آن یادگیری خودنظارتی در حل مسائل بینایی کامپیوتر افزایش یافته است.

در یادگیری خودنظارتی بدون استفاده از برچسب واقعی هر داده، با استفاده از اطلاعاتی که در خود داده وجود دارد یک برچسب مصنوعی (سیگنال نظارتی) توسط مدل تولید می‌شود. مدل (که یک شبکه عصبی عمیق می‌باشد) با بررسی ساختار تصویر ورودی به شبکه و حل یک مسئله از پیش تعریف شده[۱۱] روی تصویر ورودی، سیگنالی نظارتی جهت به روز کردن پارامترهای خود از تصویر دریافت می‌کند.

در این پژوهش مقاله چاپ شده توسط گیداریس و همکاران [2] در کنفرانس ICLR 2018 به‌روی مجموعه تصاویری از ۱۳ نوع گل مختلف اجرا شده است. در بخش۲ به بررسی کارهای انجام شده اخیر در زمینه یادگیری بدون‌نظارت و یادگیری خودنظارت

می‌پردازیم، در بخش۳ به شرح مقاله مورد بررسی و نحوه کار شبکه طراحی شده می‌پردازیم، در بخش۴ نتایج بدست آمده از انجام آزمایش‌ها گزارش شده است و در بخش۵ نتایج حاصل شده در پژوهش را جمع‌بندی و نتیجه‌گیری کرده‌ایم.

۲ . کارهای پیشین

بطور کلی کارهای پیشین در زمینه یادگیری بدون‌نظارت را می‌توان به دو دسته یادگیری خودنظارتی و یادگیری بدون‌نظارت تقسیم کرد.

۲-۱- *یادگیری خودنظارتی*

همانطور که گفته شد در یادگیری خودنظارتی یک سیگنال نظارتی از داده تولید می‌شود، این سیگنال نظارتی می‌تواند با حل یک مسئله از پیش تعریف شده توسط شبکه تولید شود. برای مثال تبدیل تصویر سیاه و سفید به رنگی و برعکس [3]، پیش‌بینی بخش حذف شده از تصویر [4] و یا در نظرگرفتن تصویر مانند یک پازل و حل پازل توسط شبکه [5] از جمله مثال‌هایی از این قبیل مسائل هستند و شبکه با استفاده از این سیگنال وزن‌های خود را به‌روز می‌کند.

۲-۲- *یادگیری بدون‌نظارت*

در یادگیری *بدون‌نظارت* آموزش مدل بدون در نظر گرفتن برچسب واقعی مجموعه داده و تولید هرگونه سیگنال نظارتی انجام می‌شود. در سال ۲۰۱۹ دوون و همکاران [6] رویکردی بدون‌نظارت جهت یادگیری نمایشی مناسب از داده‌ها ارائه دادند. روش پیشنهادی مبتنی بر حداکثر کردن اطلاعات متقابل بین نقشه ویژگی و بردار ویژگی تصویر ورودی به شبکه رمزگذار می‌باشد. نقشه ویژگی تصویر با اعمال یک سری کانولوشن روی تصویر بدست می‌آید و بردار ویژگی هر تصویر خروجی نهایی شبکه رمزگذار (که توسط لایه‌های FC شبکه تولید می‌شود) می‌باشد. جهت حداکثر کردن اطلاعات متقابل بین ورودی و خروجی شبکه رمزگذار از یک شبکه متمایزکننده استفاده شده است که نقشه ویژگی و بردار ویژگی هر تصویر را دریافت می‌کند. همانند شبکه‌های متمایزکننده در شبکه‌های مولد زوج‌های مثبت و منفی به شبکه داده می‌شود، که زوج مثبت نقشه ویژگی و بردار ویژگی بدست آمده از تصویر ورودی و زوج منفی بردار ویژگی تصویر و نقشه ویژگی تصویری متفاوت می‌باشد. شبکه متمایزکننده با تشخیص زوج‌های مثبت و منفی وزن‌های خود و شبکه رمزگذار را به‌روز می‌کند. تابع هزینه در نظر گرفته شده برای شبکه طراحی شده کران پایین واگرایی KL و با حداکثر کردن آن اطلاعات متقابل بین نقشه ویژگی تصویر و بردار ویژگی حداکثر می‌شود.

۳ . روش مورد بررسی

گیداریس و همکاران [2]رویکردی خودنظارتی جهت آموزش شبکه ارائه کردند. ابتدا به شرح ایده بکار رفته پرداخته و سپس مزیت‌های انتخاب تبدیل هندسی چرخش به عنوان یک سیگنال نظارتی برای آموزش شبکه را بررسی می‌کنیم.

۳-۱- *شرح ایده مقاله مورد بررسی*

در روش پیشنهادی توسط گیداریس و همکاران [2] روی هر تصویر تبدیل هندسی چرخش به زاویه‌های ۰، ۹۰، ۱۸۰ و ۲۷۰ زده می شود و شبکه با تشخیص زاویه چرخش اعمال شده روی تصویر آموزش داده می‌شود (شکل۱). رابطه (۱) تابع هزینه در نظرگرفته شده برای شبکه می‌باشد.

$$loss(X_i.\theta) = -\frac{1}{K}\sum_{y=1}^{K}\log(F^y(g(X_i|y)|\theta)) \qquad (۱)$$

در رابطه (۱) $g(.|y)$ تبدیل هندسی اجرا شده روی تصویر می‌باشد و y برچسب تصویر می‌باشد که مشخص می‌کند تبدیل با چه زاویه‌ایی روی تصویر اعمال شده است، $F^y(.|\theta)$ یک شبکه عمیق کانولوشنی می‌باشد که دارای وزن θ است و وزن‌های شبکه با مجموعه‌داده ورودی آموزش دیده می‌شوند.

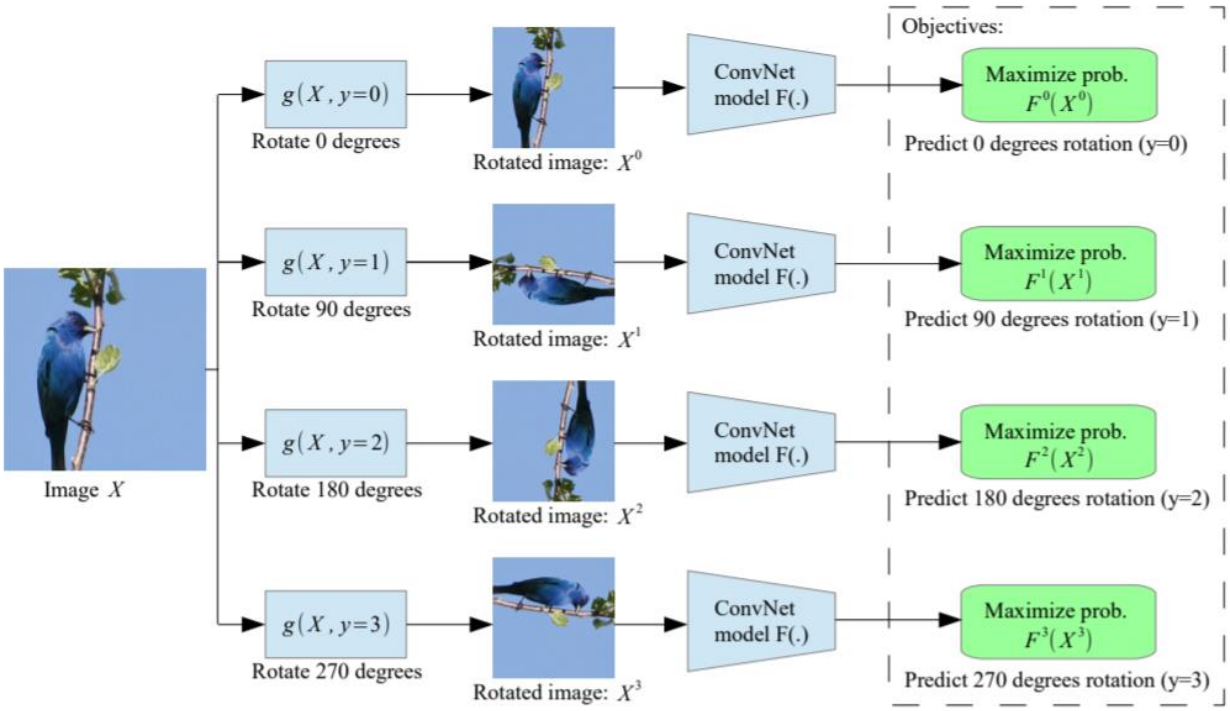
اگر $Rot(X.\phi)$ عملگری باشد که تصویر X را به اندازه ϕ بچرخاند، طبق رابطه (۲) از هر تصویر ۴ تصویر تولید می‌شود.

$$G = \{g(X|y)\}_{y=1}^{4}$$
$$g(X|y) = Rot(X.(y-1)*90) \qquad (۲)$$

۳-۲- *مزیت انتخاب تبدیل هندسی چرخش*

شبکه با تلاش برای تشخیص زاویه تبدیل چرخشی که روی تصویر انجام شده (۰، ۹۰، ۱۸۰ و یا ۲۷۰ درجه) تلاش می‌کند ویژگی‌های معنایی از تصویر را استخراج کند انجام این کار ساده به شبکه کمک می‌کند تا ساختار کُلی هر تصویر را یادبگیرد و در نتیجه شبمه در کارهایی مانند طبقه‌بندی، تشخیص شئ و تقسیم‌بندی معنایی مجموعه‌داده عملکرد خوبی داشته باشد.

تبدیل هندسی چرخش بر خلاف دیگر تبدیل‌های هندسی مانند مقیاس باعث از بین رفتن ویژگی‌های واضح در تصویر نمی‌شوند، ویژگی‌هایی که به راحتی توسط شبکه قابل آموزش هستند. برای مثال در تبدیل هندسی مقیاس با تغییر اندازه تصویر ممکن است ویژگی‌هایی که در تصویر با اندازه اصلی وجود دارد از بین برود و در نتیجه شبکه ویژگی‌های واضح در تصویر را آموزش نبیند.



شکل۱: استخراج ویژگی معنایی از هر تصویر با انجام چرخش با زاویه‌های ۰، ۹۰، ۱۸۰ و ۲۷۰ روی تصویر

۴ . نتایج ارزیابی

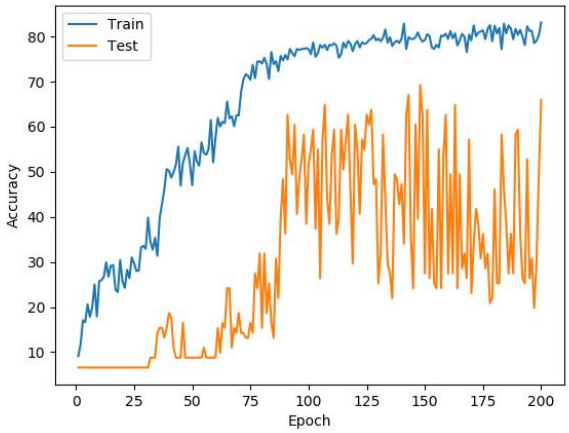
در این بخش ابتدا به توضیح مجموعه‌داده مورد آزمایش و نحوه تقسیم‌بندی مجموعه‌داده به نمونه آموزشی و نمونه آزمون می‌پردازیم و در ادامه نتایج بدست آمده از آزمایش‌ها را گزارش می‌کنیم.

۴-۱ مجموعه داده

مجموعه داده مورد بررسی شامل ۴۴۹ تصویر از ۱۳ نوع گل می‌باشد و ۸۰ درصد از داده‌ها جهت آموزش مدل و ۲۰ درصد به‌عنوان داده آزمون در نظرگرفته شده است. همه نمونه‌ها به ابعاد ۱۳۰*۱۳۰ نرمال شدند.

۴-۲ نتایج

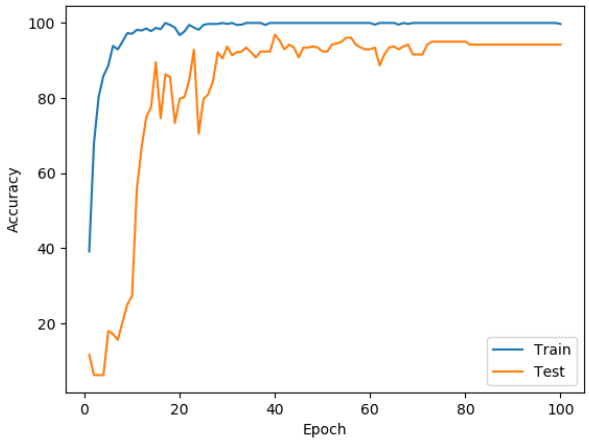
در جدول۱ نتایج ارزیابی بدست آمده روی نمونه‌های آزمون آورده شده است. جهت آموزش مدل بصورت بدون‌نظارت (با الگوریتم توضیح داده شده در بخش۳) از شبکه AlexNet استفاده شده است. جهت طبقه‌بندی نمونه‌ها از یک شبکه MLP ۳ لایه (پیشفرض مورد استفاده در مقاله) استفاده شده است. علاوه بر طبقه‌بندی نمونه‌ها بصورت بدون‌نظارت با روش پیشنهادی در پژوهش گیداریس و همکاران [2]، طبقه‌بندی بصورت نظارتی با معماری شبکه‌های VGG، AlexNet و GoogleNet نیز انجام شده است. شبکه VGG مورد استفاده از قبل رو مجموعه‌داده ImageNet آموزش داده شده است. وزن شبکه‌های AlexNet و Network in Network در رویکردهای نظارتی و بدون‌نظارت با نمونه های آموزشی تنظیم شده اند. در شکل۲ و شکل۳ دقت مدل در هرگام برای شبکه‌های AlexNet و GoogleNet نمایش داده شده است.



شکل۲: دقت در هرگام در شبکه AlexNet بصورت بدون‌نظارت

جدول ۱: **دقت طبقه‌بندی نمونه‌های آزمون با روش‌های نظارتی و بدون‌نظارتی**

روش	با/ بدون نظارت	Top-1	Top-5	تعداد گام آموزش
گیداریس و همکاران (۲۰۱۸) [2] (AlexNet)	بدون‌نظارت	**۶۹/۲**	**۱۰۰**	۲۰۰
گیداریس و همکاران (۲۰۱۸) [2] (AlexNet)	بدون‌نظارت	۶۳/۷	۹۸/۹	۱۰۰
گیداریس و همکاران (۲۰۱۸) [2] (GoogleNet)	بدون‌نظارت	-	-	-
AlexNet	بانظارت	۹۵/۶	۱۰۰	۱۰۰
GoogleNet	بانظارت	۹۳/۷	۱۰۰	۵۰
GoogleNet	بانظارت	**۹۶/۸**	**۱۰۰**	۱۰۰
VGG+MLP	بانظارت	۸۹	–	۵۰
VGG+Linear SVM	بانظارت	۹۵	–	۵۰



شکل۳: دقت در هرگام در شبکه GoogleNet بصورت بانظارت

با توجه به نتایج بدست آمده در جدول۱ بهترین دقت (Top-1) در رویکرد بدون‌نظارت ۶۹/۲ و در رویکردهای نظارتی ۹۶/۸ حاصل شد. در مقایسه با بالاترین پنج پیش‌بینی مدل رویکرد بدون‌نظارتی مورد بررسی به نتیجه ایی تقریبا به‌خوبی رویکرد نظارتی دست یافت. لازم به ذکر است به دلیل ناکافی بودن قدرت پردازنده‌های سیستم مورد استفاده جهت اجرای آزمایش‌ها، رویکرد بدون‌نظارت گیداریس و همکارن [2] با معماری GoogleNet اجرا نشد.

۵ . نتیجه‌گیری

در این پژوهش یک رویکرد بدون‌نظارت جهت آموزش مدل عمیق مورد بررسی قرار گرفت، علی رقم سادگی ایده ارئه شده در مقاله مورد بررسی، شاهد عملکرد قابل قبول مدل در دقت بدست آمده با اولین پیش‌بینی مدل (Top-1) و عملکرد تقریبا برابر با پنج پیش‌بینی اول مدل (Top-5) در مقایسه با مدل‌هایی که بصورت نظارتی آموزش داده شده بودند در طبقه‌بندی گُل‌ها بودیم.

مراجع

[1] Longlong Jing and Yingli Tian; “Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey.” TPAMI (major revision), 2019.

[2] Spyros Gidaris, Praveer Singh and Nikos Komodakis; “Unsupervised Representation Learning by Predicting Image Rotation.” ICLR, 2018.

[3] Richard Zhang, Phillip Isola, and Alexei A Efros. “Colorful image colorization.” In *European Conference on Computer Vision*, pp. 649–666. Springer, 2016.

[4] Carl Doersch, Abhinav Gupta, and Alexei A Efros. “Unsupervised visual representation learning by context prediction.” In *Proceedings of the IEEE International Conference on Computer Vision*, pp.1422–1430, 2015.

[5] Mehdi Noroozi and Paolo Favaro. “Unsupervised learning of visual representations by solving jigsaw puzzles.” In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

[6] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, and Y. Bengio, “Learning deep representations by mutual information estimation and maximization,” International Conference on Learning Representation, 2019.