تکلیف اول درس شبکه‌های عصبی

وحید ملکی
شماره دانشجویی: ۴۰۳۱۳۰۰۴

۲۶ آبان ۱۴۰۴

سوال ۴

الف: محو شدن و انفجار گرادیان چیست؟

در آموزش شبکه‌های عصبی عمیق، گرادیان تابع هزینه نسبت به وزن‌های لایه‌های اولیه با استفاده از قاعده زنجیره‌ای و به صورت ضرب مکرر مشتق توابع فعال‌سازی و وزن‌های لایه‌های بعدی محاسبه می‌شود.

- **محو شدن گرادیان (Vanishing Gradient):** زمانی رخ می‌دهد که مقادیر مطلق مشتق توابع فعال‌سازی (مانند Sigmoid و Tanh) در نواحی اشباع کمتر از ۱ باشد. با افزایش تعداد لایه‌ها، ضرب مکرر اعداد کوچکتر از ۱ باعث می‌شود گرادیان در لایه‌های ابتدایی به صورت نمایی به صفر نزدیک شود. **نتیجه عملی:** لایه‌های اولیه تقریباً هیچ به‌روزرسانی دریافت نمی‌کنند، ویژگی‌های سطح پایین (مانند لبه‌ها و گوشه‌ها در بینایی ماشین) به خوبی یاد گرفته نمی‌شوند و کل شبکه عملاً به یک شبکه کم‌عمق تبدیل می‌شود.
- **انفجار گرادیان (Exploding Gradient):** زمانی رخ می‌دهد که مقادیر مطلق مشتق‌ها یا وزن‌ها بزرگتر از ۱ باشند. ضرب مکرر این اعداد باعث رشد نمایی گرادیان می‌شود و مقدار آن به سرعت به اعداد بسیار بزرگ یا حتی NaN تبدیل می‌شود. **نتیجه عملی:** وزن‌ها تغییرات بسیار شدیدی می‌کنند، تابع هزینه نوسانات شدید پیدا می‌کند و آموزش کاملاً ناپایدار می‌شود.

ب: روش کلاسیک و بسیار موفق — نرمال‌سازی دسته‌ای (Batch Normalization)

Batch Normalization که در سال ۲۰۱۵ توسط Sergey Ioffe و Christian Szegedy معرفی شد، تا چند سال استاندارد طلایی برای حل این دو مشکل بود.
فرمول کامل BatchNorm در زمان آموزش:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \quad , \quad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \quad , \quad y_i = \gamma \hat{x}_i + \beta$$

که γ و β پارامترهای قابل یادگیری هستند و در زمان استنتاج از میانگین و واریانس متحرک استفاده می‌شود.
چرا BatchNorm این مشکلات را حل می‌کند؟

- توزیع ورودی هر لایه را به میانگین صفر و واریانس یک می‌رساند ⇒ از ورود به نواحی اشباع توابع فعال‌سازی جلوگیری می‌کند.
- گرادیان‌ها را در مقیاس مناسب نگه می‌دارد ⇒ هم از محو شدن و هم از انفجار جلوگیری می‌کند.
- اجازه استفاده از نرخ یادگیری بسیار بالاتر را می‌دهد.
- اثر منظم‌سازی دارد و نیاز به Dropout را کاهش می‌دهد.

معایب مهم BatchNorm:

- وابستگی شدید به سایز بچ (در Batch Size کوچک، تخمین μ و σ نویزدار می‌شود)
- افزایش مصرف حافظه (نیاز به ذخیره آمار بچ)
- مشکل در شبکه‌های بازگشتی و آموزش آنلاین
- تفاوت رفتار بین آموزش و استنتاج

ج: مقاله علمی پیشنهادی — حذف کامل نرمال‌سازی با روشی هوشمندانه

عنوان مقاله: High-Performance Large-Scale Image Recognition Without Normalization
نویسندگان: Andrew Brock، Soham De، Samuel L. Smith (همه از DeepMind)
کنفرانس/سال: ICML 2021
لینک کامل: `https://arxiv.org/abs/2102.06171`
این مقاله یکی از مهم‌ترین مقالات سال ۲۰۲۱ در حوزه پایداری آموزش شبکه‌های عمیق است و نشان می‌دهد که نه تنها می‌توان BatchNorm را حذف کرد، بلکه با حذف آن می‌توان به عملکرد بهتری هم رسید!

ایده اصلی مقاله: شبکه‌های بدون نرمال‌سازی (Normalizer-Free Nets یا NFNet)

نویسندگان مشاهده کردند که BatchNorm در واقع دو نقش دارد: ۱. پایدار کردن آموزش (جلوگیری از Van-ishing/Exploding Gradient) ۲. منظم‌سازی
آنها نشان دادند که نقش اول را می‌توان با روشی بسیار ساده‌تر و کارآمدتر جایگزین کرد: **برش تطبیقی گرادیان (Adaptive Gradient Clipping — AGC)**

توضیح کامل روش Adaptive Gradient Clipping (AGC)

روش کلاسیک Gradient Clipping یک آستانه ثابت (مثلاً ۰.۱) روی نرم گرادیان کل شبکه اعمال می‌کند. اما AGC هوشمندانه‌تر عمل می‌کند:
برای هر پارامتر w در هر لایه، نسبت نرم گرادیان به نرم وزن محاسبه می‌شود:

$$r = \frac{\|\nabla_w \mathcal{L}\|_2}{\|w\|_2 + \epsilon}$$

اگر این نسبت از یک آستانه λ بیشتر باشد، گرادیان آن لایه به صورت زیر کوچک می‌شود:

$$\nabla_w \mathcal{L} \leftarrow \nabla_w \mathcal{L} \times \min\left(1,\ \frac{\lambda}{r}\right)$$

به عبارت ساده‌تر: «تغییر وزن در هر قدم نباید بیشتر از λ برابر مقدار فعلی وزن باشد.»

این کار باعث می‌شود: - وزن‌ها هرگز تغییرات بسیار شدید نداشته باشند $\Rightarrow$ انفجار گرادیان غیرممکن می‌شود. - در عین حال گرادیان‌های مفید و معقول حفظ می‌شوند (برخلاف برش ثابت که ممکن است گرادیان‌های بزرگ اما مفید را هم قطع کند).

نتایج شگفت‌انگیز مقاله

- مدل NFNet-F6 بدون هیچ نوع نرمال‌سازی (نه BatchNorm، نه LayerNorm، نه GroupNorm) به دقت 88.5% Top-1 در ImageNet رسید که در زمان انتشار رکورد جدیدی بود.
- آموزش پایدار حتی با Batch Size = ۱۶ یا کمتر.
- سرعت آموزش تا ۸۰٪ بالاتر از مدل‌های معادل با BatchNorm.
- مصرف حافظه به شدت کاهش یافت.
- شبکه‌های تا بیش از ۱۰۰۰ لایه بدون مشکل آموزش داده شدند.

مزایای روش AGC و NFNet

- حذف کامل تمام لایه‌های نرمال‌سازی $\Rightarrow$ کد ساده‌تر، سرعت بالاتر، حافظه کمتر
- کاملاً مستقل از سایز بچ $\Rightarrow$ ایده‌آل برای آموزش توزیع‌شده و دستگاه‌های با حافظه محدود
- امکان استفاده از نرخ یادگیری بسیار بالا
- جلوگیری همزمان و مؤثر از هر دو مشکل Vanishing و Exploding Gradient
- دستیابی به دقت State-of-the-Art بدون هیچ ترفند اضافی
- قابل تعمیم به معماری‌های دیگر (بعداً روی Transformerها و ViTها هم موفق بوده)

محدودیت‌ها و نکات عملی

- نیاز به تنظیم دقیق λ (معمولاً بین ۰٫۰۱ تا ۰٫۱ بسته به معماری)
- محاسبه نرم وزن در هر لایه هزینه محاسباتی جزئی اضافه می‌کند (ولی بسیار کمتر از BatchNorm)
- در مقاله اصلی فقط روی شبکه‌های کانولوشنی تست شده بود (هرچند مقالات بعدی نشان دادند روی Transformerها هم عالی کار می‌کند)

جمع‌بندی نهایی

مشکلات Vanishing و Exploding Gradient از مهم‌ترین موانع آموزش شبکه‌های عمیق بودند که با Batch Normalization تا حد زیادی حل شدند. اما مقاله ICML 2021 از DeepMind با معرفی Adaptive Gradient Clipping نشان داد که می‌توان بدون هیچ نوع نرمال‌سازی و فقط با یک خط کد ساده، شبکه‌های عمیق‌تر، سریع‌تر و دقیق‌تری ساخت. این روش امروزه یکی از قدرتمندترین و مدرن‌ترین تکنیک‌های پایدارسازی آموزش شبکه‌های عصبی عمیق محسوب می‌شود و به سرعت در جامعه تحقیقاتی و صنعتی در حال گسترش است.