روش های بصری سازی Feature Map ها در CNN

1) Activation Maximization : یکی از اولین تکنیک هایی است که برای درک آنچه توسط یک فیلتر خاص در CNN فعال می‌شود، استفاده می‌شود. این روش با تولید یک تصویر که فعالیت یک واحد خاص را به حداکثر می‌رساند شروع می‌شود. این روش اغلب با استفاده از روش های بهینه سازی انجام می‌شود که تصویر ورودی را تغییر می‌دهند تا پاسخ های فیلتر را بیشینه کنند.

2) Feature Maps Visulization: این تکنیک در واقع نمایش مستقیم نقشه های ویژگی است که در آن ها می‌توان دید که هر فیلتر چگونه به بخش های مختلف تصویر واکنش نشان می‌دهد. نقشه های ویژگی به ما نشان می‌دهند که کدام ویژگی های تصویری در هر لایه شبکه فعال می‌شوند و به ما کمک می‌کنند که بفهمیم شبکه چگونه تصمیم گیری می‌کند.

3) Backpropagation & Deconvolution

این روش با بازسازی تصاویر از ویژگی های میانی شبکه، تصمیم گیری های آن را قابل فهم و تفسیر می‌کند. در این روش فعالسازی های حاصل از یک لایه مشخص در شبکه های CNN مورد استفاده قرار می‌گیرند و سپس فرآیندی معکوس انجام می‌شود تا بتوان ورودی هایی که به این فعالسازها منجر شده اند را بازسازی کرد. با این روش می‌توان دید که کدام بخش های تصویر در تصمیم گیری های مدل تأثیرگذار بوده اند.

مراحل اجرای Deconvolution :

ابتدا یک لایه از مدل CNN انتخاب می‌شود که می‌خواهیم فعال سازهای آن را بررسی کنیم. (انتخاب لایه)

در مرحله بعدی یک تصویر ورودی از شبکه عبور داده می‌شود تا فعالسازی های لایه مورد نظر بدست آیند. (Forward Pass)

مرحله بعدی معکوس کردن فعالسازی هاست. با استفاده از روش هایی مثل Transposed Conv. فعالسازی ها به سمت ورودی شبکه معکوس می‌شوند، به این معنی که سعی می‌شود تصویر ورودی اصلی که به این فعالسازی ها منجر شده است بازسازی بشود.

مرحله نهایی مرحله تحلیل و بررسی تصاویر بازسازی شده است تا متوجه شویم کدام ویژگی های تصویری در تصمیم گیری های شبکه تأثیر داشته اند.

4) Guided Backpropagation :

این متد یک روش اصلاح شده از شکل معمولی روش Back propagation است که تنها واکنش های مثبت از فعالسازی را در نظر میگیرد. این روش برای شناسایی دقیق تر قسمت های تصویر که بیش ترین تاثیر را بر تصمیمات شبکه دارند، استفاده میشود.

5) CAM و Grad-CAM

CAM (Class Activation Maps) به معنای نقشه های فعالسازی کلاس است که امکان میدهد این را میدهند که کدام نواحی از تصویر به طور خاص در تصمیم گیری های یک شبکه عصبی پیچشی (CNN) دخالت دارند. این تکنیک برای مدل هایی که شامل لایه های پیچشی و سپس لایه های گلوبال - اوریج پولینگ (GAP) میباشند، به خوبی کار میکند. CAM می تواند نقشه های حرارتی (Heat Map) تولید کند که نشان دهنده میزان تاثیر هر بخش از تصویر بر تصمیم گیری های کلاس خاص است.

Grad-CAM یا Gradient-weighted - CAM یک نسخه پیشرفته تر و عمومی تر از CAM است که با استفاده از گرادیان های محاسبه شده نسبت به نقشه های ویژگی هر لایه پیچشی، امکان تولید نقشه های حرارتی را میدهد که نشان میدهد کدام قسمت های تصویر بیشترین تاثیر را برای تشخیص یک کلاس خاص را دارند.

6) Integrated Gradients

این روش یکی از روش های محبوب برای تفسیر مدل های یادگیری عمیق است که پژوهه برای تشخیص اهمیت هر ویژگی ورودی (مثل پیکسل های تصویر) در تصمیم گیری نهایی مدل استفاده میشود. این تکنیک با محاسبه گرادیان خروجی نسبت به ورودی و ادغام این گرادیان بر روی یک مسیر از یک ورودی پایه (مثلاً تصویری که همه پیکسل های آن صفر هستند) به تصویر واقعی انجام میشود.

7) نقشه های حساسیت (Saliency Maps)

نقشه های حساسیت تکنیک هایی هستند که برای بصری سازی و تحلیل شبکه های عصبی عمیق بکار میروند این نقشه ها نشان میدهند که کدام نواحی از یک تصویر ورودی برای تصمیم گیری یک شبکه عصبی مهم تر هستند و بعبارت دیگر یک شبکه هنگام طبقه بندی یک تصویر به چه قسمت هایی از تصویر توجه بیشتری دارد.

LIME (Local Interpretable Model-agnostic Explanations) (8

این روش یک تکنیک تفسیرپذیری مدل های یادگیری ماشین است که به طور مستقل از نوع مدل، توضیحات محلی (Local) برای تصمیمات مدل فراهم می‌کند. این روش با ایجاد نمونه‌های محلی از داده‌ها و آموزش یک مدل ساده روی این نمونه‌ها، ویژگی‌هایی که بیشترین تاثیر را بر تصمیم مدل اصلی دارند، شناسایی می‌کند.