به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

Trustworthy AI

تمرین شماره ۱

نام و نام خانوادگی	مهیار ملکی
شماره دانشجویی	۸۱۰۱۰۰۴۷۶
تاریخ ارسال گزارش	۱۴۰۲/۰۱/۲۲

فهرست گزارش سوالات

فهرست شکل‌ها

فهرست جدول‌ها

پرسش ۱ – Generalization and Robustness

قسمت اول – دیتاست CIFAR10

دادگان CIFAR10 شامل ۱۰ کلاس بوده که این کلاس‌ها و تعداد تصاویر هر کدام به شرح زیر است: (لازم به ذکر است از دیتاست cifar10 ای که در سایت Kaggle[۱] قابل دسترسی است، استفاده شده‌است.)

جدول ۱- شرح دیتاست **CIFAR10**

Class Name	Train		Test
	Train	Validation	
Airplane	1000	4000	1000
Horse	1000	4000	1000
Truck	1000	4000	1000
Automobile	1000	4000	1000
Ship	1000	4000	1000
Dog	1000	4000	1000
Bird	1000	4000	1000
Frog	1000	4000	1000
Cat	1000	4000	1000
Deer	1000	4000	1000
sum	10000	40000	1000

همانطور که در جدول ۱ مشخص است، دادگان CIFAR10 شامل ۱۰ کلاس بوده که هر کدام از آن‌ها برای داده‌های آموزش[۲] و ارزیابی[۳] به ترتیب شامل ۵۰۰۰ و ۱۰۰۰ تصویر می‌باشند. مطابق خواسته سوال تنها از ۲۰ درصد داده‌های آموزش برای آموزش مدل استفاده خواهیم‌کرد و ۸۰ درصد باقی‌مانده برای اعتبارسنجی[۴] مورد استفاده قرار خواهندگرفت. چنانچه در جدول نیز مشخص است، تعداد تصاویر هر کلاس برابر بوده و تناسب بین کلاس‌ها برقرار است.

همچنین همانطور که در آخرین قسمت تمرین نیز ذکر شده، برای این که در هر بچ از داده‌ها به تعداد مساوی تصویر از هر کلاس قرار بگیرد، از یک تابع Batch Sampler استفاده کرده‌ایم.

---

[۱] https://www.kaggle.com/datasets/oxcdcd/cifar10

[۲] Train

[۳] Test

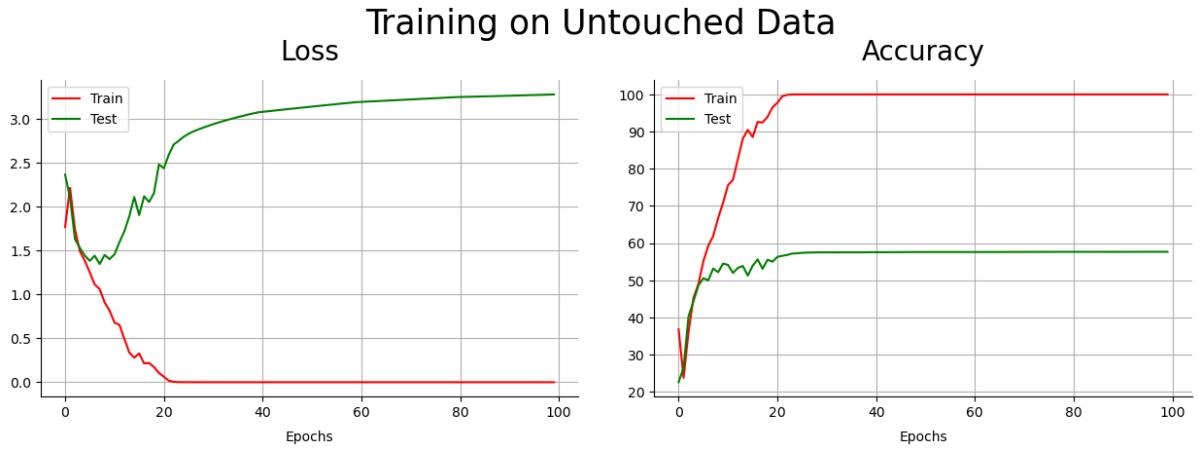
[۴] Validation

قسمت دوم – آموزش شبکه روی داده‌های دست نخورده

ابتدا مدل resnet18 را از کتابخانه torchvision بارگزاری می‌کنیم، سپس پس از تغییر ابعاد خروجی لایه‌ی تماما متصل[1] برای طبقه‌بند ۱۰ کلاسه، به آموزش مدل به روی تصاویر دست نخورده می‌پردازیم. پارامترهای آموزش در جدول ۲ و نمودارهای دقت و هزینه حین آموزش در شکل ۱ قابل مشاهده می‌باشند.

جدول ۲- پارامترهای آموزش مدل روی دادگان دست نخورده و تابع هزینه **crossentropy**

Model	Resnet18
Dataset	Cifar10 (Untouched)
Batch Size	500 (50 pic from each class)
Loss Function	Cross Entropy
Optimizer	Adam
Learning Rate	0.001
Scheduler	LR decreases by factor of 0.5 every 20 epochs
Epochs	100



شکل ۱- نمودارهای دقت و هزینه آموزش مدل روی دادگان دست نخورده و تابع هزینه **crossentropy**

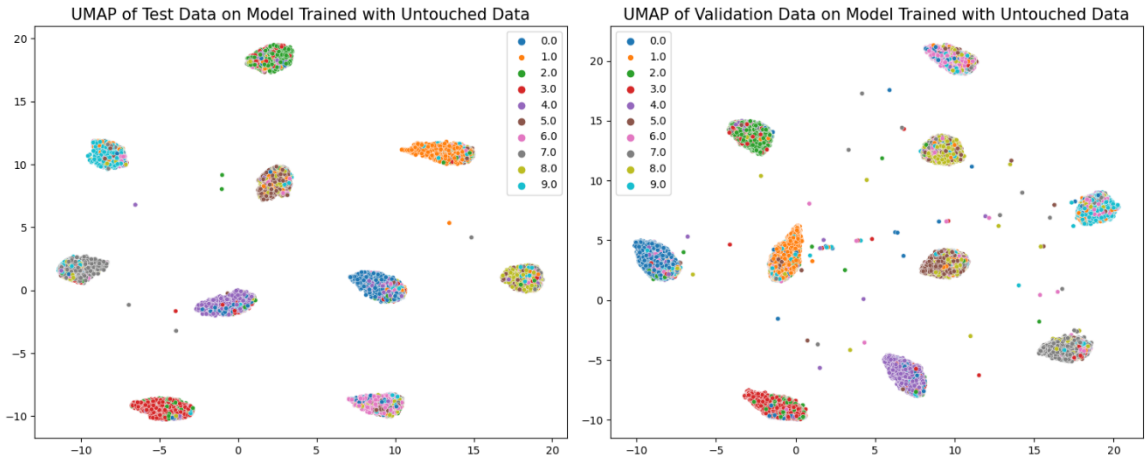
[1] Fully Connected

همانطور که در شکل ۱ قابل مشاهده است، آموزش شبکه بروی دادگان آموزش به خوبی صورت گرفته و هزینه و دقت شبکه برای داده‌های آموزش به ترتیب به مقادیر ۰ و ۱۰۰ همگرا شده‌اند. این در حالی است که عملکرد شبکه برای دادگان اعتبارسنجی متفاوت بوده و دقت خیلی کمتری (نزدیک به ۵۰ درصد) حاصل شده‌است و همچنین نمودار هزینه روندی صعودی دارد. این اتفاق می‌تواند به این دلیل باشد که تعداد داده‌های آموزش به نسبت کم است (نسبت ۱ به ۴ دادگان آموزش و اعتبارسنجی) و باعث شده است تا مدل آموزش داده‌شده، تعمیم‌پذیری[۱] خوبی نداشته‌باشد. دقت شبکه برای داده‌های دیده نشده (اعتبارسنجی و ارزیابی) در جدول ۳ قابل مشاده است.

جدول ۳- دقت مدل آموزش یافته روی دادگان دست نخورده و تابع هزینه **crossentropy** برای داده‌های دیده نشده دست نخورده

	Validation	**Test**
Accuracy	57.65%	57.72%

همچنین بازنمایی قسمت کانولوشنال[۲] شبکه که در شکل ۲ قابل مشاهده است نیز نتایج گفته‌شده را تصدیق می‌کند. لازم به ذکر است که برای بدست آوردن این بازنمایی‌ها و نمایش آن به صورت دوبعدی از روش کاهش ابعاد UMAP استفاده شده‌است. چنانچه قابل مشاهده است، با توجه به ویژگی‌های به دست آمده از قسمت کانولوشنال شبکه، تفکیک پذیری کلاس‌ها برای داده‌های دیده نشده به خوبی انجام شده‌است، ولی بین کلاس‌ها تداخل‌هایی نیز دیده می‌شود که دقت کم شبکه روی داده‌های دیده نشده می‌تواند ناشی از این امر باشد.



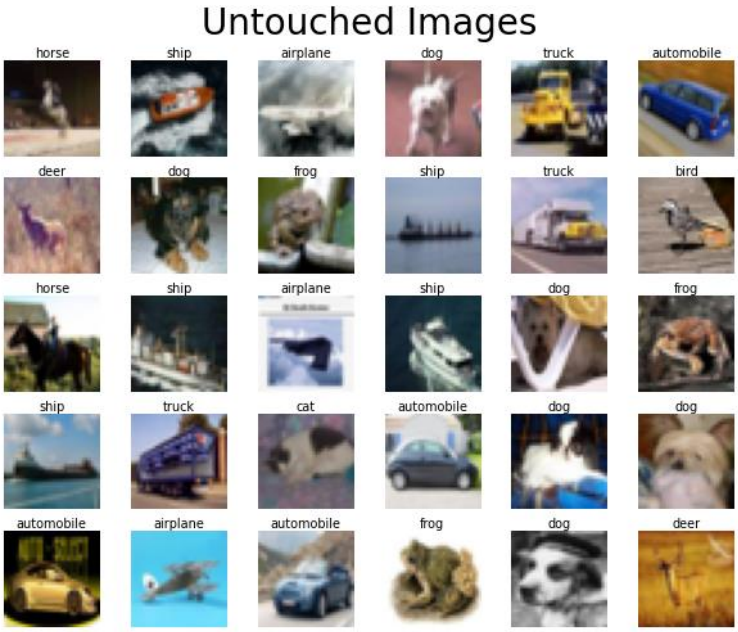
شکل ۲- بازنمایی قسمت کانولوشنال شبکه آموزش یافته روی دادگان دست نخورده با استفاده از **UMAP** برای داده‌های اعتبارسنجی و ارزیابی

[۱] Generalization

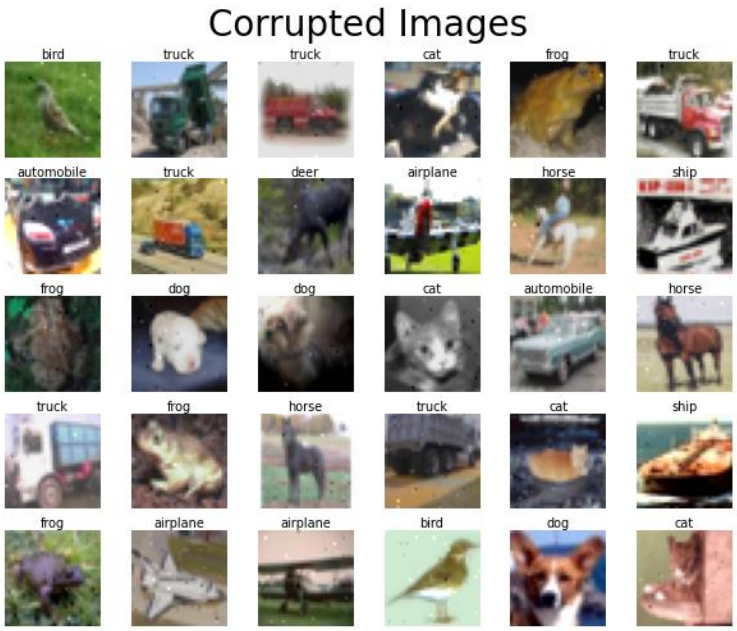
[۲] Convolutional

قسمت سوم – ارزیابی شبکه روی داده‌های اغتشاش یافته

در این قسمت همانطور که در صورت سوال گفته شده بروی داده‌ها اغتشاش ایجاد می‌کنیم. اغتشاشاتی که ما ایجاد کردیم، اعمال color jitter برای تغییر روشنایی و کنتراست تصاویر و همچنین ایجاد نویز گاوسی به صورت رندوم روی ۳۰ پیکسل از هر تصویر می‌باشد. نمونه‌هایی از تصاویر سالم و اغتشاش‌یافته در شکل‌های ۳ و ۴ قابل مشاهده‌اند. همچنین، با روش Fast Gradient یک حمله متخاصمانه نیز در حین آموزش شبکه اعمال شده‌است.



شکل ۳- تصاویر دست نخورده



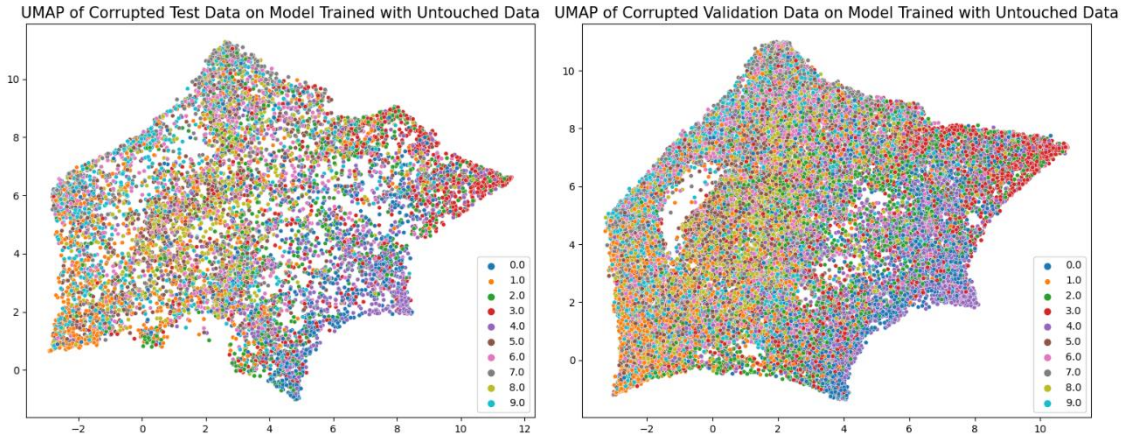
شکل ۴- تصاویر اغتشاش یافته

در قسمت قبل دیدیم که مدل ما به دلیل تعداد کم داده‌های آموزش تعمیم‌پذیری کمی داشت. در این قسمت نیز با توجه جدول ۴ مشاهده می‌کنیم که دقت مدل روی داده‌های اغتشاش یافته کاهش قابل توجهی داشته است. لذا می‌توان نتیجه گرفت که مدل آموزش داده شده در قسمت قبل از منظر مقاومت در برابر نویز[۱] نیز عملکرد خوبی ندارد.

جدول ۴- دقت مدل آموزش یافته روی دادگان دست نخورده برای داده‌های دیده نشده اغتشاش‌یافته

	Validation	**Test**
Accuracy	21.40%	21.83%

مجددا بازنمایی بخش کانولوشنال شبکه را با استفاده از UMAP ولی این بار به روی داده‌های اغتشاش یافته به دست می‌آوریم که در شکل ۵ قابل مشاهده است. همانطور که می‌بینیم، مطابق انتظار تفکیک‌پذیری بین کلاس‌ها بسیار ضعیف بوده و داده‌ها تداخل زیادی دارند. در واقع به نظر می‌رسد که ویژگی‌های بدست آمده از شبکه، به هیچ وجه ویژگی‌های مقاومی نمی‌باشند.



شکل ۵- بازنمایی قسمت کانولوشنال شبکه آموزش یافته روی دادگان دست نخورده با استفاده از **UMAP** برای داده‌های دیده نشده اغتشاش‌یافته

[۱] Robustness

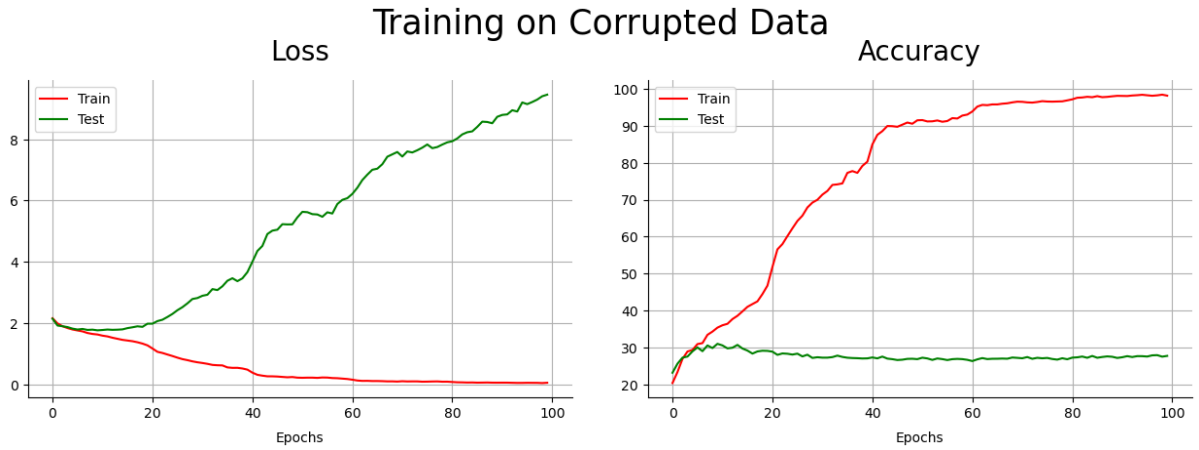
قسمت چهارم – آموزش و ارزیابی شبکه روی داده‌های اغتشاش یافته

در این قسمت شبکه را این بار با داده‌های اغتشاش یافته (adversarial example) که در قسمت قبل به دست آوردیم، آموزش می‌دهیم تا ببینیم مقاومت شبکه چگونه تغییر خواهدکرد. مشخصات پارامترهای آموزش شبکه در جدول ۵ قابل مشاهده می‌باشد.

جدول ۵- پارامترهای آموزش مدل روی دادگان اغتشاش یافته و تابع هزینه **crossentropy**

Model	Resnet18
Dataset	Cifar10 (Noisy)
Batch Size	500 (50 pic from each class)
Loss Function	Cross Entropy
Optimizer	Adam
Learning Rate	0.001
Scheduler	LR decreases by factor of 0.5 every 20 epochs
Epochs	100

نمودارهای دقت و هزینه در فرایند آموزش مدل، در شکل ۶ قابل مشاهده است. همانطور که می‌بینیم، در مقایسه با قبل، فرایند آموزش مدل سخت‌تر بوده‌است و همگرایی نمودارها با سرعت کمتری اتفاق افتاده‌است. همچنین دقت شبکه روی داده‌های اعتبارسنجی، در مقایسه با قبل، به مقدار کمتری همگرا شده‌است.



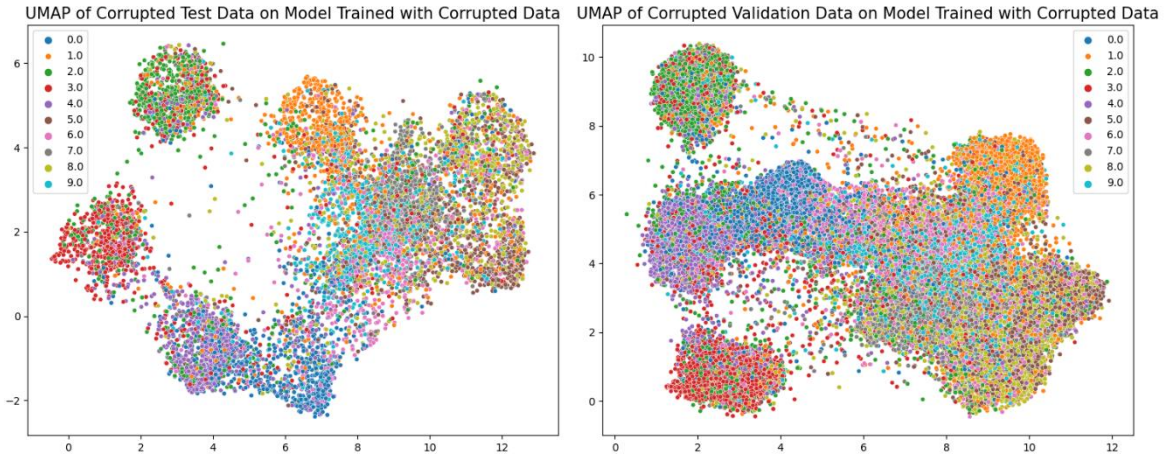
شکل ۶- نمودارهای دقت و هزینه آموزش مدل روی دادگان اغتشاش یافته و تابع هزینه **crossentropy**

با توجه به دقت مدل برای داده‌های دیده نشده که در جدول ۶ قابل مشاهده می‌باشد، می‌بینیم که دقت مدل جدید در مقایسه با مدل قبلی، کمی افزایش یافته‌است، لذا شبکه آموزش داده شده اندکی مقاوم‌تر شده‌است. البته همچنان مقادیر دقت پایین می‌باشند و مدل به اندازه کافی مقاوم نشده است.

جدول ۶- دقت مدل آموزش یافته روی دادگان اغتشاش یافته برای داده‌های دیده نشده اغتشاش‌یافته

	Validation	**Test**
Accuracy	28.25%	27.98%

همچنین مجددا با رسم نمودار بازنمایی بخش کانولوشنال شبکه که در شکل ۷ قابل مشاهده می‌باشد، می‌بینیم که تفکیک‌پذیری بین کلاس‌ها اندکی بهبود یافته و شبکه ویژگی‌های مقاوم‌تری را نسبت به قبل انتخاب کرده‌است. اما با توجه به دقت پایینی که روی داده‌های اغتشاش یافته نتیجه شده‌است، همچنان مقاومت شبکه پایین می‌باشد.

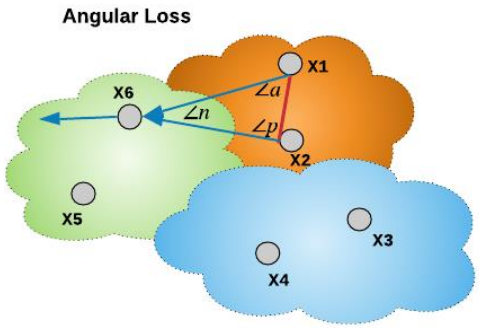


شکل ۷- بازنمایی قسمت کانولوشنال شبکه آموزش یافته روی دادگان اغتشاش‌یافته با استفاده از **UMAP** برای داده‌های دیده نشده اغتشاش‌یافته

قسمت پنجم – تابع هزینه زاویه‌ای[1]

Metric Learning

یک شبکه عمیق metric learning تصاویری که به هم شباهت دارند را روی مکان‌های نزدیک و تصاویری که مشابهت ندارند را در یک فضای embedding با فاصله از هم قرار می‌دهد. با استفاده از این روش یادگیری، ویژگی‌هایی از تصاویر به دست می‌آیند که به خوبی قابل تمییز بوده و واریانس داخلی محدودی نیز خواهندداشت. یادگیری این ویژگی‌ها این قابلیت را به شبکه می‌دهد که تعمیم‌پذیری خوبی به روی تصاویر دیده نشده داشته‌باشد و در نهایت کلاس‌های جدیدی در فضای embedding تشکیل دهد. این شبکه‌ها مشابه دیگر شبکه‌های عمیق آموزش می‌بینند، با این تفاوت که از تابع هزینه متفاوتی در آن‌ها استفاده می‌شود. این تابع هزینه، تصاویر مشابه را به هم نزدیک و تصاویر نامشابه را در فاصله از هم در فضای embedding قرار می‌دهد. به این صورت که در هر ایپاک تصاویر آموزش به شبکه داده شده و در فضای embedding نگاشت می‌شوند. سپس میزان خطای این نگاشت‌ها توسط تابع خطای گفته‌شده محاسبه شده و وزن‌های شبکه تنظیم می‌شوند. همچنین برای ارزیابی شبکه، نیاز است تا ویژگی‌های خروجی لایه embedding استخراج شده و با استفاده از یک طبقه‌بند نزدیک‌ترین همسایه، عملیات طبقه‌بندی انجام شود. توابع هزینه زیادی به این منظور معرفی شده‌اند که تابع هزینه زاویه‌ای در این قسمت مورد بحث ما می‌باشد.



شکل ۸- تابع هزینه زاویه‌ای

تابع هزینه زاویه‌ای

بر خلاف دیگر روش‌های metric learning که از معیار فاصله استفاده می‌کنند، این تابع هزینه بر مبنای زاویه بوده و پیشنهاد می‌دهد تا یک رابطه درجه سه را درون مثلثی سه‌تایی ($\Delta_{a,p,n}$) تعریف کنیم. این تابع به کمک تعریف فاصله زاویه‌ای، باعث می‌شود تا ویژگی منفی از خوشه مثبت فاصله گرفته و نقاط مثبت

[1] Angular loss

نیز به هم نزدیک‌تر شوند. (شکل ۸) در نظر گرفتن زاویه (کسینوس) باعث می‌شود که شبکه نه تنها به اندازه بلکه در برابر چرخش تصاویر هم مقاوم باشد. به عبارت دیگر، دید زاویه‌ای به ترم هزینه، مقاومت بیشتری را به تغییرات محلی ویژگی‌ها نتیجه می‌دهد.

مزایای تابع هزینه زاویه‌ای:

۱. مقاوم بودن شبکه نه تنها به اندازه بلکه در برابر چرخش
۲. قانون کسینوس‌ها هر سه ضلع مثلث را درگیر می‌کند، در مقابل تابع هزینه triplet تنها دو ضلع را درگیر می‌کند. اضافه شدن ضلع سوم باعث افزایش مقاومت شبکه می‌شود
۳. انتخاب مارجین خطای فاصله‌ای (Euclidean) برای تابع هزینه کار ساده‌ای نیست. این امر بیشتر به این دلیل است که با افزایش اندازه دیتاست، تغییرات داخل کلاسی بین کلاس‌های هدف بسیار متفاوت است و بدون یک مرجع معنی‌دار ، تنظیم چنین پارامتری دشوار است. در مقابل، انتخاب مارجین زاویه‌ای، به علت مقاوم بودن در برابر اندازه، ساده‌تر است
۴. تابع خطای زاویه‌ای به راحتی می‌تواند با دیگر توابع خطای مرسوم ترکیب شود تا عملکرد کلی را بهبود دهد

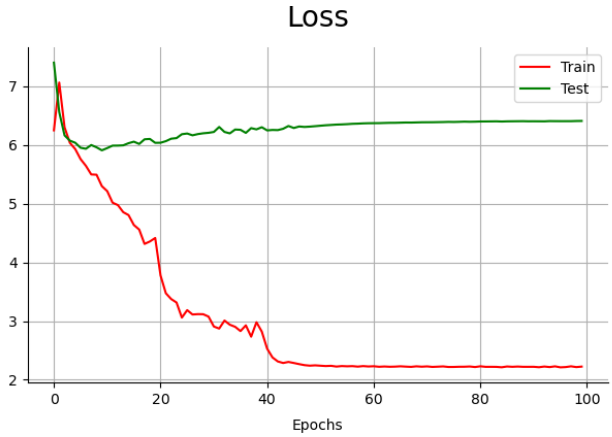
قسمت ششم – آموزش و ارزیابی شبکه روی داده‌های دست نخورده و تابع هزینه زاویه‌ای

در این قسمت با استفاده از تابع هزینه زاویه‌ای، شبکه را آموزش داده و سعی می‌کنیم تا مقاومت مدل را افزایش دهیم. پارامترهای شبکه در جدول ۷ قابل مشاهده است. لازم به ذکر است که برای ارزیابی مدل، همانطور که در قسمت ۴ گفته شد، خروجی شبکه را به یک طبقه‌بند نزدیکترین همسایه[۱] داده‌ایم. همچنین در لایه آخر از یک لایه تماما متصل برای تبدیل خروجی کانولوشنال شبکه به یک بردار ۶۴تایی استفاده کرده و مقادیر آن را نرمالایز کرده‌ایم تا در طبقه‌بند نزدیکترین همسایه، نتایج بهتری حاصل شود.

جدول ۷- پارامترهای آموزش مدل روی دادگان دست نخورده و تابع هزینه **Angularloss**

Model	Resnet18
Dataset	Cifar10 (Untouched)
Batch Size	500 (50 pic from each class)
Loss Function	Angular
Optimizer	Adam
Learning Rate	0.001
Scheduler	LR decreases by factor of 0.5 every 20 epochs
Epochs	100

نمودار هزینه در فرایند آموزش مدل، در شکل ۹ قابل مشاهده است. همانطور که می‌بینیم، هزینه آموزش پس از حدود ۵۰ ایپاک به کمینه خود همگرا شده‌است.



شکل ۹- نمودارهای دقت و هزینه آموزش مدل روی دادگان دست نخورده با تابع هزینه **AngularLoss**

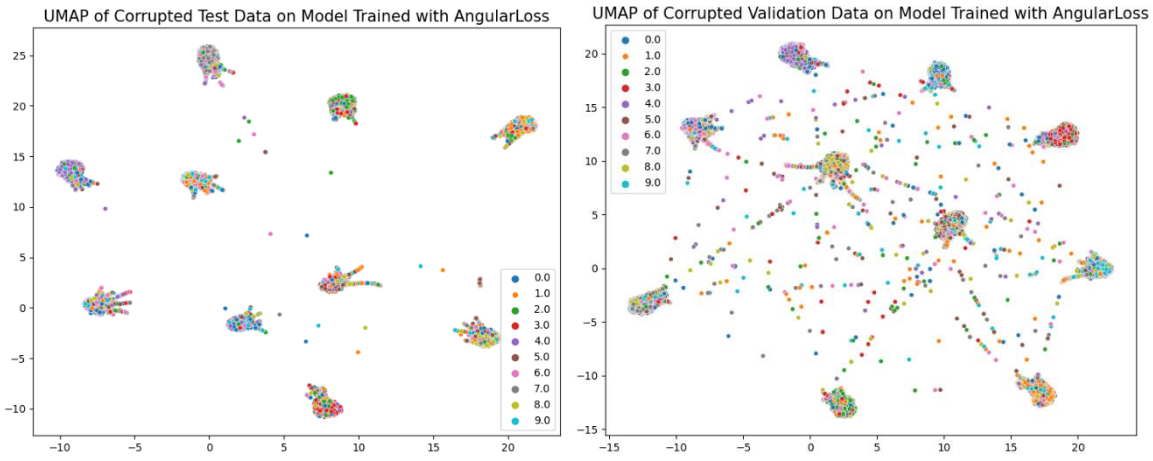
[۱] K-Nearest Neighbor

با توجه به دقت مدل برای داده‌های دیده نشده که در جدول ۸ قابل مشاهده می‌باشد، می‌بینیم که دقت مدل جدید در مقایسه با مدل قبلی، به طور قابل توجهی افزایش یافته‌است، لذا شبکه آموزش داده شده در برابر حملات مقاوم‌تر شده‌است.

جدول ۸- دقت مدل آموزش یافته روی دادگان دست نخورده با تابع هزینه **AngularLoss** برای داده‌های دیده نشده اغتشاش یافته

	Validation	**Test**
Accuracy	44.20%	44.87%

همچنین با رسم نمودار بازنمایی بخش کانولوشنال شبکه که در شکل ۱۰ قابل مشاهده می‌باشد، می‌بینیم که تفکیک‌پذیری بین کلاس‌ها تا حد زیادی بهبود یافته و شبکه ویژگی‌های مقاوم‌تری را نسبت به قبل انتخاب کرده‌است. البته همانند قسمت دوم، بین کلاس‌ها تداخل‌هایی نیز دیده می‌شود که دقت کم شبکه روی داده‌های دیده نشده می‌تواند ناشی از این امر باشد.



شکل ۱۰- بازنمایی قسمت کانولوشنال شبکه آموزش یافته روی دادگان دست نخورده با تابع هزینه **AngularLoss** با استفاده از **UMAP** برای داده‌های دیده نشده اغتشاش یافته

نتیجه گیری

جدول ۹- خلاصه و نتیجه‌گیری

مدل	داده‌های آموزش	تابع هزینه	دقت اعتبارسنجی	توضیح
اول	دست نخورده	Cross Entropy	21.40%	در این قسمت دیدیم که مدل ما در برابر نویز و حملات متخاصمانه مقاوم نبوده و دقت کمی به دست آمده‌است.
دوم	اغتشاش یافته	Cross Entropy	28.25%	در این قسمت با آموزش شبکه به روی داده‌های اغتشاش یافته، اندکی بهبود در دقت مدل مشاهده شد ولی همچنان مقاومت خوبی قابل مشاهده نیست.
سوم	دست نخورده	Angular	44.20%	در این قسمت با استفاده از روش metric learning و تابع هزینه زاویه‌ای سعی در بهبود مقاومت مدل و یافتن ویژگی‌های مقاوم‌تر داشتیم که تا حد خوبی این امر نتیجه شد.