

(i(a)

از بهينه‌سازي Adam برای به‌روزرسانی پارامترهای مدل استفاده می‌شود. این بهينه‌سازي از یک ترفند به نام "مومنتوم" استفاده می‌کند. این ترفند با نگاه داشتن یک میانگین متحرک از گرادینت‌های گذشته، به‌روزرسانی‌های مناسب تری را به دست می‌دهد. این به دلیل ایجاد واریانس کمتر در به‌روزرسانی‌ها به بهينه‌سازي کمک می‌کند تا به سرعت به حداقل تابع هزینه برسد. این ویژگی در کمک به فرآیند یادگیری و بهبود پایداری و سرعت فرآیند بهينه‌سازي مفید است. به طور کلی، ترفند مومنتوم در بهينه‌سازي Adam بهبود پایداری و سرعت فرآیند بهينه‌سازي را به دنبال دارد. این روش کمک میکند روند بهينه‌سازي سرعت مناسب پیدا کند یعنی در ابتدا که از نقطه ی بهينه دور هستیم با سرعت بیشتری به آن نزدیک می‌شویم و کم کم با نزدیک شدن به آن این روند کند تر و محتاطانه تر جلو می‌رود.

(ii(a)

Adam با نگاه داشتن یک میانگین متحرک از مقادیر مطلق گرادینت‌های گذشته، ایده نرخ یادگیری تطبیقی را با به کارگیری γ بهبود می‌بخشد. γ نشان دهنده میانگین متحرکی از مقادیر مطلق گرادینت‌های گذشته است. در Adam، به‌روزرسانی‌ها به وسیله تقسیم بر مجذور γ انجام می‌شود. به همین دلیل، پارامترهایی که دارای مقادیر مطلق بزرگتری از گرادینت هستند، به‌روزرسانی بیشتری دریافت می‌کنند. این ویژگی در کمک به فرآیند یادگیری و بهبود پایداری و سرعت فرآیند بهينه‌سازي مفید است. به طور کلی، ترفند نرخ یادگیری تطبیقی در بهينه‌سازي Adam بهبود پایداری و سرعت فرآیند بهينه‌سازي را به دنبال دارد.

اصطلاح m تقسیم بر γ نسبت سیگنال به نویز نامیده می‌شود. هنگامی که نسبت سیگنال به نویز بالا باشد، به‌روزرسانی‌ها بزرگتر خواهند بود. انتخاب مقادیر β_1 و β_2 برای کنترل این نسبت اهمیت دارد. وقتی این عبارت به مقدار صفر نزدیک می‌شود، مراحل به‌روزرسانی کندتر خواهد شد و اندازه مرحله به مقدار کوچکی نزدیک خواهد شد. نسبت سیگنال به نویز نزدیک به صفر نزدیک به بهينه می‌شود. یک ویژگی مهم از قاعده به‌روزرسانی Adam انتخاب دقیق اندازه مراحل است و باعث می‌شود که فرآیند یادگیری به مسیر درست هدایت شود و به سمت بهينه مربوطه حرکت کند.

(i(b)

$$E[\text{hdrop}] = E[\gamma d \odot h] = \gamma E[d \odot h] = \gamma E[d] \odot E[h]$$

در رابطه‌ی بالا، قابلیت انتخاب دلخواه وضعیت هر عضو مجموعه‌ی d به کار رفته است. برای این که این وضعیت‌ها به صورت تصادفی انتخاب شوند، می‌توان به هر عضو مجموعه‌ی d ، با احتمال pdrop ، مقدار 0 و به احتمال $1 - \text{pdrop}$ ، مقدار 1 اختصاص داد. بنابراین، $E[d]$ برابر با $1 - \text{pdrop}$ خواهد بود. با جایگذاری این مقدار در رابطه‌ی بالا، به دست می‌آید:

$$E[\text{hdrop}] = \gamma(1 - \text{pdrop}) \odot h$$

حال برای رسیدن به مقدار مورد انتظار باید داشته باشیم: $E[\text{hdrop}] = h$ پس $\gamma(1 - \text{pdrop})$ باید برابر با 1 باشد.

و خواهیم داشت: $\gamma = 1 / (1 - \text{pdrop})$.

(iii(b)

هدف اصلی در استفاده از دروازه خروجی، جلوگیری از بیش برآزش در شبکه‌های عصبی با کاهش هم‌آمیختگی واحدهای لایه مخفی است. با در نظر نگرفتن تصادفی واحدهای لایه پنهان در هنگام آموزش، شبکه مجبور به یادگیری ویژگی‌هایی است که مستقل از پیکربندی دقیق واحدهای دیگر هستند. این می‌تواند عملکرد تعمیم دهی مدل را بهبود بخشیده و بیش برآزش را کاهش دهد.

اما در هنگام ارزیابی، ما می‌خواهیم از تمام شبکه برای پیش‌بینی استفاده کنیم. اعمال dropout در هنگام ارزیابی باعث پیش‌بینی‌های نامنظم و ناقص می‌شود، زیرا خروجی مدل به واحدهای خاصی که در هنگام آموزش خاموش شده اند وابسته خواهد بود. در زمان ارزیابی ما می‌خواهیم مدل را به طور کامل و با در نظر گرفتن تمام وزن‌های بدست آمده بسنجیم نه بخشی از آن یا میانگینی از آن. در واقع اگر قرار است نودی در نظر گرفته نشود باید وزن آن در طول آموزش صفر شده باشد و اگر نشده یعنی وزنی که دارد خوب است پس استفاده از dropout نمی‌گذارد ما به خوبی مدل و وزن‌های بدست آمده را بسنجیم.

(ii)(f)(۲

Error Type: Propositional Phrase Attachment Error

Incorrect dependency: *named* → *Midland*

Correct dependency: *loan* → *Midland*

(iv)(f)(۲

Error Type: Modifier Attachment Error

Incorrect dependency: *elements* → *most*

Correct dependency: *crucial* → *most*

(g)(۲

تگ‌های بخش‌های گفتار (POS) برچسب‌های زبانی هستند که برای کلمات براساس متن‌شناختی و دستور زبانی آن‌ها اختصاص داده می‌شوند. استفاده از تگ‌های POS به عنوان ویژگی در پارسر، به دلیل فراهم کردن اطلاعات مفید درباره نقش کلمات در جمله، می‌تواند به تفکیک ساختار نحوی کمک کند. به عنوان مثال، در جمله "The cat sat on the mat," تگ‌های POS ممکن است به شکل زیر باشند: "The/DT cat/NN sat/VBD on/IN the/DT mat/NN". با استفاده از این تگ‌های POS، پارسر می‌تواند تشخیص دهد که "گره" و "فرش" اسم هستند، "نشست" گذشته فعل است، "روی" حرف اضافه است و "رفتن" فعل است. این اطلاعات سپس می‌تواند برای کمک به پارسر در تصمیم‌گیری‌های دقیق‌تر در پارسینگ استفاده شود. به عنوان مثال، پارسر ممکن است با استفاده از اینکه "on" حرف اضافه است، تشخیص دهد که "mat" یک عبارت حرف اضافه‌ای است که باید به "sat" متصل شود نه به cat.