شبکه های عصبی و یادگیری عمیق

تهمینه توکلی

تیر ۱۴۰۳

۱ مقدمه

در این پروژه، رویکرد جدیدی برای بهبود عملکرد طبقه بندها با استفاده از ایده‌های دو مقاله تاثیرگذار بررسی می‌شود: "یک روش یادگیری ویژگی تفکیکی برای تشخیص چهره عمیق" و "دشمن دشمن من دوست من است: بررسی دشمنان معکوس برای بهبود آموزش مقاومتی". هدف اصلی، بررسی این است که آیا می‌توان از دشمنان معکوس به جای مراکز داده ها در تابع زیان استفاده کرد تا قدرت تفکیکی مدل‌های طبقه بند را افزایش داد.

۲ پیش‌زمینه

۱.۲ تشخیص چهره عمیق با استفاده از center loss

مقاله "یک روش یادگیری ویژگی تفکیکی برای تشخیص چهره عمیق" center loss را معرفی می‌کند که هدف آن افزایش قدرت تفکیکی ویژگی‌های یاد گرفته شده توسط شبکه‌های عصبی عمیق است. این تابع زیان فاصله بین ویژگی‌های یک کلاس و مرکز آن کلاس در فضای ویژگی‌ها را کاهش می‌دهد. منظور از فضای ویژگی‌ها، خروجی لایه ماقبل آخر مدل است که توسط feature extractor یاد گرفته شده است. این کاهش واریانس درون کلاسی به بهبود استحکام و دقت مدل کمک می‌کند.

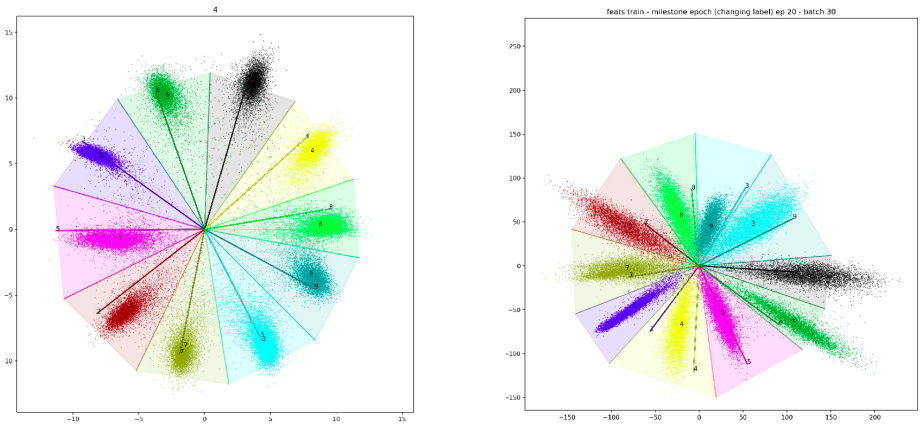
$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

در معادله ۱، x_i نشان دهنده ویژگی iام و متعلق به کلاس y_i است و c_{y_i} مرکز کلاس y_i در فضای ویژگی‌ها را نشان میدهد. مرکزها در طول یادگیری به روزرسانی می‌شوند.

شکل ۱ مربوط به دو مدل تعلیم شده تحت نظارت cross entorpy loss (سمت راست) و center loss (سمت چپ) است. با مقایسه دو تصویر در شکل ۲، متوجه میشویم که در صورتی که فقط از تابع زیان cross entorpy استفاده کنیم، ویژگی‌های عمیق آموخته شده توسط feature extractor دارای واریانس درون کلاسی بزرگی خواهند بود. همچنین برای هر کلاس بردار وزن الزاما با میانگین ویژگی های آن کلاس منطبق نیست.

در صورتی که ویژگی استخراج شده توسط مدل برای یک داده ورودی، دقیقا در امتداد بردار وزن کلاس مربوطه قرار بگیرد، خروجی احتمالی مدل برای این کلاس بیشترین مقدار خواهد بود و داده با اطمینان بالا به درستی طبقه بندی میشود. در واقع، امتداد بردار وزن هر کلاس نشان دهنده مناطقی است که با احتمال زیاد متعلق به همان کلاس طبقه بندی می‌شوند (high likelihood regions).

با افزودن جمله center loss به تابع زیان، مشکل واریانس درون کلاسی تا حد خوبی برطرف شده است (تصویر سمت چپ). اما همچنان قابل مشاهده است که محل قرارگیری میانگین ویژگی های هرکلاس با بردار وزن مربوطه فاصله دارد. این مشکل در ابعاد بالاتر شدیدتر خواهد بود.



شکل ۱: توزیع ویژگی های آموخته شده تحت نظارت cross entorpy loss (سمت راست) و center loss (سمت چپ). نقاط با رنگ های مختلف نشان دهنده ویژگی ها از کلاس های مختلف است. بردار وزن هر کلاس و میانگین ویژگی های هر کلاس مشخص شده است.

استخراج ویژگی دقیقا در امتداد بردار وزن لزوما برای feature extractor قابل انجام نیست. هدف این پروژه بررسی این است که high likelihood regions در ابعاد بالاتر در چه مکانی قرار دارند به صورتی که برای feature extractor تولید ویژگی در این مکان قابل انجام باشد. آیا high likelihood regions در مکان بردار وزن است یا میانگین ویژگی ها و یا مکانی دیگر؟

۲.۲ Inverse adversarial examples برای آموزش مقاومتی

مقاله "دشمن دشمن من دوست من است: بررسی دشمنان معکوس برای بهبود آموزش مقاومتی"، مفهوم inverse adversarial examples را معرفی می‌کند که هدف آن افزایش قدرت تفکیک مدل است. این تابع زیان فاصله بین ویژگی‌های یک کلاس و نمونه های inverse adversarial را کاهش می‌دهد.

این نمونه ها با استفاده از روش نزول گرادیان پیش‌بینی شده (PGD) اصلاح شده ساخته می‌شوند که همان معادله PGD است با این تفاوت که در جهت مینیمم شدن تابع زیان حرکت می‌کند.

این نمونه ها در high likelihood regions قرار دارند و با توجه به اینکه روش تولید آنها گرادیانی است، قدرت feature extractor برای استخراج ویژگی در این فضا درنظر گرفته شده است. بنابرین دشمنان معکوس ممکن است به عنوان مثال‌هایی معنی‌دارتر برای بهبود آموزش مدل عمل کنند.

$$\check{\mathbf{x}}^{t+1} = \Pi_{\mathbb{B}(\mathbf{x},\epsilon')}\left(\check{\mathbf{x}}^{t} - \alpha' \cdot \operatorname{sign}\left(\nabla_{\check{\mathbf{x}}^{t}} \mathcal{L}_{\text{Inv}}\left(\check{\mathbf{x}}^{t}, y\right)\right)\right), \quad (3)$$

where α' is the gradient descent step size, $\check{\mathbf{x}}^{t}$ represents t^{th} iteration update, and $\mathcal{L}_{\text{Inv}}$ denotes the loss function for the inverse adversary generation. Generally, the cross-entropy loss can be a good choice for guiding the inverse adversary

۳ هدف

هدف اصلی این پروژه بررسی این است که آیا نمونه های inverse adversarial می‌توانند جایگزین بهتری برای مراکز کلاسی در center loss باشند؟ زیرا با توجه به اینکه نمونه های inverse adversarial با در نظر گرفتن قدرت feature extractor در تولید ویژگی ساخته می‌شوند، احتمالا نزدیک کردن ویژگی ها به این مراکز برای مدل راحت تر است.

برای انجام تحلیل های لازم، دو مدل با تابع زیان عادی و center loss تعلیم داده شده اند و سپس نمونه های inverse adversarial برای هر مدل تولید شده است. میانگین این نمونه ها برای هر کلاس محاسبه شده است. با توجه به نتایج، فاصله میانگین نمونه های inverse adversarial با بردار وزن کلاس ها مقایسه شده است.

۴ روش‌شناسی

۱.۴ آموزش مدل

برای تمامی آزمایش ها از مدل TRADES روی دیتاست MNIST استفاده شده است. هر مدل ۲۵ ایپاک با نرخ یادگیری 0.01 و با روش SGD آموزش دیده است.

۱. مدل A: آموزش با cross entropy

۲. مدل B: آموزش با center loss

۲.۴ تولید inverse adversarial examples

با استفاده از مدل‌های آموزش دیده، inverse adversarial examples از طریق روش PGD اصلاح شده تولید شده اند. با توجه به پارامترها در مقاله اصلی، این روش با تعداد حلقه تکرار ۲۰ و step size برابر 0.01 و برای مقادیر مختلف پارامتر ϵ انجام شده است.

۳.۴ معیارهای مقایسه فاصله

مقایسه فاصله مراکز ویژگی‌ها با استفاده از دو معیار انجام می‌شود:

۱. شباهت کسینوسی: کسینوس زاویه بین دو بردار را اندازه‌گیری می‌کند و شباهت جهت‌گیری بین بردارهای ویژگی را نشان می‌دهد.

۲. میانگین مربعات خطا (MSE): میانگین تفاوت‌های مربعی بین عناصر متناظر دو بردار را اندازه‌گیری می‌کند و تفاوت مطلق در بزرگی را نشان می‌دهد.

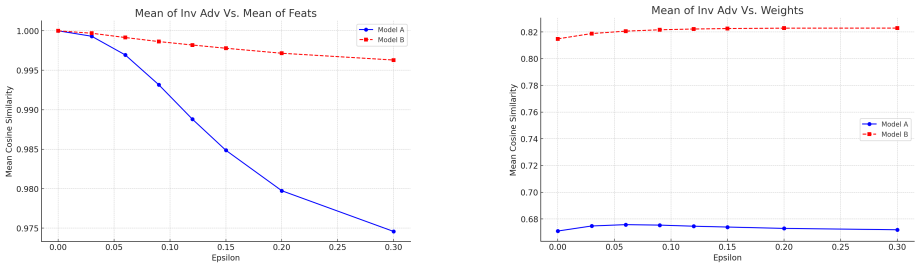
۵ آزمایش‌ها و نتایج

همه مدل ها با موفقیت همگرا شدند. ویژگی‌های استخراج شده از داده های ورودی برای محاسبه مراکز زیر استفاده شدند. برای هر کلاس، میانگین نمونه های inverse adversarial (مراکز تخاصمی) با هر سه مورد مقایسه شده است.

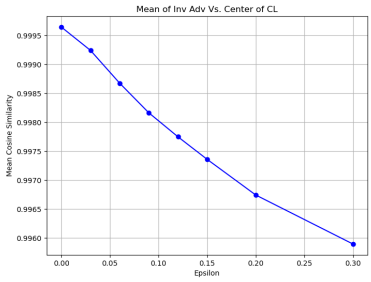
۱. میانگین خود ویژگی‌ها

۲. مراکز center loss

۳. بردار وزن



شکل ۲: میانگین شباهت کسینوسی بین مراکز تخاصمی و بردار وزن (راست) - مراکز تخاصمی و مراکز ویژگی‌ها (چپ) بر حسب ϵ در دو مدل A و B



شکل ۳: مدل B: میانگین شباهت کسینوسی بین مراکز تخاصمی و مراکز center loss بر حسب ϵ

با توجه به شکل ۲ و ۳، افزایش ϵ باعث کاهش فاصله کسینوسی و در نتیجه افزایش شباهت بین مراکز تخاصمی و میانگین ویژگی‌ها و همچنین افزایش شباهت بین مراکز تخاصمی و مراکز center loss (در مدل B) می‌شود.

در مدل B، شباهت بالا بین مراکز تخاصمی و مراکز CL نشان‌دهنده این است که نمونه های تخاصمی ویژگی‌های جهت‌گیری مراکز سنتی را حفظ می‌کنند. شباهت پایین بین مراکز تخاصمی و وزن‌های مدل نشان‌دهنده این است که وزن‌ها به طور مستقیم با مراکز فضای ویژگی‌ها هماهنگ نیستند.

علاوه بر این، شباهت با مراکز ویژگی‌ها در هر دو مدل نشان می‌دهد که نمونه های تخاصمی ویژگی‌های خاص کلاس را حفظ می‌کنند در حالی که واریانس درون کلاسی کاهش نمی‌یابد. تفاوت‌های قابل توجه با وزن‌های مدل تأیید می‌کند که ویژگی‌های فضای ویژگی‌ها با وزن‌های یاد گرفته شده متمایز هستند.

۶ نتیجه‌گیری

این پروژه نشان می‌دهد که استفاده از inverse adversarial examples به عنوان جایگزینی برای مراکز سنتی در center loss امکان‌پذیر است. با توجه به اینکه مراکز تخاصمی شباهت کسینوسی بالا و MSE پایین با مراکز سنتی center loss دارند، به ویژه در مدل A. این رویکرد نشان می‌دهد که می‌تواند استحکام مدل را افزایش داده و تفکیک‌پذیری و واریانس درون کلاسی ویژگی‌ها را نیز حفظ کند. تحقیقات بیشتر می‌تواند به بهینه‌سازی تولید inverse adversarial examples و یکپارچه‌سازی این رویکرد در چارچوب‌های دیگر یادگیری تفکیکی بپردازد.

منابع

[1] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations, ICLR 2018, 2018.

[2] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In International conference on machine learning, pages 7472–7482. PMLR, 2019.

[3] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, Xiaohua Xie. The Enemy of My Enemy Is My Friend: Exploring Inverse Adversaries for Improving Adversarial Training. CVPR, 2023.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In 6th International Conference on Learning Representations, ICLR 2018, 2018.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 499–515.

[6] Zhanglei Shi, Hao Wang , and Chi-Sing Leung , Senior Member. Constrained Center Loss for Convolutional Neural Networks. IEEE

[7] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust classification with convolutional prototype learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3474–3482.