تقطیر دانش در یک شبکه عصبی

خلاصه geoffhinton@google.com
Oriol Vinyals†
.شرکت گوگل
منظره کوهستانی
vinyals@google.com

جف دین
.شرکت گوگل
منظره کوهستانی
jeff@google.com

جفری هینتون†*
.شرکت گوگل
منظره کوهستانی

یک راه بسیار ساده برای بهبود عملکرد تقریباً هر الگوریتم یادگیری ماشینی، آموزش مدل‌های مختلف بر روی داده‌های یکسان و سپس میانگین‌گیری پیش‌بینی‌های آنها است [3]. متأسفانه، پیش‌بینی با استفاده از مجموعه‌ای از مدل‌ها دست و پا گیر است و ممکن است از نظر محاسباتی بسیار پرهزینه باشد تا امکان استقرار برای تعداد زیادی از کاربران را فراهم کند، به خصوص اگر مدل‌های جداگانه شبکه‌های عصبی بزرگ باشند. کاروانا و همکارانش [1] نشان داده‌اند که می‌توان دانش را در یک مجموعه به یک مدل واحد فشرده کرد که به کارگیری آن بسیار آسان‌تر است و ما این رویکرد را با استفاده از یک تکنیک فشرده‌سازی دست یافتیم و نشان دادیم که MNIST متفاوت توسعه می‌دهیم. ما به نتایج شگفت‌آوری در می‌توانیم مدل صوتی یک سیستم تجاری پرکاربرد را با تقطیر دانش مجموعه‌ای از مدل‌ها در یک مدل به طور قابل توجهی بهبود بخشیم. ما همچنین نوع جدیدی از گروه متشکل از یک یا چند مدل کامل و بسیاری از مدل‌های تخصصی را معرفی می‌کنیم که یاد می‌گیرند کلاس‌های ریزدانه را که مدل‌های کامل گیج می‌کنند، تشخیص دهند. بر خلاف ترکیبی از متخصصان، .این مدل های متخصص را می توان به سرعت و به صورت موازی آموزش داد

1. معرفی

بسیاری از حشرات دارای فرم لارو هستند که برای استخراج انرژی و مواد مغذی از محیط بهینه شده است و شکل بالغ کاملاً متفاوتی دارند که برای نیازهای بسیار متفاوت سفر و تولید مثل بهینه شده است. در یادگیری ماشینی در مقیاس بزرگ، ما معمولاً از مدل های بسیار مشابه برای مرحله آموزش و مرحله استقرار با وجود الزامات بسیار متفاوت آنها استفاده می کنیم: برای کارهایی مانند تشخیص گفتار و اشیا، آموزش باید ساختار را از مجموعه داده های بسیار بزرگ و بسیار زائد استخراج کند، اما این کار را نمی کند. نیاز به کار در زمان واقعی دارد و می تواند از مقدار زیادی محاسبات استفاده کند. با این حال، استقرار برای تعداد زیادی از کاربران الزامات بسیار سخت گیرانه تری در تأخیر و منابع محاسباتی دارد. قیاس با حشرات نشان می دهد که اگر این امر استخراج ساختار از داده ها را آسان تر می کند، باید مایل به آموزش مدل های بسیار دست و پا گیر باشیم. مدل دست و پا گیر می تواند مجموعه ای از مدل های آموزش دیده جداگانه یا یک مدل بسیار بزرگ منفرد باشد که با تنظیم کننده بسیار قوی مانند ترک تحصیل [9] آموزش دیده است. هنگامی که مدل دست و پا گیر آموزش داده شد، می توانیم از نوع دیگری از آموزش استفاده کنیم که آن را "تقطیر" می نامیم تا دانش را از مدل دست و پا گیر به مدل کوچکی که برای استقرار مناسب تر است، منتقل کنیم. نسخه ای از این استراتژی قبلاً توسط ریچ کاروانا و همکارانش ارائه شده است [1]. آنها در مقاله مهم خود به طور متقاعدکننده ای نشان می دهند که دانش کسب شده توسط مجموعه .بزرگی از مدل ها می تواند به یک مدل کوچک منتقل شود

یک بلوک مفهومی که ممکن است مانع از بررسی بیشتر این رویکرد بسیار امیدوارکننده شده باشد این است که ما تمایل داریم دانش را در یک مدل آموزش‌دیده با مقادیر پارامترهای آموخته شده شناسایی کنیم و این باعث می‌شود که نتوانیم ببینیم چگونه می‌توانیم شکل مدل را تغییر دهیم، اما آن را حفظ کنیم. همان دانش دیدگاه انتزاعی تر از

.دانش، که آن را از هر مصداق خاصی رها می کند، این است که دانش آموخته شده است

همچنین به دانشگاه تورنتو و موسسه تحقیقات پیشرفته کانادایی وابسته است.†سهم برابر*

نگاشت از بردارهای ورودی به بردارهای خروجی. برای مدل‌های دست و پا گیر که یاد می‌گیرند بین تعداد زیادی کلاس تمایز قائل شوند، هدف عادی آموزش به حداکثر رساندن میانگین احتمال ثبت پاسخ صحیح است، اما یک اثر جانبی یادگیری این است که مدل آموزش‌دیده احتمالات را به همه موارد نادرست اختصاص می‌دهد. پاسخ ها و حتی زمانی که این احتمالات بسیار کوچک هستند، برخی از آنها بسیار بزرگتر از دیگران هستند. احتمالات نسبی پاسخ های نادرست به ما چیزهای زیادی در مورد اینکه چگونه مدل دست و پا گیر تمایل به تعمیم دارد به ما می ممکن است تنها شانس بسیار کمی برای اشتباه گرفتن با یک ،BMW گوید. برای مثال، یک تصویر از یک .کامیون زباله داشته باشد، اما این اشتباه هنوز چند برابر بیشتر از اشتباه گرفتن آن با یک هویج است

به طور کلی پذیرفته شده است که تابع هدف مورد استفاده برای آموزش باید هدف واقعی کاربر را تا حد امکان منعکس کند. با وجود این، مدل‌ها معمولاً برای بهینه‌سازی عملکرد در داده‌های آموزشی زمانی که هدف واقعی تعمیم داده‌های جدید است، آموزش داده می‌شوند. واضح است که بهتر است مدل ها را به خوبی تعمیم دهیم، اما این نیاز به اطلاعاتی در مورد روش صحیح تعمیم دارد و این اطلاعات معمولاً در دسترس نیست. با این حال، وقتی دانش را از یک مدل بزرگ به یک مدل کوچک تقطیر می کنیم، می توانیم مدل کوچک را برای تعمیم به همان روشی که مدل بزرگ است آموزش دهیم. اگر مدل دست و پا گیر به خوبی تعمیم داده شود زیرا، برای مثال، میانگین یک مجموعه بزرگ از مدل های مختلف است، یک مدل کوچک که برای تعمیم به همان روش آموزش داده شده است، معمولاً در داده های آزمایشی بسیار بهتر از یک مدل کوچک است که در آن آموزش داده شده است. .به روش عادی در همان مجموعه تمرینی که برای تمرین گروه استفاده شد

یک راه واضح برای انتقال توانایی تعمیم مدل دست و پا گیر به یک مدل کوچک، استفاده از احتمالات کلاس تولید شده توسط مدل دست و پا گیر به عنوان "هدف های نرم" برای آموزش مدل کوچک است. برای این مرحله انتقال، می‌توانیم از همان مجموعه آموزشی یا یک مجموعه «انتقال» جداگانه استفاده کنیم. هنگامی که مدل دست و پا گیر مجموعه بزرگی از مدل های ساده تر است، می توانیم از میانگین حسابی یا هندسی توزیع های پیش بینی فردی آنها به عنوان اهداف نرم استفاده کنیم. هنگامی که اهداف نرم دارای آنتروپی بالا هستند، اطلاعات بسیار بیشتری را در هر مورد آموزشی نسبت به اهداف سخت و واریانس بسیار کمتری در گرادیان بین موارد آموزشی ارائه می دهند، بنابراین مدل کوچک اغلب می تواند بر روی داده های بسیار کمتری نسبت به مدل دست و پا گیر اصلی آموزش داده شود و با استفاده از یک نرخ یادگیری بسیار بالاتر

که در آن مدل دست و پا گیر تقریباً همیشه پاسخ صحیح را با اطمینان بسیار بالا MNIST برای کارهایی مانند تولید می کند، بیشتر اطلاعات مربوط به تابع آموخته شده در نسبت احتمالات بسیار کوچک در اهداف نرم قرار دارد. برای مثال، ممکن است به یک نسخه از 2 احتمال داده شود10^{-6}از 3 بودن و10^{-9}از 7 بودن در حالی که برای نسخه دیگر ممکن است برعکس باشد. این اطلاعات ارزشمندی است که ساختار شباهت غنی را بر روی داده ها تعریف می کند (من. ه.می گوید کدام 2 شبیه به 3 و کدام شبیه به 7 است) اما تأثیر بسیار کمی بر تابع هزینه آنتروپی متقابل در مرحله انتقال دارد زیرا احتمالات بسیار نزدیک به صفر هستند. کاروانا و همکارانش این مشکل را با استفاده از لاجیت ها (ورودی های سافت مکس نهایی) به جای احتمالات تولید شده توسط سافت مکس به عنوان اهداف یادگیری مدل کوچک دور می زنند و اختلاف مجذور بین لاجیت های تولید شده توسط مدل دست و پا گیر و دست و پا گیر را به حداقل می رسانند. لاجیت های تولید شده توسط مدل کوچک. راه‌حل کلی‌تر ما که نهایی است تا زمانی که مدل دست و پا گیر مجموعه نرم مناسبی softmax «تقطیر» نامیده می‌شود، افزایش دمای از اهداف را تولید کند. سپس هنگام آموزش مدل کوچک برای مطابقت با این اهداف نرم از همان دمای بالا استفاده .می کنیم. بعداً نشان می‌دهیم که تطبیق لجیت‌های مدل دست و پاگیر در واقع یک مورد خاص از تقطیر است

مجموعه انتقالی که برای آموزش مدل کوچک استفاده می‌شود می‌تواند کاملاً از داده‌های بدون برچسب [1] تشکیل شده باشد یا می‌توانیم از مجموعه آموزشی اصلی استفاده کنیم. ما دریافتیم که استفاده از مجموعه آموزشی اصلی به خوبی کار می کند، به خصوص اگر یک عبارت کوچک به تابع هدف اضافه کنیم که مدل کوچک را تشویق به پیش بینی اهداف واقعی و همچنین مطابقت با اهداف نرم ارائه شده توسط مدل دست و پا گیر کند. به طور معمول، .مدل کوچک نمی تواند دقیقاً با اهداف نرم مطابقت داشته باشد و خطا در جهت پاسخ صحیح مفید است

تقطیر 2

را تبدیل می کند، احتمالات کلاس logit که "softmax" شبکه های عصبی معمولاً با استفاده از یک لایه خروجی $_{من}$، با مقایسهبا$_{من}$با لاجیت های دیگرq،را تولید می کنند.با$_{من}$، برای هر کلاس به یک احتمال محاسبه می شود

$$q_{من}=\exp(z_{من}/T)$$

$$_{j}\exp(z_{j/T})(1)$$

2

جایی کهتیدمایی است که معمولاً روی آن تنظیم می شود1. استفاده از مقدار بالاتر برایتیتوزیع احتمال ملایم تری بر طبقات ایجاد می کند.

در ساده ترین شکل تقطیر، دانش با آموزش آن بر روی یک مجموعه انتقال و استفاده از توزیع هدف نرم برای هر مورد در مجموعه انتقال که با استفاده از مدل دست و پا گیر با دمای بالا در سافت مکس آن تولید می شود، به مدل تقطیر منتقل می شود. هنگام آموزش مدل مقطر از همان دمای بالا استفاده می شود، اما پس از آموزش از دمای 1 استفاده می کند.

هنگامی که برچسب های صحیح برای تمام یا برخی از مجموعه انتقال شناخته می شوند، این روش را می توان با آموزش مدل تقطیر شده برای تولید برچسب های صحیح به طور قابل توجهی بهبود بخشید. یکی از راه‌های انجام این کار استفاده از برچسب‌های صحیح برای اصلاح اهداف نرم است، اما ما دریافتیم که راه بهتر این است که به سادگی از میانگین وزنی دو تابع هدف مختلف استفاده کنیم. تابع هدف اول، آنتروپی متقاطع با اهداف نرم است و مدل تقطیر شده محاسبه می‌شود که برای تولید softmax این آنتروپی متقاطع با استفاده از همان دمای بالا در اهداف نرم از مدل دست و پا گیر استفاده شد. تابع هدف دوم، آنتروپی متقاطع با برچسب های صحیح است. این مدل تقطیر شده اما در دمای 1 محاسبه می‌شود. ما دریافتیم که softmax دقیقاً با استفاده از لجیت‌های مشابه در بهترین نتایج عموماً با استفاده از وزن بسیار کمتر در تابع هدف دوم بدست می‌آید. از آنجایی که بزرگی شیب های مهم است که آنها را در ضرب کنیمتی^{2}هنگام استفاده از اهداف سخت و T^2/تولید شده توسط اهداف نرم به اندازه1 نرم. این تضمین می کند که سهم نسبی اهداف سخت و نرم تقریباً بدون تغییر باقی می ماند اگر دمای مورد استفاده برای تقطیر در حین آزمایش با متا پارامترها تغییر کند.

تطبیق لجیت ها یک مورد خاص از تقطیر است 2.1

$_{من}$با توجه به هر لاجیت،با$_{من}$از مدلdC/dz،هر مورد در مجموعه انتقال یک گرادیان آنتروپی متقاطع کمک می کند مقطر اگر مدل دست و پا گیر دارای لاجیت باشدکه در$_{من}$که احتمالات هدف نرم را تولید می کنندپس$_{من}$و آموزش انتقال در دمای انجام می شودتی، این گرادیان توسط:

$$\frac{\partial C}{\partial z_{\text{تی}}} = \frac{1}{\text{تی}}(\text{قمن} - \text{صمن}) = \frac{1}{\text{تی}}(\text{این استکه در } j \text{ /این استبا } j \text{ تی} - \text{این استبا من/ تی})$$

اگر دما در مقایسه با بزرگی لجیت ها زیاد باشد، می توانیم تقریبی کنیم:

$$\frac{\partial C}{\partial z_{j}} \approx \frac{1}{N}\left(\frac{1}{T}\right)\left(v_{j} - v_{من} + \frac{1}{N}\sum_{j'} v_{j'}\right)$$

تی

$$\frac{\partial z1_{\text{من}}}{\partial z_{\text{تی}}} \approx \frac{1}{1 + z_{\text{تی}} / z_{\text{من}}} \qquad (\text{تی}(3 / \text{که در}_j + N$$

برای هر مورد انتقال به طوری که

اگر اکنون فرض کنیم که لجیت ها به معنای صفر بوده اند

معادله 3 ساده می کند $0 = j$ که در $j = \text{با}_j$:

$$\frac{\partial C}{\partial z1_{\text{من}}}$$

$$NT^2(\text{من})(4v - \text{با}_{\text{من}})$$

ها به $\text{logit}_{\text{من}})^2$، مشروط بر اینکه $v - z_{\text{من}})$بنابراین در حد دمای بالا، تقطیر معادل به حداقل رساندن است1/2 طور جداگانه برای هر مورد انتقال صفر معنی شوند. در دماهای پایین تر، تقطیر به لجیت های تطبیقی که بسیار ها تقریباً به طور logit منفی تر از میانگین هستند توجه کمتری می کند. این به طور بالقوه سودمند است زیرا این کامل توسط تابع هزینه استفاده شده برای آموزش مدل دست و پاگیر محدود نمی شوند بنابراین می توانند بسیار پر سر و صدا باشند. از سوی دیگر، لوجیت های بسیار منفی ممکن است اطلاعات مفیدی را در مورد دانش به دست آمده توسط مدل دست و پا گیر منتقل کنند. کدام یک از این تأثیرات غالب است، یک سؤال تجربی است. ما نشان می‌دهیم که وقتی مدل تقطیر شده برای به دست آوردن تمام دانش در مدل دست و پا گیر خیلی کوچک است، دمای متوسط بهترین کار را دارد که قویاً نشان می‌دهد که نادیده گرفتن لجیت‌های منفی بزرگ می‌تواند مفید باشد.

آزمایش مقدماتی بر روی MNIST 3

برای اینکه بفهمیم تقطیر چقدر خوب عمل می‌کند، یک شبکه عصبی بزرگ را با دو لایه پنهان از 1200 واحد مخفی خطی اصلاح‌شده در تمام 60000 مورد آموزشی آموزش دادیم. شبکه به شدت با استفاده از افت تحصیلی و محدودیت های وزنی که در [5] توضیح داده شد، منظم شد. ترک تحصیل را می توان به عنوان راهی برای آموزش مجموعه ای بزرگ از مدل هایی که وزن های مشترک دارند در نظر گرفت. علاوه بر این، تصاویر ورودی بودند

تا دو پیکسل در هر جهتی تکان می خورد. این شبکه به 67 خطای آزمایشی دست یافت، در حالی که یک شبکه کوچکتر با دو لایه پنهان از 800 واحد پنهان خطی اصلاح شده و بدون تنظیم، 146 خطا را به دست آورد. اما اگر شبکه کوچکتر صرفاً با اضافه کردن کار اضافی تطبیق اهداف نرم تولید شده توسط تور بزرگ در دمای 20 درجه تنظیم شود، به 74 خطای آزمایشی دست یافت. این نشان می‌دهد که اهداف نرم می‌توانند مقدار زیادی دانش را به مدل مقطر منتقل کنند، از جمله دانش در مورد نحوه تعمیم آن‌هایی که از داده‌های آموزشی ترجمه‌شده آموخته می‌شود، حتی اگر مجموعه انتقال حاوی هیچ ترجمه‌ای نباشد.

هنگامی که شبکه تقطیر شده 300 واحد یا بیشتر در هر یک از دو لایه پنهان خود داشت، تمام دماهای بالاتر از 8 نتایج تقریباً مشابهی داشتند. اما زمانی که این میزان به طور اساسی به 30 واحد در هر لایه کاهش یافت، دمای بین 2.5 تا 4 به طور قابل توجهی بهتر از دماهای بالاتر یا پایین تر عمل کرد.

سپس سعی کردیم تمام نمونه های رقم 3 را از مجموعه انتقال حذف کنیم. بنابراین از منظر مدل تقطیر شده، 3 یک رقم افسانه ای است که هرگز ندیده است. با وجود این، مدل تقطیر شده تنها 206 خطای آزمایشی دارد که 133 مورد آن روی 1010 سه خطای مجموعه تست است. بیشتر خطاها ناشی از این واقعیت است که سوگیری آموخته شده برای کلاس 3 بسیار کم است. اگر این سوگیری 3.5 افزایش یابد (که عملکرد کلی را در مجموعه آزمایشی بهینه می‌کند)، مدل تقطیر شده 109 خطا می‌کند که 14 مورد آن در 3 ثانیه است. بنابراین با تعصب مناسب، مدل تقطیر شده 98.6٪ از تست های 3 را درست می گیرد، علیرغم اینکه هرگز در طول تمرین 3 ندیده است. اگر از مجموعه آموزشی، مدل تقطیر شده 47.3 درصد خطای تست دارد، اما زمانی sو 8 sمجموعه انتقال شاملفقط7 که بایاس های 7 و 8 برای بهینه سازی عملکرد تست 7.6 کاهش می یابد، این به 13.2 درصد خطاهای تست می

.رسد

آزمایش در تشخیص گفتار 4

که در تشخیص (DNN) در این بخش، ما به بررسی اثرات شبیه‌سازی مدل‌های صوتی شبکه عصبی عمیق استفاده می‌شوند، می‌پردازیم. ما نشان می‌دهیم که استراتژی تقطیر که در این مقاله پیشنهاد (ASR) خودکار گفتار می‌کنیم، به اثر مطلوب تقطیر مجموعه‌ای از مدل‌ها به یک مدل واحد دست می‌یابد که به طور قابل‌توجهی بهتر از .مدلی با همان اندازه که مستقیماً از همان داده‌های آموزشی آموخته می‌شود، کار می‌کند

برای ترسیم یک بافت زمانی (کوتاه) از ویژگی‌های مشتق شده DNN در حال حاضر از ASR سیستم‌های پیشرفته استفاده می‌کنند (HMM) از شکل موج به یک توزیع احتمال بر روی حالت‌های گسسته یک مدل مارکوف پنهان یک توزیع احتمال را بر روی خوشه‌هایی از حالت‌های سه‌تلفن در هر زمان ایجاد DNN ،[4]. به طور خاص پیدا می‌کند که بهترین سازش بین استفاده از HMM می‌کند و سپس یک رمزگشا مسیری را از طریق حالت‌های .حالت‌های احتمال بالا و تولید رونویسی است که در زبان محتمل است. مدل

به گونه ای آموزش داده شود که رمزگشا (و بنابراین، مدل زبان) با به DNN (اگرچه ممکن است (و مطلوب برای انجام فریم معمول است. طبقه‌بندی DNN حاشیه راندن تمام مسیرهای ممکن در نظر گرفته شود، آموزش فریمی (محلی) با به حداقل رساندن آنتروپی متقاطع بین پیش‌بینی‌های انجام‌شده توسط شبکه و برچسب‌های ارائه‌شده :توسط هم‌ترازی اجباری با دنباله حقیقت زمینی حالت‌ها برای هر مشاهده

$$\theta = \text{ارگ حداکثر} \; P(h_{\text{تی}}|s_{\text{من}}; \text{"من"})$$

جایی کهمنپارامترهای مدل آکوستیک ما هستندپکه مشاهدات صوتی را در زمان ترسیم می کندتی،سیبه یک ساعتی، که با تراز اجباری با توالی صحیح کلمات مشخصHMM "تی; من")، از حالت "صحیحs|تیP(h،احتمال .می شود. این مدل با یک رویکرد نزولی گرادیان تصادفی توزیع شده آموزش داده شده است

ما از یک معماری با 8 لایه پنهان استفاده می کنیم که هر کدام شامل 2560 واحد خطی اصلاح شده و یک لایه است.ساعتی). ورودی 26 فریم از فیلتر 40 (HMM نهایی با 14000 برچسب (هدف های softmax را پیش بینی می 21 HMM است ضرایب بانکی با پیشرفت 10 میلی ثانیه در هر فریم و ما وضعیت Mel-scaled است. این یک نسخه کمی قدیمی از مدل آکوستیک است که توسط Mکنیمخیابانقاب تعداد کل پارامترها حدود 85 جستجوی صوتی اندروید استفاده می شود و باید به عنوان یک پایه بسیار قوی در نظر گرفته شود. برای آموزش ما از حدود 2000 ساعت داده انگلیسی گفتاری استفاده می کنیم که حدود 700 میلیون نمونه DNN مدل آکوستیک Word (WER) آموزشی به دست می دهد. این سیستم در مجموعه توسعه ما به دقت فریم 58.9% و نرخ خطای .دست می یابد 10.9%

4

WER	دقت قاب	تست سیستم
10.9 ٪	58.9٪	پایه٪
10.7%	61.1%	10xEnsemble
10.7 %	60.8%	%مدل تکی مقطر

نشان می‌دهد که مدل تک تقطیر شده تقریباً به خوبی پیش‌بینی‌های میانگین WER جدول 1: دقت طبقه‌بندی قاب و .10 مدلی را که برای ایجاد اهداف نرم استفاده شده‌اند، عمل می‌کند

نتایج 4.1

; من)، دقیقاً از همان معماری و قطار استفاده می $P(h_{تی}|s$ ما 10 مدل جداگانه برای پیش بینی آموزش دادیم $_{تی}$ کند

رویه به عنوان خط پایه مدل‌ها به‌طور تصادفی با مقادیر پارامترهای اولیه مختلف مقداردهی اولیه می‌شوند و متوجه می‌شویم که این تنوع کافی در مدل‌های آموزش‌دیده ایجاد می‌کند تا به پیش‌بینی‌های میانگین مجموعه اجازه می‌دهد تا به طور قابل‌توجهی از مدل‌های فردی بهتر عمل کنند. ما افزودن تنوع به مدل‌ها را با تغییر مجموعه داده‌هایی که هر مدل می‌بیند بررسی کرده‌ایم، اما متوجه شدیم که این به طور قابل‌توجهی نتایج ما را تغییر نمی‌دهد، بنابراین رویکرد ساده‌تری را انتخاب کردیم. برای تقطیر ما دمای آن را امتحان کردیم[1، 2، 5، 10]و از وزن نسبی 0.5 بر روی آنتروپی متقاطع برای اهداف سخت استفاده کرد که در آن فونت پررنگ بهترین مقدار استفاده شده برای جدول 1 را نشان می دهد.

جدول 1 نشان می دهد که، در واقع، رویکرد تقطیر ما قادر است اطلاعات مفیدتری را از مجموعه آموزشی به جای استفاده از برچسب های سخت برای آموزش یک مدل واحد استخراج کند. بیش از 80 درصد از بهبود دقت طبقه‌بندی فریم که با استفاده از مجموعه‌ای از 10 مدل به دست می‌آید، به مدل تقطیر شده منتقل می‌شود که شبیه به مشاهده کردیم. این مجموعه به دلیل عدم تطابق در تابع MNIST بهبودی است که در آزمایش‌های اولیه خود در در یک مجموعه آزمایشی 23 هزار کلمه‌ای) ایجاد می‌کند، اما مجدداً،) WER هدف، بهبود کمتری در هدف نهایی به دست آمده توسط گروه به مدل تقطیر شده منتقل می‌شود WER بهبود.

ما اخیراً از کار مرتبط با یادگیری یک مدل آکوستیک کوچک با تطبیق احتمالات کلاس یک مدل بزرگتر که قبلاً آموزش دیده است، آگاه شده ایم [8]. با این حال، آنها تقطیر را در دمای 1 با استفاده از یک مجموعه داده بزرگ بدون برچسب انجام می‌دهند و بهترین مدل تقطیر شده آنها تنها 28 درصد از فاصله بین نرخ خطای مدل‌های بزرگ و کوچک را کاهش می‌دهد. آموزش دیده با برچسب های سخت

گروه آموزشی از متخصصان در مورد مجموعه داده های بسیار بزرگ 5

آموزش مجموعه ای از مدل ها یک راه بسیار ساده برای استفاده از محاسبات موازی است و با استفاده از تقطیر می توان با این ایراد معمول که یک مجموعه نیاز به محاسبات بیش از حد در زمان آزمایش دارد، مقابله کرد. با این حال، یک ایراد مهم دیگر به مجموعه‌ها وجود دارد: اگر مدل‌های منفرد شبکه‌های عصبی بزرگ باشند و مجموعه داده بسیار بزرگ باشد، مقدار محاسبات مورد نیاز در زمان آموزش بیش از حد است، حتی اگر موازی کردن آن آسان باشد.

در این بخش، نمونه‌ای از چنین مجموعه داده‌هایی را ارائه می‌دهیم و نشان می‌دهیم که چگونه یادگیری مدل‌های تخصصی که هر کدام بر روی یک زیرمجموعه گیج‌انگیز متفاوت از کلاس‌ها تمرکز می‌کنند، می‌تواند کل مقدار محاسبات مورد نیاز برای یادگیری یک مجموعه را کاهش دهد. مشکل اصلی متخصصانی که بر روی ایجاد تمایزهای ریز تمرکز می‌کنند این است که آنها به راحتی بیش از حد قرار می‌گیرند و ما توضیح می‌دهیم که چگونه می‌توان با استفاده از اهداف نرم از این بیش از حد برازش جلوگیری کرد.

5.1 مجموعه داده JFT

یک مجموعه داده داخلی گوگل است که دارای 100 میلیون تصویر برچسب دار با 15000 برچسب است. JFT یک شبکه عصبی کانولوشن عمیق [7] بود که برای JFT وقتی ما این کار را انجام دادیم، مدل پایه گوگل برای حدود شش ماه با استفاده از نزول گرادیان تصادفی ناهمزمان بر روی تعداد زیادی از هسته‌ها آموزش داده شده بود. در این آموزش از دو نوع موازی [2] استفاده شد. اول، کپی‌های زیادی از شبکه عصبی وجود داشت که روی مجموعه‌های مختلف هسته‌ها اجرا می‌شد و مینی‌بچ‌های مختلفی را از مجموعه آموزشی پردازش می‌کرد. هر ماکت، گرادیان متوسط را در مینی دسته فعلی خود محاسبه می کند و این گرادیان را به یک سرور پارامتر تقسیم شده ارسال می کند که مقادیر جدیدی را برای پارامترها ارسال می کند. این مقادیر جدید منعکس کننده همه گرادیان های دریافت شده توسط سرور پارامتر از آخرین باری است که پارامترها را به ماکت ارسال کرده است. دوم، هر ماکت با قرار دادن زیرمجموعه های مختلف نورون ها روی هر هسته، روی چندین هسته پخش می شود. آموزش گروهی هنوز سومین نوع موازی سازی است که می تواند پیچیده شود

JFT 1: مهمانی چای؛ عید پاک؛ دوش عروس؛ حمام نوزاد؛ اسم حیوان دست اموز عید پاک؛ ...
JFT 2: پل؛ پل کابلی؛ پل معلق؛ پل راهاهن؛ دودکش؛ ...
JFT 3: اوپل آسترا؛ خانواده مزدا ;E100; Opel Signum; تویوتا کرولا ;...

جدول 2: کلاس های نمونه از خوشه های محاسبه شده توسط الگوریتم خوشه بندی ماتریس کوواریانس ما

در اطراف دو نوع دیگر، اما تنها در صورتی که تعداد هسته های بیشتری در دسترس باشد. انتظار برای چندین سال برای آموزش مجموعه‌ای از مدل‌ها گزینه‌ای نبود، بنابراین ما به یک راه بسیار سریع‌تر برای بهبود مدل پایه نیاز داشتیم.

5.2 مدل های تخصصی

وقتی تعداد کلاس‌ها بسیار زیاد است، منطقی است که مدل دست و پا گیر مجموعه‌ای باشد که شامل یک مدل عمومی آموزش‌دیده بر روی همه داده‌ها و بسیاری از مدل‌های «تخصصی» است، که هر کدام بر روی داده‌هایی آموزش داده شده‌اند که به شدت غنی شده‌اند. نمونه هایی از یک زیرمجموعه بسیار گیج کننده از کلاس ها (مانند انواع مختلف قارچ). سافت مکس این نوع متخصص را می توان با ترکیب تمام کلاس هایی که به آنها اهمیتی نمی دهد در یک کلاس زباله گرد بسیار کوچکتر کرد.

برای کاهش اضافه برازش و به اشتراک گذاشتن کار یادگیری آشکارسازهای ویژگی سطح پایین تر، هر مدل تخصصی با وزن های مدل عمومی مقداردهی اولیه می شود. سپس این وزن‌ها با آموزش متخصص با نیمی از نمونه‌های آن از زیرمجموعه ویژه و نیمی از نمونه‌های تصادفی از بقیه مجموعه آموزشی، کمی اصلاح می‌شوند. کلاس زباله‌دان به نسبت نسبتی که logit پس از آموزش، ما می‌توانیم مجموعه آموزش مغرضانه را با افزایش کلاس تخصصی بیش از حد نمونه برداری می‌شود، تصحیح کنیم.

5.3 اختصاص کلاس به متخصصان

به منظور استنتاج گروه بندی دسته بندی اشیا برای متخصصان، تصمیم گرفتیم بر روی دسته هایی تمرکز کنیم که شبکه کامل ما اغلب آنها را اشتباه می گیرد. حتی اگر می‌توانستیم ماتریس سردرگمی را محاسبه کرده و از آن به عنوان راهی برای یافتن چنین خوشه‌هایی استفاده کنیم، رویکرد ساده‌تری را انتخاب کردیم که برای ساخت خوشه‌ها به برچسب‌های واقعی نیازی ندارد.

به طور خاص، ما یک الگوریتم خوشه‌بندی را برای ماتریس کوواریانس پیش‌بینی‌های مدل عمومی خود اعمال می‌کنیم، به طوری که مجموعه‌ای از کلاس‌هااسدمتر که اغلب با هم پیش بینی می شوند، به عنوان هدف برای یکی از را برای ستون‌های K-means مدل های تخصصی ما استفاده می شوند،متر. ما یک نسخه آنلاین از الگوریتم ماتریس کوواریانس اعمال کردیم و خوشه‌های معقولی به دست آوردیم (نشان داده شده در جدول 2). ما چندین الگوریتم خوشه‌بندی را امتحان کردیم که نتایج مشابهی را ایجاد کرد.

5.4 انجام استنباط با گروه های متخصص

قبل از بررسی اینکه در هنگام تقطیر مدل‌های تخصصی چه اتفاقی می‌افتد، می‌خواستیم ببینیم که گروه‌های حاوی متخصصان چقدر خوب عمل می‌کنند. علاوه بر مدل‌های تخصصی، ما همیشه یک مدل عمومی داریم تا بتوانیم با کلاس‌هایی برخورد کنیم که متخصصی برای آن‌ها نداریم و تصمیم بگیریم از چه متخصصانی استفاده کنیم. با توجه به تصویر ورودیایکس، ما طبقه بندی اول را در دو مرحله انجام می دهیم:

مرحله 1: برای هر مورد آزمایشی، ما را پیدا می کنیم مجموعهnمحتمل ترین کلاس ها بر اساس مدل عمومی. این کلاس ها را صدا بزنیدک. در آزمایشات خود استفاده کردیم n = 1.

مرحله 2: سپس همه مدل های تخصصی را انتخاب می کنیم،متر، که زیر مجموعه ویژه ای از کلاس های گیج کننده،اسدمتر، دارای یک تقاطع غیر خالی باکو این را مجموعه فعال متخصصان بنامیمآی(توجه داشته باشید که این مجموعه ممکن است خالی باشد). سپس توزیع احتمال کامل را پیدا می کنیم qدر تمام کلاس هایی که حداقل می کند:

$$KL\ (\text{صدمتر}، q)(5)$$

$$KL\ (\text{ص}^g، q) + m \in A_\text{ک}$$

توزیع احتمال یک مدل تخصصی یا مدل کامل کلی را نشان می دهد.gوپمتر پ KL جایی کهدرنشان دهنده واگرایی توزیعپمتر توزیعی بر روی تمام طبقات تخصصی استمتربه علاوه یک کلاس سطل زباله، بنابراین هنگام محاسبه توزیع به تمام کلاس های موجودqتوزیع ما همه احتمالات کامل را جمع می کنیمqآن از کامل KL واگرایی اختصاص می دهدمترسطل زباله

6

دقت تست شرطی سیستم دقت تست
٪پایه 43.1٪ 25.0
%مدل تخصصی 45.9% 26.1 61 +

JFT. جدول 3: دقت طبقه بندی (1 بالا) در مجموعه توسعه

%تغییر دقت نسبی صحیح 0 350037 0 0.0 top1 تعداد متخصصان پوشش # نمونه های آزمایشی دلتا در
1 141993 + 1421 + 3.4٪
2 67161 + 1572 + 7.4٪
3 38801 + 1124 + 8.8٪
4 26298 + 835 + 10.5٪
5 16474 + 561 + 11.1٪
6 10682 + 362 + 11.3٪
7 7376 + 232 + 12.8٪
8 4703 + 182 + 13.6٪
9 4706 + 208 + 16.6٪
٪یا بیشتر 9082 + 324 + 14.1 10

پوشش JFT جدول 4: بهبود دقت برتر 1 بر اساس # مدل های تخصصی که کلاس صحیح را در مجموعه تست می دهند.

معادله 5 یک راه حل فرم بسته کلی ندارد، اگرچه زمانی که همه مدل ها یک احتمال واحد برای هر کلاس تولید می کنند، بسته به اینکه ما از آن استفاده کنیم، جواب یا میانگین حسابی یا هندسی است KL(q, p))یاKL(p, q)). ها از گرادیان نزول استفاده می logit و برای بهینه سازی (T = 1 باq = soft tmax(z)پارامتر می کنیم معادله 5. توجه داشته باشید که این بهینه سازی باید برای هر تصویر انجام شود .w.r.tکنیمبا.

نتایج 5.5

با شروع از شبکه کامل پایه آموزش دیده، متخصصان بسیار سریع تمرین می کنند (چند روز به جای چندین هفته همچنین تمامی متخصصان به صورت کاملا مستقل آموزش می بینند. جدول 3 دقت آزمون مطلق را .(JFT برای برای سیستم پایه و سیستم پایه همراه با مدل های تخصصی نشان می دهد. با 61 مدل تخصصی، 4.4 درصد بهبود نسبی در دقت آزمون به طور کلی وجود دارد. ما همچنین دقت آزمون مشروط را گزارش می‌کنیم، که دقتی است که فقط با در نظر گرفتن نمونه‌های متعلق به کلاس‌های تخصصی، و محدود کردن پیش‌بینی‌های خود به آن زیرمجموعه از کلاس‌ها انجام می‌شود.

مدل متخصص را آموزش دادیم که هر کدام دارای 300 کلاس بود 61 ،JFT برای آزمایش‌های تخصصی (به‌علاوه کلاس زباله‌دان). از آنجایی که مجموعه‌های کلاس‌های متخصصان از هم جدا نیستند، اغلب متخصصان متعددی داشتیم که یک کلاس تصویری خاص را پوشش می‌دادند. جدول 4 تعداد نمونه های مجموعه آزمایشی، تغییر در تعداد نمونه های صحیح در موقعیت 1 هنگام استفاده از متخصص(ها) و درصد بهبود نسبی در دقت نشان می دهد که بر اساس تعداد متخصصان تحت پوشش تفکیک شده است. JFT را برای مجموعه داده top1 کلاس. ما از این روند کلی تشویق می‌شویم که وقتی متخصصان بیشتری داریم که یک کلاس خاص را پوشش می‌دهند، بهبود دقت بیشتر می‌شود، زیرا آموزش مدل‌های متخصص مستقل بسیار آسان است.

هدف نرم به عنوان تنظیم کننده 6

یکی از ادعاهای اصلی ما در مورد استفاده از اهداف نرم به جای اهداف سخت این است که بسیاری از اطلاعات مفید را می توان در اهداف نرمی که احتمالاً نمی توان با یک هدف سخت رمزگذاری کرد حمل کرد. در این بخش ما نشان می‌دهیم که این یک اثر بسیار بزرگ با استفاده از داده‌های بسیار کمتر برای تناسب با پارامترهای 85 میلیونی مدل گفتار پایه که قبلاً توضیح داده شد، است. جدول 5 نشان می دهد که تنها با 3٪ از داده ها (حدود 20 میلیون نمونه)، آموزش مدل پایه با اهداف سخت منجر به اضافه کردن شدید می شود (ما توقف اولیه را انجام دادیم، زیرا دقت پس از رسیدن به 44.5٪ به شدت کاهش می یابد)، در حالی که همان مدل آموزش دیده است. با اهداف نرم قادر است تقریباً تمام اطلاعات را در مجموعه تمرینی کامل بازیابی کند (حدود 2٪ خجالتی). حتی قابل توجه تر است که توجه داشته باشیم که ما مجبور به توقف اولیه نبودیم: سیستم با اهداف نرم به سادگی به 57% "همگرا" شد. این نشان می دهد که اهداف نرم راه بسیار موثری برای برقراری ارتباط با نظم های کشف شده توسط یک مدل آموزش دیده بر روی تمام داده ها به مدل دیگری هستند.

7

سیستم و مجموعه آموزشی	Train Frame Accuracy	Test Frame Accuracy
پایه (100% مجموعه آموزشی)%	63.4%	58.9
پایه (3% مجموعه آموزشی)%	67.3%	44.5
اهداف نرم (3% از مجموعه آموزشی)%	65.4%	57.0

جدول 5: اهداف نرم به یک مدل جدید اجازه می دهد تا تنها از 3 درصد مجموعه آموزشی به خوبی تعمیم یابد. اهداف نرم با تمرین در مجموعه تمرینی کامل به دست می آیند.

استفاده از اهداف نرم برای جلوگیری از نصب بیش از حد متخصصان 6.1

استفاده کردیم، تمام کلاس‌های غیرتخصصی خود را JFT متخصصانی که در آزمایش‌های خود روی مجموعه داده کامل در تمام کلاس‌ها داشته باشند، Softmax در یک کلاس زباله جمع کردند. اگر به متخصصان اجازه دهیم ممکن است راه بسیار بهتری برای جلوگیری از تطبیق بیش از حد آنها نسبت به توقف اولیه وجود داشته باشد. یک متخصص بر روی داده هایی که در کلاس های خاص خود بسیار غنی شده است آموزش دیده است. این بدان معنی است که اندازه موثر مجموعه آموزشی آن بسیار کوچکتر است و تمایل زیادی به اضافه کردن بر روی کلاس های ویژه خود دارد. این مشکل را نمی توان با کوچکتر کردن متخصص بسیار حل کرد زیرا در این صورت اثرات انتقال بسیار مفیدی را که از مدل سازی همه کلاس های غیرتخصصی بدست می آوریم از دست می دهیم.

آزمایش ما با استفاده از 3 درصد از داده‌های گفتاری قویاً نشان می‌دهد که اگر یک متخصص با وزن‌های معمم اولیه اولیه شود، می‌توانیم با آموزش آن با اهداف نرم برای افراد غیر، تقریباً تمام دانش خود را در مورد کلاس‌های غیر ویژه حفظ کنیم. کلاس های ویژه علاوه بر آموزش آن با اهداف سخت. اهداف نرم را می توان توسط متخصص ارائه کرد. ما در حال حاضر در حال بررسی این رویکرد هستیم.

ارتباط با ترکیبی از خبرگان 7

استفاده از متخصصانی که بر روی زیرمجموعه هایی از داده ها آموزش دیده اند، شباهت هایی به ترکیبی از متخصصان [6] دارد که از یک شبکه دروازه برای محاسبه احتمال انتساب هر مثال به هر متخصص استفاده می کنند. همزمان با یادگیری برخورد با نمونه هایی که به آنها اختصاص داده شده است، شبکه دروازه در حال یادگیری انتخاب کارشناسانی است که هر نمونه را بر اساس عملکرد نسبی تبعیض آمیز کارشناسان برای آن مثال، به آنها اختصاص دهد. استفاده از عملکرد متمایز متخصصان برای تعیین تکالیف آموخته شده بسیار بهتر از خوشه‌بندی ساده بردارهای ورودی و اختصاص یک متخصص به هر خوشه است، اما موازی‌سازی آموزش را سخت می‌کند: اول، مجموعه آموزشی وزنی برای هر متخصص مدام در حال تغییر است. روشی که به همه کارشناسان دیگر بستگی دارد و دوم اینکه، شبکه گیت باید عملکرد کارشناسان مختلف را در یک مثال مقایسه کند تا بداند چگونه احتمالات تخصیص خود را اصلاح کند. این مشکلات به این معنی است که ترکیبی از متخصصان به ندرت در رژیمی استفاده می شود که ممکن است سودمندترین باشد: وظایفی با مجموعه داده های عظیم که شامل

.زیر مجموعه های متفاوتی هستند

موازی کردن آموزش چندین متخصص بسیار ساده تر است. ابتدا یک مدل کلی را آموزش می دهیم و سپس از ماتریس سردرگمی برای تعریف زیرمجموعه هایی استفاده می کنیم که متخصصان بر روی آنها آموزش دیده اند. هنگامی که این زیر مجموعه ها تعریف شدند، متخصصان می توانند کاملاً مستقل آموزش ببینند. در زمان آزمون می‌توانیم از پیش‌بینی‌های مدل عمومی استفاده کنیم تا تصمیم بگیریم کدام متخصصان مرتبط هستند و فقط این .متخصصان باید اجرا شوند

بحث 8

ما نشان داده‌ایم که تقطیر برای انتقال دانش از یک مجموعه یا از یک مدل بزرگ بسیار منظم به یک مدل تقطیر حتی زمانی که مجموعه انتقالی که برای آموزش مدل MNIST کوچک‌تر بسیار خوب عمل می‌کند. تقطیر در تقطیر شده استفاده می‌شود، فاقد هر گونه نمونه از یک یا چند کلاس باشد، بسیار خوب عمل می‌کند. برای یک مدل آکوستیک عمیق که نسخه‌ای از مدل مورد استفاده در جستجوی صوتی اندروید است، نشان داده‌ایم که تقریباً تمام پیشرفت‌هایی که با آموزش مجموعه‌ای از شبکه‌های عصبی عمیق به دست می‌آید را می‌توان در یک شبکه عصبی .با همان اندازه تقطیر کرد. استقرار بسیار آسان تر است

برای شبکه‌های عصبی واقعاً بزرگ، حتی آموزش یک مجموعه کامل غیرممکن است، اما ما نشان داده‌ایم که عملکرد یک شبکه واقعاً بزرگ که برای مدت بسیار طولانی آموزش داده شده است، می‌تواند به طرز چشمگیری با یادگیری تعداد زیادی از آنها بهبود یابد. شبکه های تخصصی، که هر کدام یاد می گیرند بین کلاس ها در یک خوشه بسیار گیج کننده تمایز قائل شوند. ما هنوز نشان نداده‌ایم که می‌توانیم دانش متخصصان را در یک شبکه .بزرگ تقطیر کنیم

قدردانی ها

و ایلیا سوتسکور و یورام سینگر برای ImageNet از یانگ کینگ جیا برای کمک به مدل‌های آموزشی در .بحث‌های مفید تشکر می‌کنیم

منابع

[1] C. Buciluˇa، R. Caruana، و A. Niculescu-Mizil. مدل فشرده سازی درمجموعه مقالات دوازدهمین کنفرانس بین المللی *ACM SIGKDD* کشف دانش و داده کاوی, KDD '06, pages 535–541, New York, NY, USA, 2006. ACM.

[2] J. Dean, G.S. Corrado, R. Monga, K. Chen, M. Devin, Q.V Le, M.Z., M. Ranzato, A. Senior, P. Tucker, K. Yang, and A.Y. شبکه های عمیق توزیع شده در مقیاس بزرگ. که در *NIPS*، 2012.

[3] تی جی دیتریش. روش های مجموعه ای در یادگیری ماشینی که درسیستم های طبقه بندی کننده چندگانه، صفحات 1-15. اسپرینگر، 2000.

[4] G. E. Hinton، L. Deng، D. Yu، G. E. Dahl، A. Mohamed، N. Jaitly، A. Senior، V. Vanhoucke، P. Nguyen، T. N Sainath و B. Kingsbury. شبکه‌های عصبی عمیق برای مدل‌سازی صوتی در تشخیص گفتار: دیدگاه‌های مشترک چهار گروه تحقیقاتیمجله پردازش سیگنال *IEEE*، 29 (6): 82-97، 2012.

[5] G. E. Hinton، N. Srivastava، A. Krizhevsky، I. Sutskever و R. R. Salakhutdinov. من شبکه های عصبی را با جلوگیری از انطباق همزمان آشکارسازهای ویژگی اثبات می کنم.پیش چاپ *arXiv* *arXiv:1207.0580*، 2012.

[6] R. A. Jacobs، M. I. Jordan، S. J. Nowlan و G. E. Hinton. ترکیبی تطبیقی از کارشناسان محلی.محاسبات عصبی، 3 (1): 79-87، 1991.

[7] A. Krizhevsky، I. Sutskever و G. E. Hinton. طبقه بندی شبکه تصویری با شبکه های عصبی

.کانولوشنال عمیق که درپیشرفت در سیستم های پردازش اطلاعات عصبی، صفحات 1097-1105، 2012

[8] J. Li، R. Zhao، J. Huang، و Y. Gong. آموزش dnn در اندازه کوچک با معیارهای مبتنی بر توزیع. صفحات 1910-1914، 2014، *Proceedings Interspeech 2014*که در.

[9] N. Srivastava، G.E. هینتون، آ. کریژفسکی، آی. سوتسکور، و آر. آر. سالاخوتدینوف. Dropout: یک راه ساده برای جلوگیری از برازش شبکه های عصبی.مجله تحقیقات یادگیری ماشین، 15 (1): 1929-1958، 2014.