تفسیرپذیری در یادگیری عمیق

سجاد سبزی
محمدرضا احمدی تشنیزی

۱۷ دسامبر ۲۰۲۳

مقدمه

مکانیسم‌های attention در یادگیری عمیق به عنوان یک جزء بنیادی در طراحی معماری شبکه‌های عصبی تبدیل شده‌اند. این امکان را به مدل‌ها می‌دهند که به طور انتخابی بر روی بخش‌های معینی از داده ورودی تمرکز کنند، فرآیند یادگیری را در وظایفی مانند تشخیص تصاویر، پردازش زبان طبیعی و سایر وظایف بهبود ببخشند. مفهوم "attention" در این زمینه از توجه شناختی انسان‌ها الهام گرفته است؛ به معنای اینکه در حین محیط اطراف به برخی جنبه‌ها بیشتر تمرکز می‌کنند و از دیگران توجه می‌شوند.

استفاده از attention در یادگیری عمیق

attention به طور گسترده در یادگیری عمیق استفاده می‌شود، به ویژه در وظایفی که داده‌های متوالی مانند پردازش زبان یا تحلیل داده‌های سری زمانی در آن وجود دارد. مکانیسم attention به معنای اصولی به مدل این امکان را می‌دهد که وزن‌های مختلفی را به بخش‌های مختلفی از داده ورودی اختصاص دهد و در یک لحظه خاص تاکید کند که چه چیزی را مهم می‌داند. این اغلب با استفاده از بردار متناسب، که جمع‌کننده‌ای از ویژگی‌های ورودی است، پیاده‌سازی می‌شود و وزن‌های آن توسط ارتباط هر ویژگی با سیاق کنونی تعیین می‌شود.

یک مورد کاربرد نمونه در مدل‌های دنباله به دنباله است، که attention به مدل کمک می‌کند تا در تولید هر کلمه از دنباله خروجی بر روی بخش‌های خاصی از دنباله ورودی تمرکز کند. این به ویژه در وظایفی مانند ترجمه ماشینی مفید است، جایی که بخش‌های مختلفی از جمله ورودی ممکن است در مراحل مختلف تولید جمله خروجی مهم‌تر باشند.

انواع مکانیسم‌های attention

مکانیسم‌های attention می‌توانند بر اساس نحوه attention به بردارهای ویژگی و پرسش‌های مدل دسته‌بندی شوند. به طور عام، آن‌ها را می‌توان به سه گروه اصلی تقسیم کرد:

مکانیسم‌های attention مرتبط به ویژگی‌ها

این مکانیسم‌ها بر اساس ویژگی‌های داده ورودی تأسیس شده‌اند. آن‌ها می‌توانند بر اساس تعداد ویژگی‌ها، سطوح ویژگی‌ها یا نمایندگی ویژگی‌ها بیشتر تقسیم شوند.

مکانیسم‌های attention مرتبط به پرسش‌ها

این مکانیسم‌ها بر اساس انواع پرسش‌هایی که توسط مدل انجام می‌شود طراحی شده‌اند. به عنوان مثال، برخی از مدل‌ها ممکن است بر اساس نوع وظیفه‌ای که انجام می‌دهند، مکانیسم‌های attention مختلفی استفاده کنند، مثل تشخیص تصاویر در مقایسه با ترجمه متن.

مکانیسم‌های attention عمومی

این مکانیسم‌ها به یک مدل خاص و یا مدل پرسش عمومی نیستند و بیشتر برای انواع مختلف داده و وظایف قابل استفاده هستند.

یک مثال از مکانیسم attention خاص، مکانیسم هم‌توجه (co-attention) است که در مواقعی استفاده می‌شود که مدل نیاز به attention به چندین نوع ورودی به طور همزمان دارد، مثل تصویر و سوال در وظایف پاسخ به سوال‌های تصویری.

چالش‌ها و محدودیت‌ها

هر چند مکانیسم‌های attention قابلیت‌های مدل‌های یادگیری عمیق را به شدت افزایش داده‌اند، اما چالش‌ها و محدودیت‌هایی هم دارند. یکی از چالش‌های اصلی، پیچیدگی محاسباتی است که مکانیسم‌های attention به مدل اضافه می‌کنند. این می‌تواند فرآیند آموزش و پیش‌بینی را کند کند، به ویژه برای مجموعه‌داده‌های بزرگ یا مدل‌های پیچیده.

چالش دیگر، قابل فهمی این مدل‌ها است. هر چند وزن‌های attention برخی از نظراتی را در مورد اینکه مدل به چه چیزی تمرکز می‌کند، ارائه می‌دهند، اما درک دقیق دلایل دقیق این وزن‌ها ممکن است مشکل باشد. این عدم شفافیت می‌تواند مشکل مهمی در زمینه‌هایی باشد که توجیه قابلیت‌ها مهم است، مانند حوزه بهداشت.

علاوه بر این، مکانیسم‌های attention گاهی ممکن است منجر به مدل‌هایی شوند که در آموزش از داده‌های آموزشی اختلال ایجاد می‌کنند، به ویژه در مواردی که داده‌های آموزشی به میزان کافی گوناگون نیستند یا پیچیدگی کامل وظیفه را نمایش نمی‌دهند.

نتیجه‌گیری

در نتیجه، مکانیسم‌های attention نقش حیاتی در یادگیری عمیق مدرن ایفا می‌کنند و به مدل‌ها امکان می‌دهند تا بر روی بخش‌های مهم‌تر داده ورودی برای یک وظیفه خاص تمرکز کنند. هر چند با چالش‌هایی مانند افزایش پیچیدگی محاسباتی و احتمال ایجاد انطباق از داده‌های آموزشی همراه هستند، اما مزایای آن‌ها در بهبود عملکرد و قابلیت‌های مدل بسیار مهم هستند. در حالی که یادگیری عمیق به تدریج تکامل می‌کند، توسعه و بهبود مکانیسم‌های attention احتمالاً یکی از حوزه‌های کلیدی تحقیق و نوآوری خواهد ماند.

References

[۱] Analytics Vidhya - "Attention Mechanism In Deep Learning". Link

[۲] ar۵iv.org - "A General Survey on Attention Mechanisms in Deep Learning". https://ar5iv.org/pdf/2203.14263