

پیشنهاد پژوهشی:

استفاده از بردارهای تخمین حرکت برای کاهش پیچیدگی شبکه های عصبی
کانولوشنی در پردازش ویدئو

**Using motion estimation to reduce the complexity of convolutional
neural networks in video processing**

ارائه شده به: پژوهشکده علوم کامپیوتر، پژوهشگاه دانشهای بنیادی

توسط:

مهدی مدرسی

دانشکده مهندسی برق و کامپیوتر، دانشکده فنی دانشگاه تهران، تهران

modarressi@ut.ac.ir

زمستان ۱۴۰۳

چکیده. با گسترش حجم داده‌ها و ابعاد شبکه‌های عصبی کانولوشنی، تعداد پارامترها و حجم محاسبات به طور چشمگیری افزایش می‌یابد. این افزایش، نیاز به مصرف انرژی و پهنای باند حافظه بسیار زیاد را به دنبال دارد و به چالشی بزرگ برای زیرساخت‌های سخت‌افزاری تبدیل می‌شود. همچنین، اجرای این شبکه‌ها بر روی پردازنده‌های عمومی با سربرار نرم افزاری قابل توجهی همراه است که منجر به افزایش زمان مورد نیاز برای پردازش می‌شود. بنابراین استفاده مجدد از محاسبات در این گونه از شبکه‌ها مورد توجه محققان قرار گرفته است. یکی از انواع استفاده‌های مجدد از محاسبات، استفاده مجدد مبتنی بر ورودی است که با توجه به شباهت ورودی‌ها، محاسبات را مجدداً استفاده می‌کنند. این پیشنهاد پژوهشی بر اساس این مشاهده شکل گرفته است که می‌توان با توجه به خاصیت‌های ورودی‌های ویدئویی، به طور خاص تخمین حرکت (**motion estimation**) و استخراج بردارهای حرکت، قسمت‌های مشابه در دو قاب پشت سر هم را بدون سربرار محاسبات تشخیص داد و محاسبات مربوط به آنها را مورد استفاده مجدد قرار داد. با توجه به درصد بالای مشتبه‌ت در داده‌های ویدئویی، این امر باعث کاهش چشمگیر محاسبات و دسترسی‌ها به حافظه برون تراشه و همچنین توان مصرفی خواهد شد. همچنین با انجام تقریب‌هایی در اینگونه شبکه‌ها، می‌توان درصد استفاده مجدد از محاسبات را با کاهش جزئی دقت افزایش داد.

توصیف طرح

شبکه‌های عصبی نقش مهمی در پیشبرد علوم مختلف ایفا می‌کنند و یکی از زیرمجموعه‌های مهم آنها یادگیری عمیق است که از طریق شبکه‌های عصبی عمیق به حل مسائل پیچیده می‌پردازد. این شبکه‌ها در زمینه‌های مختلفی مانند روباتیک، کنترل، پردازش سیگنال، پردازش تصویر و تشخیص الگو به کار می‌روند. اهمیت روزافزون شبکه‌های عصبی عمیق ناشی از نیاز به عملکرد هوشمند در طیف گسترده‌ای از سیستم‌های کامپیوتری، از حسگرها و سیستم‌های نهفته کوچک تا سرورهای بزرگ داده است. طبق پیش‌بینی‌ها و آمارهای موجود، تا سال ۲۰۲۴ تعداد دستگاه‌های متصل به اینترنت اشیاء جهانی به ۸۳ میلیارد خواهد رسید که نشان‌دهنده گسترش قابل توجهی در استقرار اینترنت اشیاء در بخش‌های مختلف است و انتظار می‌رود که میزان هزینه‌ها در این حوزه تا سال ۲۰۲۸ به ۹۹۱ میلیارد دلار برسد. به‌طور کلی، پردازش صوت و تصویر، که مبنای بسیاری از خدمات هوشمند و نوین در دستگاه‌های قابل حمل است، به بهترین شکل از طریق شبکه‌های عصبی پیاده‌سازی می‌شود. موفقیت این شبکه‌ها در بینایی ماشین، پیشرفت‌های قابل توجهی را در حوزه‌های تشخیص چهره، خودرو خودران، تشخیص‌های پزشکی پیشرفته و ... امکان‌پذیر کرده است که قبلاً غیرممکن تصور می‌شد. یکی از زیرمجموعه‌های عظیم و پایه‌ای شبکه‌های عصبی، به شبکه‌های عصبی کانولوشنی معروف هستند.

Convolutional Neural Networks (CNN) یکی از مهم‌ترین معماری‌های مورد استفاده در یادگیری عمیق هستند. ساختار کلی یک CNN شامل چندین لایه متفاوت است که هر کدام وظیفه خاصی در استخراج و پردازش ویژگی‌ها از تصاویر ورودی دارند. معمولاً یک CNN از چند لایه کانولوشنی آغاز می‌شود که وظیفه اصلی آن‌ها اعمال فیلترها به تصویر ورودی برای استخراج ویژگی‌های ابتدایی مانند لبه‌ها، بافت‌ها و الگوهای محلی است. این لایه‌ها اغلب با لایه‌های پولینگ (**pooling**) همراه هستند که ابعاد خروجی‌ها را کاهش می‌دهند و باعث می‌شوند شبکه مقاوم‌تر نسبت به جابجایی‌های کوچک در تصویر باشد. پس از این مراحل، لایه‌های کانولوشنی و پولینگ متعدد ممکن است برای استخراج ویژگی‌های پیچیده‌تر در سطوح مختلف اضافه شوند. در نهایت، داده‌ها به لایه‌های کاملاً متصل (**fully connected**) ارسال می‌شوند که وظیفه آن‌ها ترکیب و طبقه‌بندی ویژگی‌های استخراج‌شده است.

کاربردهای شبکه‌های عصبی کانولوشنی

CNN‌ها به خاطر قابلیت‌های برجسته‌شان در پردازش داده‌های تصویری و دیگر انواع داده‌های ساختاریافته، کاربردهای متنوعی دارند. برخی از وظایف اصلی که CNN‌ها برای آن‌ها طراحی شده‌اند، شامل موارد زیر است:

*تشخیص اشیاء: تشخیص اشیاء به معنای شناسایی و تعیین موقعیت اشیاء مختلف در یک تصویر است. این وظیفه معمولاً با استفاده از چارچوب ۱۸هایی مانند (YOLO) Once Look Only You انجام می‌شود.

*تشخیص الگو: تشخیص الگو شامل دسته‌بندی و تحلیل الگوها و ویژگی‌های موجود در تصاویر است. این وظیفه به طور عمده در تشخیص چهره، تشخیص دست‌خط و سایر کاربردها استفاده می‌شود.

*بخش‌بندی تصویر: بخش‌بندی تصویر به معنای تقسیم یک تصویر به نواحی مختلف بر اساس ویژگی‌های تصویری است. این وظیفه برای ایجاد نقشه‌های دقیق از ویژگی‌های مختلف تصویر مانند سلول‌های پزشکی یا مناطق مختلف در تصاویر ماهواره‌ای استفاده می‌شود.

*کلاس‌بندی تصویر: کلاس‌بندی تصویر به شناسایی و تخصیص یک برچسب یا کلاس خاص به هر تصویر پرداخته و می‌تواند شامل تشخیص اشیاء، تشخیص چهره، یا هر نوع دیگر از طبقه‌بندی تصاویر باشد. این وظیفه به طور عمده با استفاده از معماری‌های مختلف شبکه‌های عصبی کانولوشنی مانند LeNet، MobileNet انجام می‌شود.

*تبدیل تصویر به تصویر: تبدیل تصویر به تصویر شامل تبدیل یک نوع تصویر به نوع دیگر است. برای مثال، تبدیل تصویر رنگی به تصویر سیاه‌وسفید یا برعکس، یا حتی تبدیل تصاویر با کیفیت پایین به کیفیت بالا.

*پیش‌بینی و تولید تصویر: پیش‌بینی و تولید تصویر شامل ایجاد تصاویر جدید بر اساس الگوهای یادگرفته شده است. این وظیفه با استفاده از شبکه‌های (GAN) Networks Adversarial Generative انجام می‌شود که به تولید تصاویر واقع‌گرایانه کمک می‌کند.

*استخراج ویژگی: استخراج ویژگی شامل استخراج ویژگی‌های مهم از داده‌های تصویری است که برای وظایف دیگر مانند دسته‌بندی یا تشخیص اشیاء استفاده می‌شود.

*تشخیص فعالیت: تشخیص فعالیت به شناسایی و تحلیل فعالیت‌های مختلف در یک ویدئو یا توالی تصاویر اشاره دارد. این وظیفه برای تحلیل ویدئو و شناسایی فعالیت‌های خاص در صحنه‌های مختلف کاربرد دارد.

ویژگی‌های شبکه‌های عصبی کانولوشنی

پیچیدگی محاسباتی بسیار زیاد شبکه‌های عصبی کانولوشنی، مستلزم صرف توان مصرفی، پهنای باند حافظه و قدرت پردازشی فراوانی است و از این رو چالشی جدی برای زیرساخت‌های سخت‌افزاری محسوب می‌شود. CNN ها به دلیل نیاز به انجام عملیات‌های پیچیده و تعداد بالای پارامترها، به طور معمول به منابع محاسباتی زیادی نیاز دارند. این نیاز به منابع زیاد به ویژه در

کاربردهای پردازش تصویر و ویدئو که نیازمند انجام محاسبات بر روی داده‌های حجیم و با وضوح بالا هستند، بیشتر به چشم می‌آید.

با این وجود، به دلیل ماهیت تقریبی و محاسبات اضافی این شبکه‌ها، بهینه‌سازی سخت‌افزاری به یکی از اولویت‌های اصلی تبدیل شده است. یکی از ویژگی‌های کلیدی شبکه‌های عصبی کانولوشنی که عنوان شد، ماهیت تقریبی محاسبات آن‌هاست. به عبارت دیگر، عملیات ریاضی لازم برای شبکه‌های عصبی می‌توانند با دقتی معقول و در بسیاری از مواقع با خطاهای قابل قبول در انجام محاسبات انجام شوند، حال آنکه خروجی نهایی مدل دقت بسیار کم و ناچیزی را از دست می‌دهد. این ویژگی باعث می‌شود که در بسیاری از کاربردها بتوان از دقت محاسباتی کمی استفاده کرد و در عین حال عملکرد کلی سیستم حفظ شود.

در نهایت، با توجه به رشد روزافزون نیازها و کاربردهای شبکه‌های عصبی کانولوشنی، اهمیت بهینه‌سازی سخت‌افزاری این شبکه‌ها روز به روز بیشتر می‌شود. بهبود تکنیک‌های بهینه‌سازی و استفاده از فناوری‌های جدید می‌تواند به کاهش چالش‌های مربوط به منابع سخت‌افزاری و بهبود عملکرد کلی سیستم‌های مبتنی بر شبکه‌های عصبی کمک کند.

اهمیت کاهش محاسبات در شبکه‌های عصبی کانولوشنی

شبکه‌های عصبی کانولوشنی به دلیل توانایی‌شان در یادگیری و استخراج الگوهای پیچیده از داده‌ها، انقلابی در حوزه بینایی کامپیوتری و بسیاری دیگر از حوزه‌ها ایجاد کرده‌اند. با این حال، یکی از چالش‌های مهم مرتبط با شبکه‌های عصبی کانولوشنی، نیاز عظیم به محاسبات است که عمدتاً به دلیل تعداد زیاد عملیات مورد نیاز برای پیچش‌ها و فضای پارامتری بزرگ است. این بار محاسباتی نه تنها زمان و انرژی مورد نیاز برای آموزش و استنتاج را افزایش می‌دهد، بلکه استقرار CNN‌ها بر روی دستگاه‌های با منابع محدود مانند تلفن‌های همراه و دستگاه‌های اینترنت اشیا را نیز محدود می‌کند. به همین دلیل، پژوهش‌های زیادی بر روی توسعه روش‌هایی برای کاهش پیچیدگی محاسباتی CNN‌ها بدون کاهش عملکرد آن‌ها متمرکز شده است. این روش‌ها شامل هرس مدلی، کمی‌سازی استفاده مجدد از محاسبات، طراحی معماری‌های کارآمد و استفاده از شتابدهنده‌های سخت‌افزاری تخصصی می‌باشند. پرداختن به ناکارآمدی‌های محاسباتی در CNN‌ها برای پیشرفت کاربردهای عملی آن‌ها و بهبود کارایی‌شان در سناریوهای واقعی بسیار حائز اهمیت است. با توجه به اینکه CNN‌ها به بخش جدایی‌ناپذیر برنامه‌های بی‌درنگ تبدیل شده‌اند، نیاز به تسریع عملکرد آن‌ها به ویژه برای سیستم‌های نهفته در دستگاه‌های موبایل و لبه شبکه افزایش یافته است. از این رو کاهش مؤثر محاسبات برای دستیابی به عملکرد بی‌درنگ و تضمین امکان اجرای برنامه‌ها بر روی سیستم‌های نهفته حیاتی است. با کاهش چشمگیر بار محاسباتی، پردازش داده‌ها سریع‌تر انجام می‌شود و به این ترتیب، نیازهایی نظیر تأخیر کم که برای برنامه‌های بی‌درنگ ضروری است، برآورده می‌شود.

ویژگی‌های ویدئو به عنوان ورودی شبکه عصبی کانولوشنی

استفاده از ویدئو به عنوان ورودی شبکه عصبی کانولوشنی دارای چندین ویژگی منحصر به فرد است که می‌تواند عملکرد شبکه را بهبود بخشد. این ویژگی‌ها عبارتند از:

• اطلاعات زمانی: برخلاف تصاویر ثابت، ویدئو تداوم زمانی را فراهم می‌کند. این ویژگی امکانی را به CNN می‌دهد تا الگوها و وابستگی‌ها را در طول زمان یاد بگیرد که برای وظایفی مانند تشخیص عمل، خلاصه‌سازی ویدئو و ردیابی اشیاء بسیار مفید است.

• اطلاعات حرکتی: ویدئوها حرکت اشیاء را ضبط می‌کنند که برای درک صحنه‌های پویا بسیار مهم است. شبکه‌های عصبی کانولوشنی می‌توانند از جریان نوری و ویژگی‌های مبتنی بر حرکت مانند تخمین حرکت برای بهبود توانایی خود در تشخیص و ردیابی اشیاء در طول زمان و همچنین کاهش محاسبات استفاده کنند.

• اطلاعات زمینه‌ای: ویدئوها اغلب زمینه غنی‌تری نسبت به تصاویر تکی ارائه می‌دهند، زیرا نشان می‌دهند که صحنه‌ها چگونه تکامل می‌یابند. این زمینه می‌تواند به CNN ها کمک کند تا با فهم تنظیمات گسترده‌تری که در آن اشیاء یا اعمال رخ می‌دهند، پیش‌بینی‌های دقیق‌تری انجام دهند.

• افزایش داده‌ها: ویدئوها فرصت‌های بیشتری برای افزایش داده‌ها نسبت به تصاویر فراهم می‌کنند. با انتخاب فریم‌های مختلف و ترکیب فریم‌ها یا تغییر سرعت پخش، نمونه‌های آموزشی متنوع‌تری می‌توان تولید کرد که باعث بهبود مقاومت و تعمیم‌پذیری CNN می‌شود.

بنابراین از ویژگی‌های حرکتی ویدئوها می‌توان در کاربردهای بسیاری بهره برد. در قاب‌های متوالی، شباهت‌های بسیاری یافت می‌شود که همین شباهت‌ها مقدمه‌ای برای استفاده هر چه زیاده‌تر از شبکه‌ها برای استنتاج از این نوع ورودی خاص می‌شود.

راه حل پیشنهادی

استفاده مجدد از محاسبات یکی از راهکارهای مناسب برای کاهش پیچیدگی و تعداد محاسبات CNN ها و در نهایت ذخیره پهنای باند مورد استفاده شبکه‌ها است. استفاده مجدد از محاسبات در شبکه‌ها به دو دسته اصلی (و چندین دسته فرعی) مبتنی بر وزن‌ها و مبتنی بر ورودی‌ها تقسیم می‌شود. استفاده مجدد مبتنی بر وزن‌ها به معنی استفاده از تکرار وزن‌ها برای کاهش محاسبات و استفاده مجدد مبتنی بر ورودی‌ها به معنی استفاده از شباهت و تکرار ورودی‌ها به منظور کاهش محاسبات در شبکه‌های عصبی است. استفاده مجدد مبتنی بر شباهت ورودی‌ها بسیار مؤثر بوده و این نوع استفاده مجدد در ورودی‌هایی که از نوع ویدئو هستند به دلیل شباهت بسیار زیاد قاب‌های متوالی به‌خاطر پیوستگی آنها، بسیار دیده می‌شود. همین امر باعث ایجاد محاسبات تکراری و سربارهای محاسباتی بسیاری می‌شود. با توجه به ماهیت تقریبی محاسبات در شبکه‌های عصبی کانولوشنی، امکان استفاده بهینه از بسیاری از محاسبات ورودی‌های پیشین وجود دارد. این به آن معنا است که با استفاده از تکنیک‌های بهینه‌سازی، می‌توان بخشی از محاسبات

را مجدداً مورد استفاده قرار داد یا از محاسبات ساده‌تر برای دستیابی به نتایج مشابه استفاده کرد. به این ترتیب، مصرف منابع سخت‌افزاری کاهش یافته و توان مصرفی و پهنای باند حافظه ذخیره می‌شود.

اکثر پژوهش‌ها نتایج ورودی قبل را در یک حافظه ذخیره می‌کنند و در صورت تساوی ورودی‌ها، محاسبات جاری حذف شده و نتایج از حافظه خوانده می‌شوند. پژوهش‌ها و مقالات این حیطه یا از سربار زمانی مقایسه که بسیار زیاد است و بخش زیادی از سود ناشی از استفاده مجدد را از بین می‌رود رنج می‌برند و صرفاً فقط محاسبات مربوط به پیکسل‌های متناظر در دو قاب پشت سرهم با شرط مساوی بودن را استفاده مجدد می‌کنند و یا با استنتاج از روی اطلاعات حرکتی دقت را کاهش داده‌اند و الگوریتم‌های پیچیده برای ایده‌شان مطرح کرده‌اند. در اینجا ما به دنبال راهی هستیم که علاوه بر نگه‌داشتن دقت در سطحی معقول، سربار زمانی مقایسه را نیز حذف کنیم. همچنین باتوجه به ویژگی‌هایی که در ویدئوها نظیر تخمین حرکت وجود داشت، به این نتیجه رسیدیم که از این نوع ویژگی‌ها برای استفاده مجدد از محاسبات بدون انجام مقایسه می‌توانیم استفاده کنیم. در مرحله فشرده‌سازی ویدئوها، ابتدا عملیات تخمین حرکت انجام می‌شود. در این مرحله برای تمامی بلوک‌ها باتوجه به بهترین گزینه برای این بلوک در قاب قبل، بردار حرکت‌ای تعریف می‌شود. این بهترین گزینه، همان عدد تشابه است که معیار بی‌نظیری برای کشف تشابه و بازاستفاده از نتایج آن بلوک برای کاهش محاسبات شبکه عصبی به حساب می‌آید که بدون محاسبات خاصی در اختیار ما است.

براین اساس نوآوری‌های این طرح عبارتند از:

• استفاده از تخمین حرکت برای کاهش سربار محاسباتی و استفاده مجدد

• استفاده از ترکیب برای افزایش اثربخشی روش پیشنهادی

به این ترتیب، ابتدا الگوریتمی برای یافتن بلوک‌های ویدئویی با پتانسیل استفاده مجدد از محاسبات و روش‌هایی برای تقریب محاسبات آن‌ها بر اساس بردارهای حرکت کدگذاری شده پیشنهاد می‌دهیم. سپس برای بلوک‌هایی در عملیات استفاده مجدد شرکت ندارند و بایستی تمام محاسبات برای آنها اعمال شود، از روش لایه‌های ترکیبی برای کاهش انتقال داده‌های نگاشت ویژگی استفاده می‌کنیم. این روش به الگوریتم شتابدهی اشاره دارد که وظیفه دارد مراحل مختلف یک CNN را با هم ترکیب کرده و پردازش آن‌ها را بهینه و سریع‌تر انجام دهد. علاوه بر این، از خطاهای تخمین حرکت (مشتق‌شده از ویدئوی فشرده) برای کنترل ریسک تقریب استفاده می‌کنیم.

این روش می‌تواند محاسبات قابل توجهی را صرفه‌جویی کند، درحالی که منجر به کاهش ناچیز خطای نهایی نگاشت ویژگی شود. این اقدامات باعث کاهش توان مصرفی در کاربردهای نهفته خواهد شد.