تفسیرپذیری در یادگیری عمیق

سجاد سبزی
محمدرضا احمدی تشنیزی

۱۷ دسامبر ۲۰۲۳

مقدمه

استفاده از تفسیرپذیری در یادگیری عمیق به یک نقطه مهم در حوزه هوش مصنوعی تبدیل شده است. این نیاز به درک و اعتماد به فرآیندهای تصمیم‌گیری پیچیده مدل‌های یادگیری عمیق را دنبال می‌کند. تفسیرپذیری ارجاع به میزان درک یک انسان از علت یک تصمیم انجام شده توسط یک مدل یادگیری ماشین را درک کند.

استفاده از تفسیرپذیری در یادگیری عمیق

مدل‌های یادگیری عمیق، که به دلیل پیش‌بینی‌های قدرتمند خود شناخته می‌شوند، اغلب به دلیل پیچیدگی آنها به عنوان "جعبه‌های سیاه" دیده می‌شوند. برای کاهش این مسئله، روش‌های تفسیرپذیری برای افزایش قابلیت درک و اعتمادپذیری این مدل‌ها استفاده می‌شوند. دو رویکرد معروف در این زمینه عبارتند از توضیحات مدل محلی قابل تفسیر و بی‌تبعی LIME و Shapley-Values

توضیحات مدل قابل تفسیر محلی(LIME)

- LIME یک روش برای تفسیر مدل‌های جعبه‌های سیاه است که با تقریب آنها به صورت محلی با مدل‌های قابل تفسیر انجام می‌شود. این روش بر روی وفاداری محلی تمرکز دارد و اطمینان حاصل می‌کند که توضیح مطابق با نحوه عملکرد مدل در مجاورت مثالی که پیش‌بینی می‌شود، باشد. LIME به ویژه برای وظایفی مانند طبقه‌بندی متن و تصویر که در درک اینکه کدام ویژگی‌های داده به پیش‌بینی‌های خاصی منجر می‌شوند، مفید است.

Shapley-Values

- Shapley-Values برای ارزیابی نقش هر ویژگی در پیش‌بینی یک مدل استفاده می‌شوند. این روش بی‌تبعی است و معمولاً برای اندازه‌گیری اهمیت ویژگی‌ها در یک مدل آموزش دیده استفاده می‌شود. با این حال، Shapley-Values به منابع محاسباتی قابل توجهی نیاز دارند و کارایی آنها در توضیح تعاملات پیچیده در شبکه‌های عصبی عمیق محدود است.

چالش‌ها در تفسیرپذیری

مدل‌های یادگیری عمیق پیچیده هستند و دارای لایه‌های چندگانه و تعاملات غیرخطی هستند که آنها را به صورت ذاتی دشوار به تفسیر می‌کنند. این پیچیدگی چالش مهمی را در رسیدن به تفسیرپذیری کامل ایجاد می‌کند. به علاوه، تعادل بین وفاداری (دقت تفسیر نسبت به عملکرد مدل) و سادگی (آسانی درک تفسیر) وجود دارد. روش‌هایی مانند LIME و Shapley-Values، در عمل، ممکن است همیشه توضیح کامل یا وفادار به صورت جهانی از رفتار مدل را فراهم نکنند.

علاوه بر این، اطمینان از قابلیت اعتماد به الگوریتم‌های تفسیر امر حیاتی است. این الگوریتم‌ها باید به درستی علت پشت تصمیمات یک مدل را آشکار کنند. اعتمادپذیری به ویژه زمانی مهم است که تفسیر توسط یک الگوریتم خارجی ارائه می‌شود که ممکن است جزء مدل مورد تفسیر نباشد.

نتیجه‌گیری

تفسیرپذیری در یادگیری عمیق برای درک، اعتماد و استفاده مؤثر از سیستم‌های هوش مصنوعی، به ویژه در برنامه‌های حیاتی، بسیار حیاتی است. روش‌هایی مانند LIME و Shapley-Values در این راستا گام‌های مهمی برداشته‌اند. با این حال، پیچیدگی ذاتی مدل‌های یادگیری عمیق و چالش تعادل بین وفاداری و تفسیرپذیری به این معناست که این یک زمینه تحقیقاتی در حال ادامه است. پیشرفت‌های آینده در این زمینه برای توسعه سیستم‌های هوش مصنوعی ضروری هستند که نه تنها قدرتمند بلکه شفاف و قابل اعتماد هم باشند.

References

- For a detailed exploration of LIME and its application in interpretability، see "Why Should I Trust You؟" Explaining the Predictions of Any Classifier)https://ar5iv.labs.arxiv.org/html/1602.04938(

- For a comprehensive understanding of interpretability and trustworthiness in deep learning، refer to Interpretable Deep Learning: Interpretation، Interpretability، Trustworthiness، and Beyond)https://ar5iv.labs.arxiv.org/html/2103.10689(

- Additional insights into interpretability challenges and methodologies can be found in DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems)https://ar5iv.labs.arxiv.org/html/1906.10263(

- For further exploration of interpretability methods in Natural Language Processing، the article Model Explainability in Deep Learning Based Natural Language Processing provides valuable information)https://ar5iv.labs.arxiv.org/html/2106.07410(