Transfer Learning Optimization
بررسی تکنیک‌های نرمال‌سازی و دینامیک گرادیان در MobileNetV2 بر روی CIFAR-10

هیوا ابوالهادی‌زاده
۴۰۰۴۰۵۰۰۴
مباحث ویژه در هوش مصنوعی
استاد: جناب آقای دکتر افتخاری

۲۲ فروردین ۱۴۰۴

چکیده

در این پژوهش، تأثیر سه روش نرمال‌سازی شامل BatchNorm، LayerNorm و Filter Response Normalization (FRN) در چارچوب یادگیری انتقالی با مدل MobileNetV2 بر مجموعه داده CIFAR-10 بررسی شده است. همچنین نقش Gradient Clipping در پایداری گرادیان‌ها تحلیل شده است. نتایج نشان می‌دهد که ساختار FRN بدون Clipping بهترین عملکرد را ارائه می‌دهد وClipping در ترکیب با BatchNorm موجب بهبود همگرایی می‌شود. LayerNorm نیز بدون نیاز بهClipping عملکردی پایدار از خود نشان داده است. این یافته‌ها انتخاب نرمال‌سازی مناسب را به‌عنوان عاملی کلیدی در موفقیت fine-tuning برجسته می‌سازد.

فهرست مطالب

۱ مقدمه

یادگیری انتقالی [۱] یکی از رویکردهای کلیدی در یادگیری ماشین مدرن است که به مدل‌ها امکان می‌دهد دانش حاصل از حل یک مسئله را برای حل مسئله‌ای دیگر به کار ببرند. این روش به‌ویژه در شرایطی که داده‌های آموزشی محدود هستند، بسیار مؤثر عمل می‌کند. به جای آموزش مدل از ابتدا، از یک مدل از پیش آموزش‌دیده استفاده شده و آن را برای وظیفه‌ی جدید بهینه‌سازی می‌کنیم. این فرآیند که اصطلاحاً fine-tuning نام دارد، در سال‌های اخیر به‌طور گسترده در کاربردهای بینایی ماشین، پردازش زبان طبیعی، و سایر حوزه‌ها مورد استفاده قرار گرفته است.

در مسیر آموزش مدل‌های عمیق، یکی از عوامل تعیین‌کننده در پایداری و سرعت همگرایی، نرمال‌سازی داده‌ها و لایه‌ها است. تکنیک‌هایی نظیر Batch Normalization، Layer Normalization و Filter Response Normalization (FRN)، با تنظیم آماری ویژگی‌ها در طول آموزش، باعث می‌شوند مدل بتواند سریع‌تر و پایدارتر یاد بگیرد. با این حال، رفتار این روش‌های نرمال‌سازی در زمینه‌ی یادگیری انتقالی، به‌ویژه هنگام تطبیق مدل‌های از پیش‌آموزش‌دیده با داده‌های جدید، هنوز نیازمند بررسی دقیق‌تر است.

هدف این پروژه، تحلیل تجربی و سیستماتیک عملکرد روش‌های مختلف نرمال‌سازی در فرآیند انتقال یادگیری است. در این راستا، از مدل MobileNetV2 به عنوان مدل پایه استفاده شده است؛ مدلی سبک و بهینه برای دستگاه‌های با توان پردازشی پایین که بر روی مجموعه‌های داده‌ی بزرگ مانند ImageNet آموزش دیده است. داده‌های مورد استفاده در این پروژه، مربوط به مجموعه‌ی تصویری CIFAR-10 هستند که شامل ۶۰۰۰۰ تصویر رنگی در ۱۰ کلاس مختلف با ابعاد ۳۲×۳۲ پیکسل می‌باشند. برای سازگاری با ورودی MobileNetV2، این تصاویر به ابعاد ۲۲۴×۲۲۴ بازنمونه‌گیری شده‌اند.

در این گزارش، سه نوع head نرمال‌سازی مختلف طراحی و با مدل پایه تلفیق شده‌اند. همچنین، تأثیر تکنیک Gradient Clipping نیز بررسی شده تا تأثیر آن بر پایداری و رفتار گرادیان‌ها در طی فرآیند fine-tuning مشخص گردد. برای تحلیل جامع، از ابزارهایی مانند تحلیل چشم‌انداز تابع هزینه [۲] و پایش گرادیان استفاده شده است.

سوالات اصلی این پروژه عبارت‌اند از:

کدام نوع نرمال‌سازی در انتقال یادگیری بر روی CIFAR-10 عملکرد بهتری دارد؟

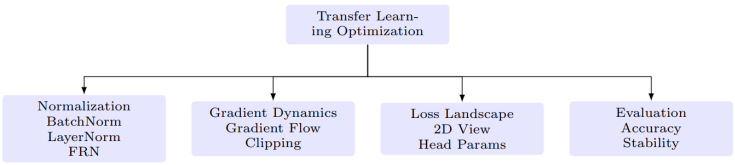
آیا استفاده از Gradient Clipping باعث بهبود پایداری آموزش می‌شود؟

رفتار گرادیان‌ها در لایه‌های مختلف مدل چه تفاوت‌هایی در شرایط نرمال‌سازی مختلف دارد؟

---

[۱] Transfer Learning

[۲] Loss Landscape



شکل ۱: نمای کلی اجزای مورد بررسی در این گزارش

۲ مروری بر کارهای پیشین

یادگیری انتقالی، به‌ویژه در حوزه بینایی ماشین، در سال‌های اخیر مورد توجه گسترده‌ای قرار گرفته است. استفاده از مدل‌های از پیش آموزش‌دیده مانند ResNet، EfficientNet و MobileNet در مسائل جدید، موجب صرفه‌جویی در منابع محاسباتی و بهبود عملکرد مدل شده است. با این حال، چگونگی تطبیق این مدل‌ها با داده‌های جدید و نقش تکنیک‌های نرمال‌سازی در فرآیند fine-tuning، همچنان از موضوعات باز و پرچالش در ادبیات هوش مصنوعی هستند.

در مطالعه‌ی [Singh and Krishnan, 2020]، نویسندگان تکنیک جدیدی به نام Filter Response Normalization (FRN) را معرفی می‌کنند که برخلاف Batch Normalization، مستقل از اندازه‌ی batch و توزیع آماری داده‌ها عمل می‌کند. این ویژگی، FRN را به گزینه‌ای مناسب برای یادگیری انتقالی، به‌ویژه در شرایط با batch size کوچک یا داده‌های نامتوازن، تبدیل کرده است.

از سوی دیگر، تحلیل چشم‌انداز تابع هزینه [۳] به‌عنوان ابزاری برای درک بهتر رفتار مدل‌ها در طول آموزش پیشنهاد شده است. مقاله‌ی مشهور [Li et al., 2018] نشان می‌دهد که شکل و همواری نواحی کمینه‌ی تابع هزینه می‌تواند با قابلیت تعمیم مدل ارتباط مستقیم داشته باشد.

در زمینه‌ی پایداری آموزش شبکه‌های عمیق، تکنیک Gradient Clipping نیز به‌عنوان روشی برای کنترل gradient explosion مطرح شده است. یافته‌های مطالعاتی مانند [Zhang et al., 2019] نشان می‌دهد که در شرایطی خاص، به‌ویژه هنگام fine-tuning لایه‌های عمیق، clipping می‌تواند منجر به همگرایی پایدارتر و کاهش نوسانات در مقدار گرادیان‌ها شود.

مطالعه‌ی حاضر، در ادامه‌ی این پژوهش‌ها، تلاش دارد تا با تمرکز بر ترکیب MobileNetV2، داده CIFAR-10، و تکنیک‌های نرمال‌سازی متنوع، به ارزیابی تجربی نقش نرمال‌سازی و clipping در یادگیری انتقالی بپردازد. در این مسیر، از تحلیل loss landscape و پراکندگی گرادیان نیز برای درک عمیق‌تر استفاده خواهد شد.

---

[۳] Loss Landscape

۳ روش‌شناسی

این بخش به تشریح فرآیند آماده‌سازی داده، طراحی مدل انتقال یادگیری، پیاده‌سازی Headهای نرمال‌سازی، و تنظیمات دقیق آزمایش‌ها می‌پردازد. کلیه مراحل در محیط Google Colab با استفاده از کتابخانه‌های PyTorch و Torchvision پیاده‌سازی شده‌اند.

۱.۳ معماری مدل و تغییرات

مدل پایه استفاده‌شده در این پروژه MobileNetV2 است که از پیش بر روی مجموعه داده ImageNet آموزش دیده و به‌عنوان Feature Extractor عمل می‌کند[۴]. برای تطبیق این مدل با وظیفه‌ی طبقه‌بندی تصاویر CIFAR-10، بخش classifier آن حذف شده و خروجی آن با Head جدیدی ترکیب شده است. این Head مسئول تولید خروجی ۱۰کلاسه نهایی است.

سه نوع Head مختلف با هدف بررسی تأثیر روش‌های نرمال‌سازی طراحی و پیاده‌سازی شده‌اند:

- **BatchNormHead**: این Head شامل یک لایه Fully Connected با خروجی ۲۵۶، یک لایه Batch Normalization[۵]، تابع فعال‌ساز ReLU، Dropout و در نهایت یک لایه FC خروجی است. میانگین و واریانس را در طول batch محاسبه می‌کند. نرمال‌سازی به صورت $\hat{x} = \frac{x-\mu_{\text{batch}}}{\sqrt{\sigma^2_{\text{batch}}+\epsilon}}$ انجام شده و سپس توسط مقیاس و انتقال (γ, β) تنظیم می‌شود.
- **LayerNormHead**: ساختاری مشابه BatchNormHead دارد، با این تفاوت که از Layer Normalization استفاده می‌کند. آماره‌ها به صورت مستقل از batch و در طول ویژگی‌ها (در هر نمونه) محاسبه می‌شوند، و نرمال‌سازی روی feature dimension داخلی انجام می‌گیرد.
- **FRNHead**: در این ساختار از تکنیک Filter Response Normalization (FRN) استفاده شده است که به‌طور خاص طراحی شده تا از وابستگی به آماره‌های دسته جلوگیری کند. لایه FRN در این Head به‌صورت سفارشی پیاده‌سازی شده و با یک پارامتر τ ترکیب می‌شود: $\max(\gamma x + \beta,\ \tau)$. برخلاف دو مورد قبلی، هیچ وابستگی به batch یا channel خاص ندارد.

از منظر یادگیری انتقالی، BatchNorm در شرایطی با batch size کوچک[۶] با نوسان آماره‌ها دچار افت عملکرد می‌شود[۷]. FRN به دلیل مستقل بودن از می‌تواند در این شرایط عملکرد

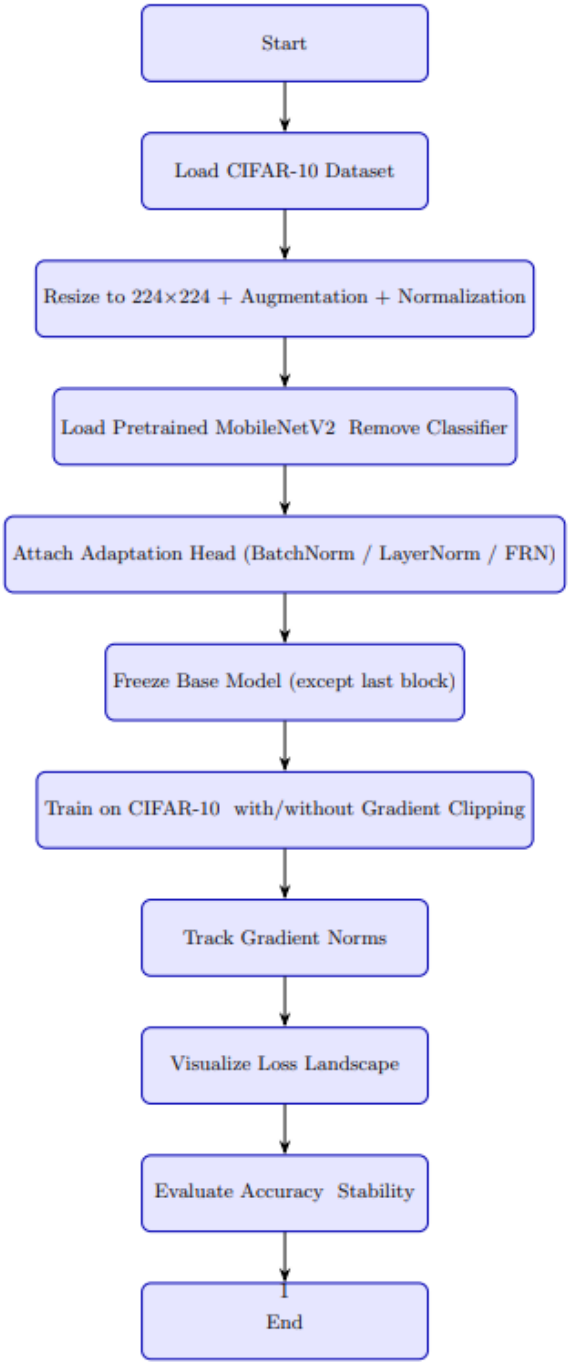
---

[۴] استفاده از مدل‌های از پیش آموزش‌دیده یکی از رایج‌ترین روش‌ها در یادگیری انتقالی است [Pan and Yang, 2010].

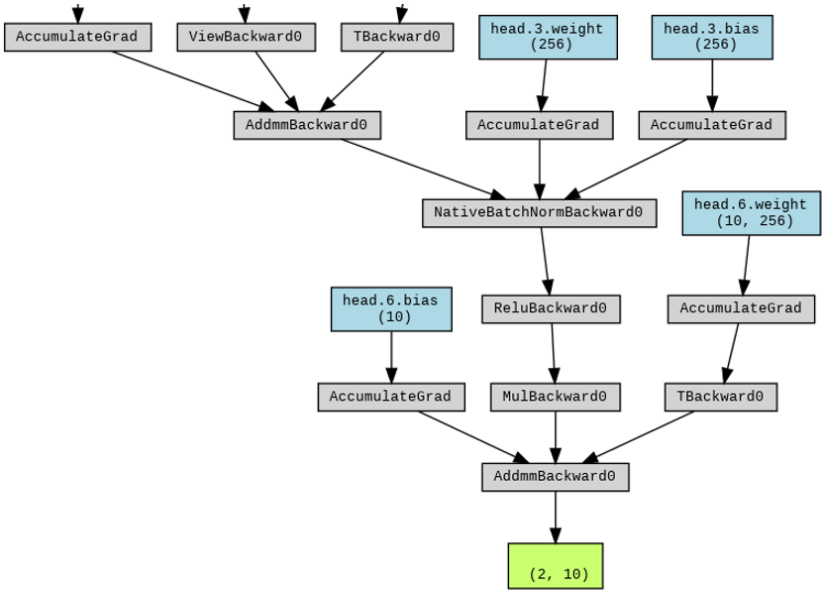
[۵] نرمال‌سازی دسته‌ای نقش مهمی در پایداری گرادیان‌ها و تسریع همگرایی دارد [Singh and Krishnan, 2020].

[۶] مانند fine-tuning با داده کم

[۷] رجوع شود به [Singh and Krishnan, 2020]



شکل ۲: جریان کار



شکل ۳: بخشی از معماری مدل

بهتری داشته باشد. در مقابل، LayerNorm گرچه بیشتر در حوزه‌های مبتنی بر توالی مانند NLP رایج است، اما در برخی ساختارهای Vision Transformer نیز موفق بوده و در این پروژه نیز برای مقایسه استفاده شده است. در همه‌ی Headها، ابتدا روی خروجی Feature Extractor عملیات Global Average Pooling انجام می‌شود، سپس ویژگی‌ها با Flatten به برداری با اندازه ثابت تبدیل شده و به لایه FC ورودی Head داده می‌شوند.

برای جلوگیری از Overfitting و حفظ ویژگی‌های عمومی آموخته‌شده توسط MobileNetV2، همه‌ی لایه‌های مدل پایه فریز شده‌اند به‌جز بلوک انتهایی آن (features.18)، که برای فرآیند fine-tuning باز شده و با نرخ یادگیری پایین‌تر آموزش دیده است.

شکل ۴: معماری BatchNorm Head

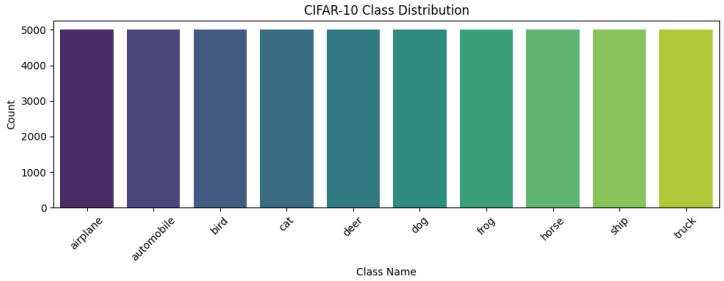
شکل ۵: معماری LayerNorm Head

شکل ۶: معماری FRN Head

۲.۳ تنظیمات آزمایش‌ها

مجموعه داده مورد استفاده در این پروژه، CIFAR-10 است که شامل ۶۰۰۰۰ تصویر رنگی ۳۲×۳۲ در ده کلاس است. برای سازگاری با ورودی مدل MobileNetV2، تمام تصاویر به اندازه ۲۲۴×۲۲۴ بازنمونه‌گیری شده‌اند[۸].

شکل ۷: نمونه‌هایی از داده‌ها



شکل ۸: توزیع داده‌ها به ازای هر کلاس

برای داده‌های آموزش، عملیات Data Augmentation شامل موارد زیر انجام شده است:

- RandomHorizontalFlip
- RandomCrop با padding برابر ۴
- نرمال‌سازی با میانگین و انحراف معیار CIFAR-10:

$$\mu = (0.4914, 0.4822, 0.4465)$$
$$\sigma = (0.2023, 0.1994, 0.2010)$$

[۸] ورودی اصلی MobileNetV2 برای تصویر با اندازه ۲۲۴×۲۲۴ طراحی شده است.

برای داده‌های اعتبارسنجی و تست، تنها Resize و نرمال‌سازی اعمال شده است. تقسیم‌بندی داده‌ها به‌صورت ۸.۰ آموزش و ۲.۰ اعتبارسنجی انجام شده است. یک کلاس اختصاصی به‌نام CIFAR10DataLoader مسئول بارگذاری، تقسیم، و بصری‌سازی داده‌هاست (از جمله توزیع کلاس‌ها و پیش‌نمایش نمونه‌ها). آزمایش‌های طراحی‌شده به‌صورت ترکیبی از سه نوع Head[۹] و دو حالت استفاده یا عدم استفاده از Gradient Clipping (با مقدار آستانه ۱/۰) می‌باشند. در نتیجه ۶ حالت مختلف بررسی شده‌اند.

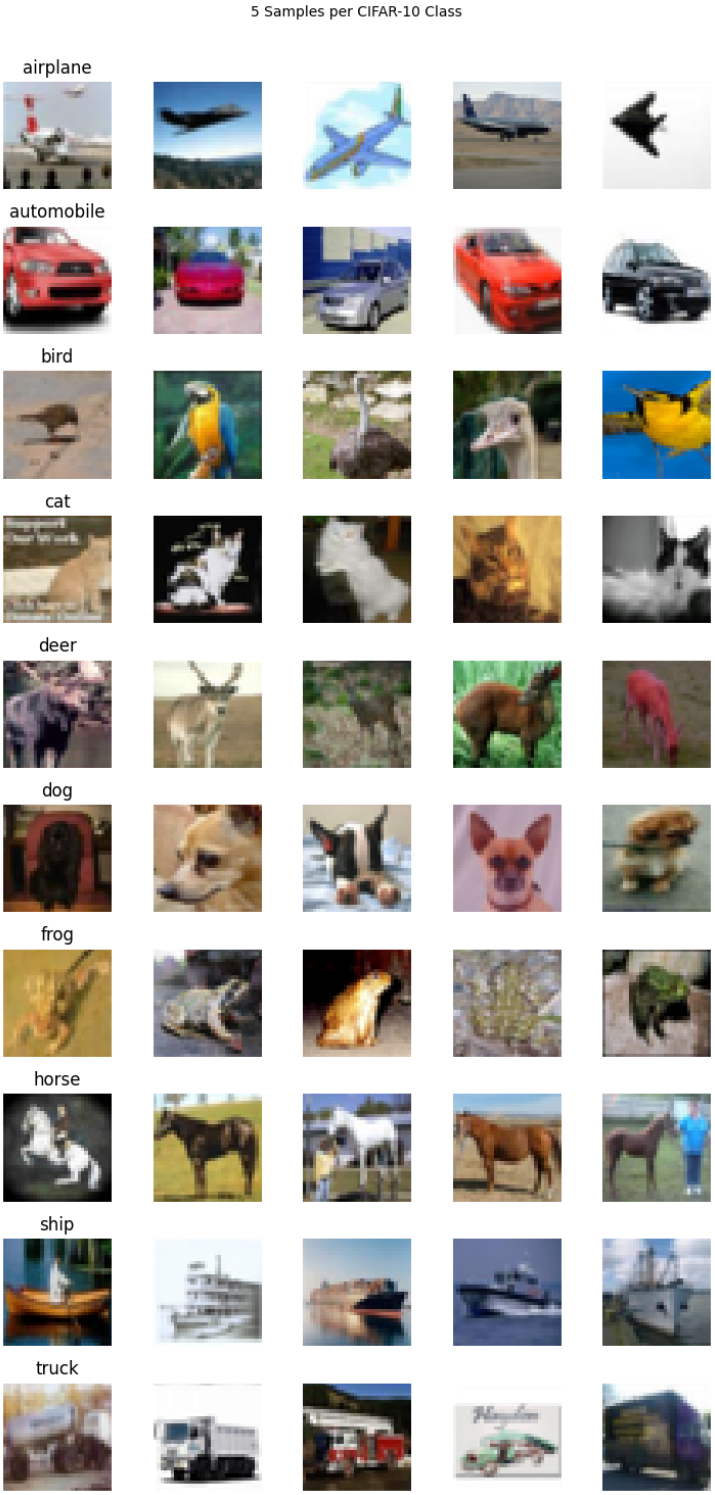
برای آموزش از تابع هزینه CrossEntropyLoss و بهینه‌ساز Adam استفاده شده است. نرخ یادگیری برای لایه‌های Head برابر 10^{-3} و برای لایه‌ی باز MobileNetV2 برابر 10^{-5} در نظر گرفته شده است.

پارامترهای کلیدی آزمایش‌ها:

- تعداد epoch: ۱۵
- batch size: ۶۴ (و ۳۲ برای تحلیل گرادیان)
- بهینه‌ساز: Adam با دو نرخ یادگیری مجزا
- Gradient Clipping: فقط در نیمی از آزمایش‌ها با مقدار ۱/۰

فرآیند آموزش شامل پایش دقت و مقدار Loss در هر epoch، ذخیره‌ی نتایج، و در صورت نیاز، تحلیل گرادیان‌ها و Loss Landscape برای مدل‌های مختلف است. همچنین، ابزار اختصاصی برای ثبت گرادیان لایه‌ها و مقایسه‌ی توزیع آن‌ها در دو حالت clipping و بدون clipping توسعه داده شده است[Zhang et al., 2019, Li et al., 2018].

[۹] BatchNorm, LayerNorm, FRN



شکل ۹: نمونه هایی از هر کلاس

۴ تحلیل تجربی

۱.۴ دقت و خطا

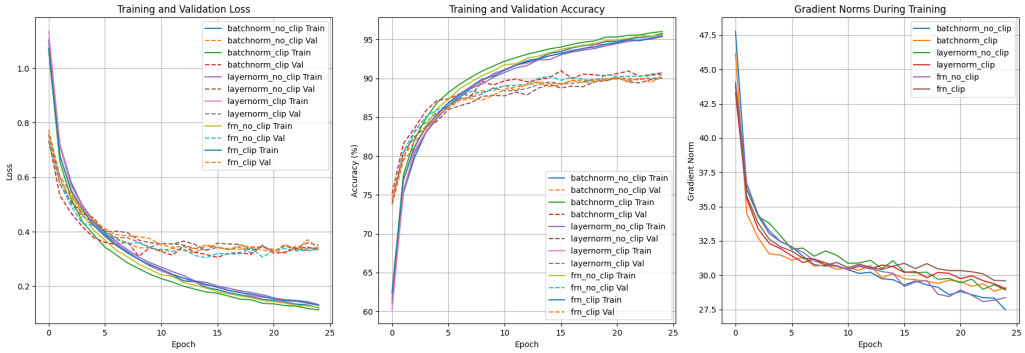
برای ارزیابی اثر ساختارهای مختلف نرمال‌سازی، شش آزمایش مجزا شامل ترکیب سه روش نرمال‌سازی [۱۰] با یا بدون تکنیک Gradient Clipping انجام شد. تمامی مدل‌ها روند صعودی نسبتاً همواری در دقت اعتبارسنجی طی ۲۵ دوره آموزش نشان دادند.

به‌طور خاص، بالاترین دقت نهایی متعلق به FRN-NoClip با مقدار 90.57% بود. پس از آن، مدل BatchNorm-Clip با 90.71% و مدل LayerNorm-NoClip با 90.22% قرار گرفتند. پایین‌ترین دقت نهایی مربوط به BatchNorm-NoClip با 89.64% بود.

از نظر همگرایی، مدل‌های دارای Clipping معمولاً در دوره‌های ابتدایی همگرایی سریع‌تری داشتند. به‌طور مثال، دقت BatchNorm-Clip در اولین اپک 75.16% بود، درحالی‌که نسخه بدون کلیپینگ 73.73% را ثبت کرد. با این حال، در دوره‌های پایانی، تفاوت عملکرد کاهش یافته و در برخی موارد مانند LayerNorm اثر Clipping حتی ناچیز یا منفی بود.

۲.۴ بررسی چشم‌انداز تابع هزینه

با وجود خطای فنی در رسم نمودار چشم‌انداز تابع هزینه ، روند همگرایی پایدار مدل‌های FRN و LayerNorm بدون نیاز به Clipping می‌تواند نشانه‌ای از ساختار نرم‌تر loss surface در این معماری‌ها باشد. همچنین ثبات نُرم گرادیان در طی زمان (کاهش تدریجی از حدود ۴۴ به زیر ۲۹) در این ساختارها مؤید همین موضوع است.



شکل ۱۰: آنالیز آموزش مدل ها

[۱۰] BatchNorm، LayerNorm، FRN

۵ بحث

نتایج تجربی نشان دادند که نوع نرمال‌سازی تأثیر مستقیمی بر پایداری و کیفیت آموزش دارد. ساختار FRN بدون کلیپینگ بالاترین دقت را به‌دست آورد و نشان داد که به‌طور ذاتی پایدار است.

مدل‌های LayerNorm نیز به‌ویژه در حالت بدون Clipping رفتار همگرای روان و پایداری را از خود نشان دادند، که می‌توان آن را به ماهیت مستقل بودن آن از آماره‌های Batch نسبت داد. روند تغییر نُرم گرادیان در این ساختار نیز بسیار منظم بود.

در مقابل، BatchNorm بدون کلیپینگ دارای نوسانات بیشتری بود. استفاده از Clipping در این ساختار باعث بهبود پایداری گرادیان و افزایش دقت نهایی از 89.64% به 90.71% شد. به‌طور مشخص، نُرم گرادیان در BatchNorm-Clip کنترل‌شده‌تر بود (از حدود ۱۵.۴۶ تا ۰۸.۲۹).

نقش گرادیان کلیپینگ در تمامی ساختارها یا مثبت بود (مانند BatchNorm) یا تأثیر خنثی داشت (مانند LayerNorm) و در هیچ‌کدام کاهش عملکرد ایجاد نکرد.

۶ نتیجه‌گیری و پیشنهادات

یافته‌های اصلی به‌شرح زیر است:

- ساختار FRN-NoClip بهترین عملکرد نهایی را با دقت 90.57% نشان داد.
- Clipping در ساختار BatchNorm منجر به بهبود عملکرد شد (افزایش 1.07%+).
- ساختارهای FRN و LayerNorm بدون نیاز به Clipping به عملکرد پایدار و دقیقی رسیدند.

محدودیت‌ها:

- آموزش فقط روی بخش Head مدل، با فریز بودن لایه‌های پایه
- استفاده از منابع سخت‌افزاری محدود (مانند Google Colab)
- عدم امکان ترسیم چشم‌انداز تابع هزینه برای برخی مدل‌ها به دلیل خطای list index out of range

پیشنهادهای توسعه آتی:

- اعمال نرمال‌سازی ترکیبی یا بررسی تکنیک‌های جدید مانند GroupNorm
- استفاده از مدل‌های سبک‌تر مانند EfficientNet-Lite
- انجام Fine-Tuning کامل و نه فقط Head
- ارزیابی روی داده‌های پیچیده‌تر مانند Tiny-ImageNet

۷ خلاصه مدیریتی

در این پروژه، سه ساختار نرمال‌سازی متفاوت به همراه یا بدون گرادیان کلیپینگ در فرآیند انتقال یادگیری ارزیابی شدند. یافته‌های کلیدی عبارت‌اند از:

- ساختار FRN-NoClip بهترین دقت را ارائه داد (90.57%).
- گرادیان کلیپینگ در ساختار BatchNorm باعث بهبود عملکرد شد (+1.07%).
- LayerNorm و FRN در اکثر موارد بدون نیاز به Clipping عملکرد پایداری داشتند.

پاسخ به سه سؤال کلیدی:

۱. بهترین تکنیک نرمال‌سازی: FRN-NoClip با دقت نهایی 90.57%

۲. تأثیر Clipping: در BatchNorm مؤثر و مثبت، در سایر روش‌ها تأثیر کم یا خنثی

۳. راهکارهای عملی: استفاده از LayerNorm/FRN در پروژه‌های واقعی با batch کوچک؛ فعال‌سازی Clipping در BatchNorm؛ توجه به محدودیت منابع در طراحی معماری Head

۸ ضمائم

لینک گوگل کولب تست‌ها

۹ منابع

References

Hao Li, Zheng Xu, Gavin Taylor, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS*, 2018.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Aarush Singh and Dilip Krishnan. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. In *CVPR*, 2020.

Zihan Zhang, Hao He, and Ruslan Salakhutdinov. Understanding and improving gradient clipping. *arXiv preprint arXiv:1905.11881*, 2019.