دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی

گرایش مهندسی فناوری اطلاعات

عنوان:

بازشناسی کنش انسان از داده‌های اسکلتی توسط شبکه‌های عصبی گراف ـ پیچشی زمان ـ مکانی با مدل توجه

نگارش:

رضا رحیمی آذغان

استاد راهنما:

دکتر کسایی

شهریور ۱۳۹۸


بسم الله الرحمن الرحیم

به نام خدا

دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی

عنوان: بازشناسی کنش انسان از داده‌های اسکلتی توسط شبکه‌های عصبی گراف_ پیچشی زمان_ مکانی با مدل توجه

نگارش: رضا رحیمی آذغان

کمیته‌ی ممتحنین

استاد راهنما: دکتر کسایی امضاء:

استاد مدعو: دکتر همت‌یار امضاء:

تاریخ:

سپاس

از استاد بزرگوارم که با کمک‌ها و راهنمایی‌های بی‌دریغشان، بنده را در انجام این پروژه یاری داده‌اند، تشکر و قدردانی می‌کنم. هم‌چنین از سرکار خانم اسدی که در هر مرحله از پروژه بنده را راهنمایی کردند و راه صحیح را به نشان دادند از صمیم قلب سپاس‌گزارم.

چکیده

در سال‌های اخیر، بازشناسی کنش انسان[۱] از عمده‌ترین زمینه‌های مورد بحث و تحقیق در دنیای علوم و مهندسی کامپیوتر بوده است.[۱] با افزایش چشم‌گیر اطلاعات در دسترس و پیشرفت روزافزون شبکه‌های عصبی[۲]، جلوه‌ی جدیدی به موضوع بازشناسی کنش انسان داده شده است. از مقدمات این موضوع، بحث نحوه‌ی نمایش داده‌های بدن انسان در دو بعد زمان و مکان است. با ظهور سنسورهای کینکت، دو روش عمده برای بهینه‌کردن هرچه بیش‌تر نمایش این داده‌ها وجود دارد. این دو روش شامل استفاده از اطلاعات RGB-D و استفاده از اطلاعات سه‌بعدی اسکلت‌های بدن هستند.[۲] اخیرا، به‌دلیل کم‌بودن حجم داده‌های اسکلتی، اطلاعات معنائی[۳] بالا و نیز خاصیت مقیاس‌پذیری آن‌ها، مطالعات بسیاری بر روی نمایش داده‌ها بر این روش صورت گرفته است.

در این پروژه مطالعه بر روی نمایش داده‌های سه‌بعدی اسکلتی و استفاده از آن برای بازشناسی کنش انسان ادامه پیدا می‌کند. شبکه‌ی مورداستفاده در این پروژه شبکه‌ی پیچشی‌ـ گرافی زمان‌ـ مکانی [۴] است که تعمیمی بر شبکه‌ی پیچشی‌ـ گرافی[۵] است که آن نیز تعمیمی بر شبکه‌ی پیچشی[۶] است. هم‌چنین سعی بر آن بوده است که با ارائه‌ی یک مدل توجه[۷] و با استفاده از معیار فلان شبکه‌های عصبی موجود را بهبود بخشید. در کارهای آینده نیز، می‌توان مدل‌های توجه بهینه‌تری را معرفی نمود و تاثیر هرکدام را بر شبکه‌ی مورد استفاده در این پروژه مطالعه نمود.

کلیدواژه‌ها: بازشناسی کنش انسان، اطلاعات سه‌بعدی اسکلت، شبکه‌های گراف‌ـ پیچشی، مدل توجه

---

[۱] Human Action Recognition
[۲] Neural Networks
[۳] Semantic Information
[۴] Spatial Temporal Graph Convolutional Networks
[۵] Graph Convolutional Networks
[۶] Convolutional Networks
[۷] Attention Model

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل ۱

مقدمه

۱ـ۱ تعریف مسئله

یکی از حوزه‌های بسیار پرکاربرد و پرمباحثه در زمینه‌ی پردازش تصویر و بینایی کامپیوتر، بازشناسی کنش انسان است. از دهه‌ی ۸۰ میلادی، این حوزه به‌دلیل کاربرد بسیار بالایی که در زمینه‌های تعامل انسان و کامپیوتر و هم‌چنین پزشکی می‌تواند داشته باشد، توجه بسیاری از افراد فعال در علوم کامپیوتر را به خود جلب کرده است.[۳] در این حوزه، با استفاده از اطلاعات استخراج‌شده از ویدیو، کنش صورت گرفته در آن ویدیو شناسایی می‌شود. در حقیقت می‌توان گفت که این حوزه، تعمیمی بر مساله‌ی دسته‌بندی تصاویر[۱] است که در آن اشیای موجود در یک تصویر شناسایی شده و به دسته‌ی خاص خودشان نسبت داده می‌شوند. همان‌گونه که در حوزه‌ی دسته‌بندی تصاویر، مساله‌ی چگونگی نمایش تصویر مطرح است، در حوزه‌ی بازشناسی کنش انسان نیز یکی از مسائل اساسی نحوه‌ی نمایش اطلاعات و به زبان دقیق‌تر، نحوه‌ی نمایش حرکت بدن انسان موجود در ویدیو است. دو راه‌حل مهم برای رفع این مشکل به‌صورت بهینه، نمایش اطلاعات RGB-D و اطلاعات سه‌بعدی اسکلت انسان هستند.[۲][۱] تفاوت این دو روش در شکل ۱ـ۱ به‌وضوح به تصویر کشیده شده است.

نمایش اطلاعات در قالب مختصات سه‌بعدی اسکلت بدن، روشی است که در این پروژه مورد استفاده قرار می‌گیرد. بازشناسی کنش انسان با این روش نمایش، مدت زمان زیادی است که در حوزه‌ی بینایی رایانه‌ای مورد کند و کاو قرار گرفته است. الگوریتم‌های قدیمی‌تر و مبتنی بر روش‌های دست‌ساز

[۱] Image Classification

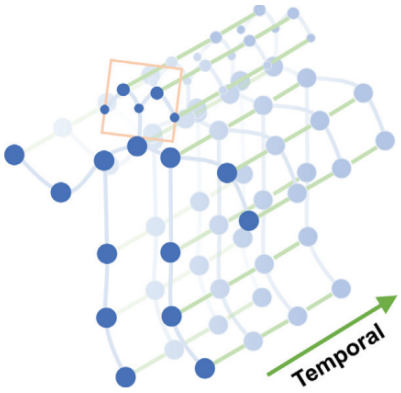
الف ب ج

شکل ۱ ــ ۱: الف ــ نمایش اسکلت‌های بدن که از تصویر استخراج شده‌اند، ب ــ تصویر ژرفا[۲]، ج ــ تصویر رنگی [۴]

بیش‌تر از یک‌سری قوانین و روش‌های نسبتا ثابت و انعطاف‌ناپذیر استفاده می‌کردند. به همین دلیل میزان خطای آن‌ها بالا بود و برای برخی موارد خاص و پیچیده به هیچ وجه قابل پیاده‌سازی نبودند.[۵] با رشد روزافزون یادگیری ژرف[۳] و هم‌چنین افزایش اطلاعات در دسترس، شبکه‌های عصبی سرتاسر[۴] روش‌های جدیدتر و بهتری برای مسائلی هم‌چون بازشناسی کنش فراهم آوردند. در این‌گونه شبکه‌ها ورودی به لایه‌ی ابتدایی شبکه داده شده و خروجی از انتهای آن دریافت می‌گردد. هم‌چنین با یک معیار مناسب و مقایسه‌ی خروجی با این معیار، پارامترهای برنامه (توسط الگوریتم‌های بهینه‌سازی) به‌گونه‌ای تغییر می‌کنند که خروجی به معیار نزدیک و نزدیک‌تر شود. به این الگوریتم، الگوریتم یادگیری گفته می‌شود. در عین این‌که این الگوریتم به‌صورت کلی توضیح داده شد، معماری‌های شبکه‌ی[۵] بسیار متنوعی وجود دارند که برای حل مسائل مختلف مورد استفاده قرار می‌گیرند.

یکی از معماری‌های بسیار پرکاربرد، شبکه‌های عصبی پیچشی[۶] هستند. این شبکه‌ها بیش‌تر در مسائل مربوط به پردازش تصویر استفاده می‌شوند. بسیاری از حوزه‌های معروف پردازش تصویر، از قبیل آشکارسازی اشیا[۷]، دسته‌بندی تصاویر و ... با استفاده از این معماری روش حل بسیار بهینه‌تری پیدا کرده‌اند. شبکه‌های عصبی پیچشی ــ گرافی[۸] تعمیمی بر این معماری است. این مدل، به‌جای یک تصویر، گرافی را به‌عنوان ورودی گرفته و الگوریتم را بر روی آن اجرا می‌کند. این شبکه نیز کاربرد زیادی در مسائلی چون دسته‌بندی تصاویر، پردازش متن و ... دارد.[۶] شبکه‌ی استفاده شده در این پروژه،

[۳] Deep Learning
[۴] End to End Neural Networks
[۵] Network Architecture
[۶] Convolutional Neural Networks
[۷] Object Detection
[۸] Graph Convolutional Neural Network



شکل ۱ ـ ۲: نمایشی از ابعاد زمان‌ـ‌مکانی اسکلت انسان که ST-GCN بر روی آن کار می‌کند. [۵]

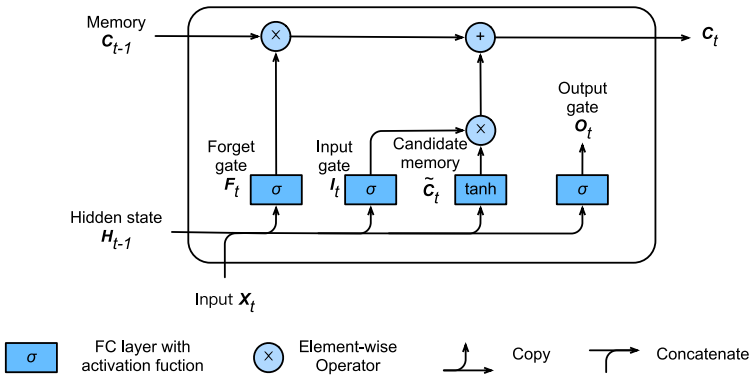
شبکه‌ی عصبی پیچشی‌ـ گرافی زمان‌ـ مکانی[۹] است که به اختصار ST-GCN نام دارد. شکل ۱ ـ ۲ (که نشان‌دهنده‌ی ورودی ST-GCN است) ایده‌ی کلی و مختصری از چگونگی کارکرد این شبکه را نمایش می‌دهد. توضیحات هر کدام از این معماری‌ها در ادامه‌ی پروژه به تفصیل آمده‌اند.

۱ ـ ۲ اهمیت موضوع

همان‌گونه که قبلا اشاره شد، بازشناسی کنش انسان یکی از پرکاربردترین مباحث در حوزه‌ی بینایی ماشین است. کاربردهای این حوزه از مسائلی هم‌چون سرگرمی تا موارد پزشکی متغیر است. می‌توان ادعا کرد که بازشناسی کنش انسان، هدف اصلی سیستم‌های هوشمند ویدیویی[۱۰] است.[۱] در حوزه‌هایی هم‌چون تعامل انسان با ماشین، با بازشناسی کنش انسان، می‌توان عکس‌العملی در خور عمل صورت گرفته انجام داد. در موارد پرشکی، با شناخت دقیق کنش، امکان فیزیوتراپی برای بیماران با مشکلات جسمانی وجود دارد. در موارد امنیتی[۱۱]، می‌توان بدون این‌که ناظر انسانی وجود داشته باشد، با اکتفا بر کامپیوترها نظارت ویدیویی لازم را اعمال کرد.

با توجه به این کاربردهای گسترده، وجود روشی برای بهینه‌سازی بازشناسی کنش انسان از ملزومات این بحث تلقی می‌شود. بسیاری از سیستم‌هایی که نیازمند به داشتن ویژگی بازشناسی کنش هستند، بایستی بصورت بی‌درنگ[۱۲] عمل کنند. به همین دلیل زمان موجود برای محاسبات و نیز ضریب خطا،

[۹] Spatial Temporal Graph Convolutional Neural Network
[۱۰] Intelligent Video Systems
[۱۱] Surveillance
[۱۲] Real Time



شکل ۱ـ۳: نمای کلی از یک LSTM [۹]

تا جای ممکن بایستی کاهش یابد. روش مورد استفاده در این پروژه، داده‌های کم‌تری نسبت به سایر روش‌ها استفاده می‌کند و به همین دلیل علاوه بر انعطاف پذیری بالا، سرعت و اطمینان بالایی را نیز تامین می‌کند.

۱ـ۳ ادبیات موضوع

یکی از روش‌های قدیمی برای مسائل در حوزه‌ی پردازش تصویر مانند بازشناسی کنش، استفاده از روش‌های مبتنی بر ویژگی‌های دست‌ساز[۱۳] است.[۷][۵][۸] در این روش، با کمک آشکارسازها برخی از نواحی تصویر را شناسایی کرده و با استفاده از این نواحی دسته کنش صورت گرفته را شناسایی می‌کنند. [۷] برای داده‌های ورودی کم، این روش به‌خوبی و با درصد موفقیت بالاتری عمل می‌کند. هرچند در صورتی که داده‌ی ورودی به فراوانی در دسترس باشد، استفاده از شبکه‌های ژرف بهینه‌ترین راه‌حل موجود است.

یکی از روش‌های معروف برای بازشناسی کنش انسان، استفاده از شبکه‌های عصبی بازگشتی [۱۴] است که به اختصار RNN نامیده می‌شوند.[۱۰] در بسیاری از مقالات از مدل تعمیم‌یافته‌ی این شبکه‌ی عصبی، که به حافظه‌ی کوتاه مدت بلند[۱۵] یا LSTM معروف است استفاده می‌کنند.[۱۰][۱۱] تصویر ۱ـ۳ نمایی از یک سلول LSTM را نشان می‌دهد. در این شبکه‌ی عصبی، ورودی‌ها، که قاب‌های[۱۶]

[۱۳] Handcrafted Features
[۱۴] Recurrent Neural Network
[۱۵] Long Short-Term Memory
[۱۶] Frames

جدول ۱ـ۱: نتایج برخی آزمایش‌ها بر روی پایگاه داده‌ی NTU RGB+D [۱۱]

روش	CS	CV
Skeletal Quads	٪۳۸/۶	٪۴۱/۴
Lie Group	٪۵۰/۱	٪۵۲/۸
Dynamic Skeletons	٪۶۵/۲	٪۶۰/۲
HBRNN	٪۵۹/۱	٪۶۴
Deep RNN	٪۵۶/۳	٪۶۴/۱
Deep LSTM	٪۶۰/۷	٪۶۷/۳
Part-aware LSTM	٪۶۲/۹	٪۷۰/۳
JTM CNN	٪۷۳/۴	٪۷۵/۲
SkeletonNet	٪۷۵/۹	٪۸۱/۲
Visualization CNN	٪۷۶	٪۸۲/۶

ویدیو هستند، به صورت سری داده می‌شوند. شبکه تعدادی دروازه[۱۷] دارد که با یادگیری(بهبود) آن‌ها به مرور زمان می‌تواند به یاد داشته باشد که کدام قاب‌ها اطلاعات بیش‌تری در اختیار شبکه می‌گذارند تا بر روی آن‌ها تمرکز بیش‌تری بگذارد و آن قاب‌ها را در طول زمان بیش‌تر به یاد داشته باشد و سایر اطلاعات را فراموش کند.[۱۲]

در برخی از مقالات با تغییراتی بر روی این شبکه نتایجی حاصل شده است که به برخی از آن‌ها در جدول ۱ـ۱ ذکر شده است.

۱ـ۴ چالش‌ها

برخی از چالش‌های موجود در بازشناسی کنش از روی داده‌های اسکلتی به شرح زیر می‌باشند:

- بزرگ‌ترین مساله (که پیش‌روی هرگونه شبکه‌ی سرتاسر قرار دارد) وجود مجموعه‌داده‌ی مناسب برای تضمین عملکرد بهینه است. همان‌گونه که اشاره شد، برتری اصلی این شبکه‌ها نسبت به

[۱۷] Gate

ویژگی‌های دست‌ساز، زمانی حاصل می‌شود که داده‌ی کافی در اختیار شبکه باشد.

- ورودی هرگونه شبکه‌ی بازشناسی کنش، دنباله‌ای از قاب‌ها در حوزه‌ی زمان است. این موضوع (بخصوص برای مجموعه‌داده‌های با وضوح بالاتر) باعث افزایش قدرت پردازشی و حافظه‌ی موردنیاز می‌شود. علاوه بر اندازه‌ی مجموعه‌ی داده، اندازه و تعداد پارامترهای شبکه از مواردی است که پیچیدگی زمانی به روش حل وارد می‌کنند.
- ناهنجاری و پیچیدگی در مجموعه‌داده‌های موجود، سرعت آموزش در شبکه‌ی عصبی را کاهش می‌دهد. به‌عنوان مثال زوایای متفاوت برای ویدیوهای مختلف، تعداد افراد حاضر در یک ویدیو، سرعت متفاوت انجام کنش توسط افراد مختلف و ... از جمله مواردی هستند که کیفیت شبکه را کاهش می‌دهند.

۱ـ۵ فرضیات

برای روشی که در این پروژه انتخاب شده است، برخی فرضیات از ابتدا در نظر گرفته شده است.

- با استفاده از مجموعه داده‌ی NTU RGB+D[۱۳]، داده‌های اسکلتی آماده هستند و نیازی به استفاده از الگوریتم‌هایی مانند تخمین حالت [۱۸]، برای استخراج این داده‌ها نیست.[۱۳][۵]
- در مجموعه داده‌ی اشاره‌شده، تنها یک کنش صورت می‌گیرد و اگر در قابی بیش از یک کنشگر موجود باشد، کنش (تعامل) بین این دو (و نه به صورت جداگانه) انجام خواهد شد.
- در این مجموعه داده، در هر قاب حداکثر دو کنشگر موجود است و پس‌زمینه نیز به‌دلیل استفاده از سنسورهای کینکت[۱۹] حذف شده‌اند.

۱ـ۶ اهداف تحقیق

در این پایان‌نامه، سعی شده است که بازشناسی کنش با استفاده از شبکه‌های ST-GCN و لایه‌های توجه انجام گیرد. برای یادگیری، از مجموعه‌داده‌ی NTU RGB+D[۱۳] استفاده می‌شود که بزرگترین

[۱۸] Pose Estimation
[۱۹] Kinect Sensors

مجموعه‌داده‌ی شامل اطلاعات سه‌بعدی اسکلت بدن هستند. این مجموعه‌داده از دو سنجه[۲۰] تشکیل یافته‌است که جزئیات هرکدام در انتهای پایان‌نامه تشریح خواهد شد. برای سنجش خروجی نیز از معیار تابع هزینه‌ی آنتروپی متقابل[۲۱] استفاده می‌شود.

۱ـ۷ ساختار پایان‌نامه

این پایان‌نامه شامل پنج فصل است. فصل دوم دربرگیرنده‌ی ادبیات مربوطه با پایان‌نامه است. در فصل سوم روش‌های پیاده‌سازی‌شده در این پروژه به تفصیل بیان گردیده است. فصل چهارم شامل نتایج تجربی به دست آمده از آزمودن روش پیشنهادی و مقایسه این نتایج با نتایج برخی روش‌های قبلی که روی مجموعه‌داده‌ی NTU RGB+D پیاده‌سازی و آزمون شده‌اند، است. بالاخره، جمع‌بندی کلی و راه‌کارهای ممکن برای ادامه‌ی این پروژه در فصل پنجم آورده شده است.

[۲۰] Benchmark
[۲۱] Cross Entropy Loss Function

فصل ۲

ادبیات مربوطه

۲ـ۱ شبکه‌های عصبی پیچشی

«پیچش» یک عمل‌گر خطی است که برای توابع n بعدی تعریف می‌شود. مقدار آن برای توابع تک متغیره از فرمول

$$f(t) * g(t) = \int_{-\infty}^{+\infty} f(\tau).g(t-\tau)d\tau \qquad (۲ـ۱)$$

محاسبه می‌شود.[۱۴] همین فرمول با اندکی تغییر برای توابع با دو متغیر (ماتریس‌های دوبعدی) مانند تصاویر، به شکل

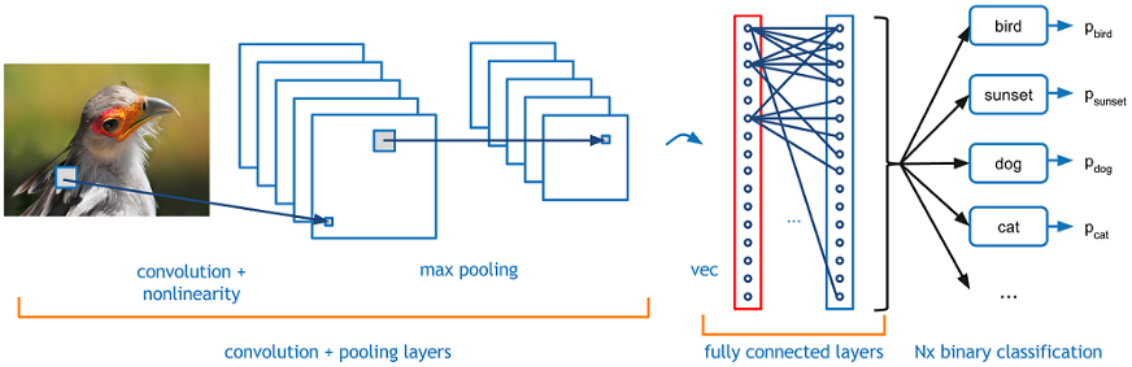
$$x[m,n] * k[m,n] = \sum_{j=-\infty}^{+\infty} \sum_{i=-\infty}^{+\infty} k[i,j].x[m-i,n-j] \qquad (۲ـ۲)$$

تعریف می‌شود.[۱۵]

در فرمول ۲ـ۲ اگر $x[m,n]$ تصویر ورودی باشد، به عمل‌وند $k[m,n]$ هسته [۱] گفته می‌شود. هسته مربعی به ابعاد ۳×۳، ۵×۵ یا ... است که با انتخاب مناسب محتوا و ابعاد آن، تصویر خروجی می‌تواند دارای ویژگی‌های خاص مربوط به تصویر اصلی باشد. مثلا با انتخاب هسته‌ی خاص، می‌توان لبه‌ها را در تصویر شناسایی کرد.

s

[۱] Kernel



شکل ۲ ـ ۱: نمای کلی از یک شبکه‌ی پیچشی [۱۶]

ایده‌ی اصلی در CNN این است که مقادیر هسته ثابت در نظر گرفته نشود و به‌عنوان پارامترهای شبکه در هر مرتبه[۲] به‌روزرسانی شوند.[۱۷] به‌عنوان مثال، پیچش[۳] تصویر با ابعاد ۱۹۲۰×۲۱۸۰ با هسته‌ای به ابعاد ۳×۳ را در نظر بگیرید. این شبکه‌ی ساده و تک‌لایه ۹ پارامتر دارد که می‌توان با تعریف تابع هزینه‌ی مناسب، برای رسیدن به خروجی مطلوب، مدام این ۹ پارامتر را تغییر داد. شکل ۲ ـ ۱ نمای کلی از یک CNN نشان می‌دهد.

یک CNN اغلب از سه دسته لایه تشکیل می‌شود. لایه‌های پیچشی[۴]، لایه‌های ادغام[۵] و لایه‌های کاملا متصل[۶]. توضیح هر کدام از این لایه‌ها به اختصار آمده است.

ایده‌ی کلی پشت لایه‌های پیچشی، همان ایده‌ی اصلی CNN است. به ازای هر لایه در این دسته از لایه‌ها، یک یا چند هسته وجود دارد که تصویر را فیلتر می‌کنند. در نتیجه‌ی این فیلتر برخی از ویژگی‌های تصویر استخراج می‌شود و تصویر کاهش اندازه می‌دهد. هرچند اگر تصویر از چندین کانال[۷] تشکیل شده باشد، تعداد این کانال‌ها رفته رفته بیش‌تر خواهد شد. در ابتدا اکثر تصاویر شامل سه کانال قرمز، سبز و آبی[۸] هستند.

برخی از لایه‌ها عمل «ادغام»[۹] را نیز بر روی تصویر انجام می‌دهند. ادغام انواع مختلفی همانند

[۲] Iteration
[۳] Convolve
[۴] Convolutional Layers
[۵] Pooling Layers
[۶] Fully Connected Layers
[۷] Channel
[۸] RGB Channels .
[۹] Pooling

«ادغام بیشینه»[۱۰]، «ادغام میانگین»[۱۱] و ... دارد. به عنوان مثال اگر بخواهیم با یک فیلتر به‌اندازه‌ی ۳×۳، تصویری به اندازه‌ی H×W را ادغام بیشینه کنیم، بایستی از گوشه‌ی سمت چپ تصویر شروع کرده و فیلتر را بر روی تصویر بگذاریم. بیشینه مقدار پیکسل‌هایی از تصویر که زیر فیلتر قرار گرفته‌اند، پیکسل اولِ خروجی خواهد بود. سپس فیلتر را یک پیکسل به راست انتقال می‌دهیم و ... . دقت شود که محتویات هسته در عمل ادغام اهمیتی ندارد. به همین دلیل در ادغام کردن، پارامتری برای یادگیری وجود نخواهد داشت.

لایه‌های کاملا متصل بعد از لایه‌های پیچشی و لایه‌های ادغام در CNN قرار می‌گیرند. بعد از چندین لایه‌ی پیچشی و ادغام، تمامی پیکسل‌های خروجی را به یک شبکه‌ی کاملا متصل (مانند شبکه‌ی رگرسیون خطی[۱۲]) می‌دهند تا خروجی نهایی بعد از چندین لایه‌ی کاملا متصل به دست آید. الگوریتم یادگیری برای این لایه‌ها مانند الگوریتم‌های معمول برای شبکه‌های عصبی ساده‌ای چون رگرسیون خطی است.

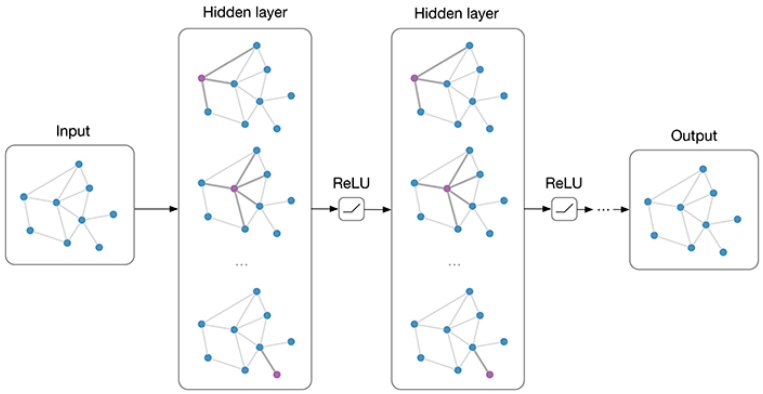
۲ـ۲ شبکه‌های عصبی گراف‌ـ‌پیچشی

امروزه بسیاری از مجموعه داده‌های موجود، مثل اطلاعات شبکه‌های اجتماعی، شبکه‌ی اینترنت و ... به شکل گراف هستند.[۶] همان‌گونه که پیش از این ذکر شد، شبکه‌های عصبی گراف‌ـ‌پیچشی، که به اختصار GCN نامیده می‌شوند، تعمیمی بر CNN هستند که در آن ورودی به‌جای تصویر، یک گراف است. هرچند تفاوت‌ها در همین‌جا به انتها نمی‌سد و پیچیدگی‌های ساختاری یک گراف، دشواری‌های خاصی را به این شبکه‌ها تحمیل کرده است. شکل ۲ـ۲، نمایی کلی از یک GCN را نمایش می‌دهد.

به‌صورت دقیق هر GCN دو ماتریس را به‌عنوان ورودی دریافت می‌کند.

- ماتریس X به ابعاد $N \times F$، که N تعداد رئوس و F، تعداد ویژگی‌های ورودی[۱۳] برای هر راس است.
- ماتریس مجاورتِ A به ابعاد $N \times N$ که ساختار کلی گراف را مشخص می‌کند.

[۱۰] Max Pooling
[۱۱] Average Pooling
[۱۲] Linear Regression
[۱۳] Input Features



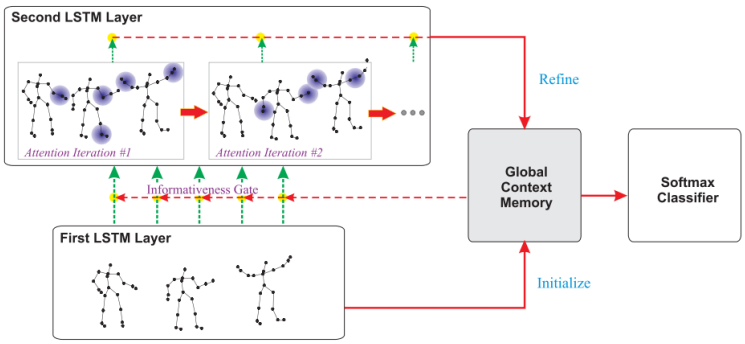
شکل ۲ ـ ۲: نمای کلی از یک GCN [۱۸]

حال با تعریف وزن (هسته‌ی) مناسب برای هر لایه از شبکه، می‌توان شبکه را آموزش[۱۴] داد. هرچند هنوز هم برخی از مشکلات برای شبکه‌های نسبتا بزرگ وجود دارد. مشکلاتی که در این پروژه به آن‌ها برخورد شد و راه‌حل پیشنهادی به شرح زیر هستند.

- همان‌گونه که ذکر شد، یکی از ورودی‌های شبکه، ماتریس مجاورت گراف است. در صورتی که یک راس یال بازگشتی به خودش نداشته باشد، درایه‌ی نظیر آن راس در ماتریس مجاورت صفر خواهد بود. همین موضوع باعث می‌شود که در مسیر لایه‌های شبکه، تنها ویژگی‌های ورودی رئوس مجاور آن راس در درایه‌ی نظیر آن وجود داشته باشند. به‌عبارت دیگر بعد از طی یک لایه، ویژگی‌های رئوس بدون یال بازگشتی تقریبا فراموش خواهند شد. به همین دلیل، قبل از هرکاری به تمامی رئوس گراف یک یال بازگشتی اضافه می‌شود. این کار با جمع کردن ماتریس مجاورت با ماتریس واحد (I) انجام می‌گیرد.

- رئوس با درجه‌ی بالاتر(پایین‌تر) رفته رفته اندازه‌ی بزرگ‌تری (کوچک‌تری) خواهند داشت. به این موضوع انفجارگرادیان[۱۵] و میرایی گرادیان[۱۶] گفته می‌شود. به همین دلیل عادی‌سای[۱۷] ورودی شبکه یک موضوع اجتناب‌ناپذیر است. برای این کار از روش موجود در [۶] استفاده شده است.

[۱۴] Train
[۱۵] Exploding Gradient
[۱۶] Vanishing Gradient
[۱۷] Normalization



شکل ۲ـ۳: مدل توجه با استفاده از حافظه‌ی زمینه‌ی سراسری در یک LSTM [۱۱]

۲ـ۳ مدل‌های توجه

معماری‌ها و مدل‌هایی که پیش‌تر معرفی شد، تفاوتی بین نقاط مختلف یک تصویر یا ویدیو قائل نبودند. درحالی که برای انجام یک الگوریتم بر روی تصویر، برخی جزئیات نه‌تنها مهم نیستند، بلکه در نتیجه‌ی نهایی خلل ایجاد می‌کنند.[۱۱] به‌عنوان مثال در فرآیند تشخیص تصویر، پس‌زمینه‌ی شی موردنظر اهمیتی ندارد. هم‌چنین در مثال خاص این پروژه، وقتی که کنش صورت‌گرفته دست‌زدن است، مفاصل پای یک شخص اهمیت چندانی ندارد. به همین دلیل، برای بهینه‌سازی بیش‌تر و درصد خطای پایین‌تر، بهتر است که شبکه رفته رفته متوجه شود که به کدام یک از جزئیات تصویر یا ویدیو بیش‌تر از باقی اجزا اهمیت قائل شود. یکی از مدل‌های توجه استفاده‌شده، مدل موجود در [۱۱] است که در معماری LSTM به کار گرفته شده است. در این مدل علاوه بر دروازه‌های موجود معماری LSTM ، یک حافظه‌ی زمینه‌ی سراسری[۱۸] هم اضافه شده است. هم‌چنین از دو لایه LSTM استفاده شده است که لایه‌ی اول این حافظه را مقداردهی اولیه میکند و لایه‌ی دوم آن را بهبود می‌بخشد. در نهایت مقدار این حافظه است که به دسته‌بند بیشینه‌ی هموار[۱۹] داده می‌شود تا خروجی مورد نظر حاصل شود. طریقه‌ی مقداردهی و بهبودبخشی به حافظه، روشی مشابه دروازه‌های معمول LSTM دارد. تصویر ۲ـ۳ استفاده از این مدل توجه را نمایش می‌دهد.

روش دیگر به‌کارگیری مدل توجه (که در این پروژه هم از آن استفاده شده است) استفاده از پارامترهای قابل آموزش به‌ازای هر مفصل است. این مدل مناسب شبکه‌های CNN یا مشتقات آن مانند GCN

[۱۸] Global Context Memory
[۱۹] Softmax Classifier

است.[۵] در این مدل، از پارامترهای منسوب به وزن‌های اهمیت مفصل [۲۰] استفاده می‌شود که به هر مفصل یک وزن مشخص می‌دهد. این وزن بعد از آموزش کل شبکه مقدار بهینه پیدا می‌کند. سپس در هنگام ارزیابی، این وزن بر روی هسته‌ی موجود در شبکه‌ی GCN ضرب می‌شود تا هر یال تاثیر مشخصی بر روی جواب نهایی شبکه داشته باشد.

روش دیگری که در حوزه‌ی پردازش تصویر بسیار جدید است، استفاده از لایه‌ی خاصی به اسم لایه‌ی ادغام توجه[۲۱] است. اولین استفاده از این روش در بازشناسی کنش انسان در [۱۹] صورت گرفته است. در این روش، لایه‌های کاملا متصل از لایه‌های پیچشی تغذیه نمی‌شوند. بلکه قبل از لایه‌های کاملا متصل، خروجی‌های لایه‌های پیچشی به لایه‌های ادغام توجه داده می‌شود. در این گونه از لایه‌های ادغام، کل ورودی تحت ادغام قرار می‌گیرد و خروجی آن به لایه‌های کاملا متصل یا به یک لایه‌ی دسته‌بند بیشینه‌ی هموار داده می‌شود. (به تفاوت این نوع ادغام با گونه‌های قبلا معرفی شده توجه شود که در گونه‌های قبلی، مربع‌های $k \times k$ از گوشه‌ی سمت بالا_چپ تصویر انتخاب شده و تحت ادغام قرار می‌گیرند). جزئیات این روش در فصل آتی توضیح داده خواهند شد.

[۲۰] Edge Importance Weighting
[۲۱] Attention Pooling Layer

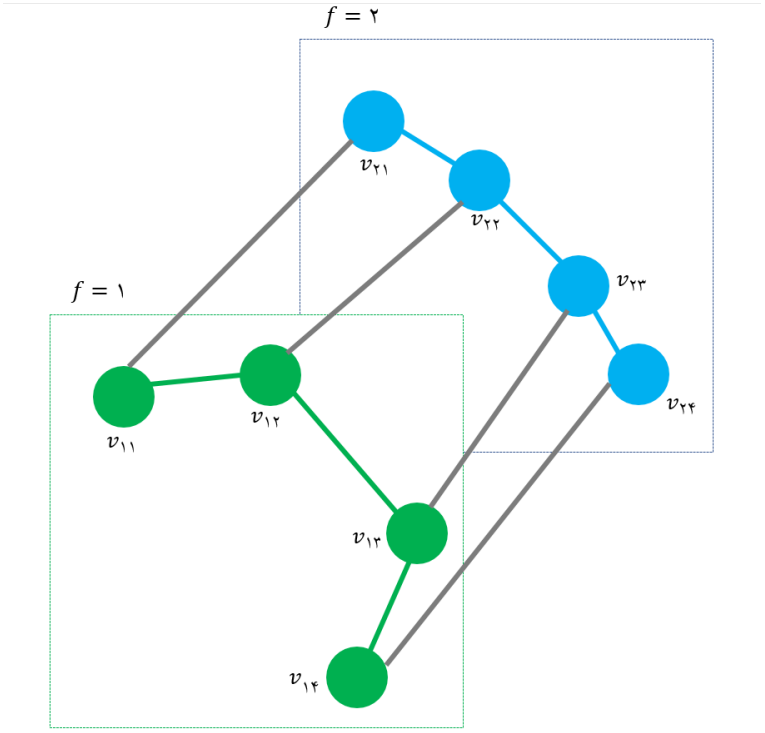
فصل ۳

روش‌های پیشنهادی

۳ـ۱ مقدمه

این پروژه ادامه‌ای است بر کار [۵] و [۱۹]. هردوی این مقالات، با استفاده از شبکه‌های عصبی سرتاسر، به بازشناسی کنش پرداخته‌اند. در این پایان‌نامه نیز سعی شده است که با استفاده از روش‌های هر دوی این مقاله‌ها، درصد خطای پایین‌تری را به دست آورد. در ابتدای این فصل کلیت شبکه و برخی ریزه‌کاری‌ها که بایستی اعمال شوند توضیح داده می‌شود. در ادامه مدل توجه مورد استفاده و چگونگی کار آن شرح داده خواهد شد. در انتها نیز معیاری که برای سنجش کار استفاده خواهیم کرد معرفی می‌شود. تصویر ۳ـ۱ نموداری مختصر از کل کار انجام‌گرفته را شرح می‌دهد.

شکل ۳ـ۱: نمودار بلوکی برای شبکه‌ی ST-GCN با مدل توجه ادغامی



شکل ۳ـ۲: مثالی از یک گراف زمان‌ـمکانی

۳ـ۲ شبکه‌ی استفاده شده

شبکه‌ی مورد استفاده در این پروژه، شبکه‌ی گراف‌ـ پیچشی زمان‌ـ مکانی (ST-GCN) است. بزرگ‌ترین تفاوت این شبکه با شبکه‌ی گراف‌ـ پیچشی، در نوع ورودی‌ای است که به آن داده می‌شود. ورودی st-gcn یک گراف $G = (V, E)$ است که در آن V مجموعه‌ی رئوس و E مجموعه‌ی یال‌های گراف است. مجموعه‌ی V تعریفی مانند

$$V = \{v_{fi} | f = ۱, ..., F, i = ۱, ..., N\} \quad (۱ـ۳)$$

دارد. [۵]

در رابطه‌ی ۳ـ۱ ، f شاخص شماره‌ی قاب[۱] و i شاخص شماره‌ی راس (مفصل) در یک قاب است. همان‌گونه که از این رابطه مشخص است، رئوس گراف تمامی مفاصل در تمامی قاب‌ها را شامل می‌شود و محدود به یک قاب نیست. به همین دلیل تعریف یال نیز بایستی شامل تمامی ارتباطات متصل کننده‌ی این رئوس باشد. [۵] مجموعه‌ی یال‌ها را به دو زیرمجموعه‌ی یال‌های یک قاب (E_s) و یال‌های بین

[۱] Frame

دو قاب (E_t) تفکیک کرده و هرکدام را به شکل

$$E_s = \{v_{fi}v_{fj}|(i,j) \in H\},$$
$$E_t = \{v_{fi}v_{(f+1)i}\} \quad (۲\_۳)$$

تعریف می‌کند.[۵]

در رابطه‌ی ۳_۲ مجموعه‌ی H مجموعه‌ی شامل مفاصل مجاور بدن انسان هستند. توجه کنید که چگونه مجموعه یال‌های E_t مفاصل متناظر در دو قاب را به هم‌دیگر متصل می‌کند.

برای انجام عمل پیچش در این گراف، بایستی رابطه‌ی همسایگی را برای هر راس در گراف تعریف کنیم. چرا که پارامتر وزن در عمل پیچش، به‌ازای هر راس، بر روی همسایه‌های آن راس شناور خواهد بود. برای هر راس مانند v_{fi} همسایه‌های آن با رابطه‌ی

$$N(v_{fi}) = \{v_{qj}|d(v_{fj}, d_{fi}) \leqslant K, |q-f| \leqslant \left\lfloor \frac{\Gamma}{۲} \right\rfloor\} \quad (۳\_۳)$$

تعریف شده‌اند.[۵]

در رابطه‌ی ۳_۳، تابع d کوتاه‌ترین مسیر بین دو راس ورودی آن را مشخص می‌کند. هم‌چنین متغیر K حداکثر فاصله بین رئوس همسایه در یک قاب و متغیر Γ حداکثر فاصله بین دو راس همسایه در دوقاب مختلف را بیان می‌کند. به بیان دیگر، می‌توان گفت که K اندازه‌ی ماتریس وزن در بعد مکان و Γ اندازه‌ی آن در بعد زمان است. برای افزایش سرعت کار در این پروژه، $K = ۱$ و $\Gamma = ۲$ در نظر گرفته شده است. مقادیر بالاتر از این می‌تواند در کارهای آتی مورد بررسی قرار گیرد.

شکل ۳_۲ روابطی که تا این‌جای کار بیان شد را ترسیم کرده است. در این شکل، یال‌های یک قاب به‌صورت رنگی و یال‌های بین دو قاب بی‌رنگ هستند. هم‌چنین به‌ازای $K = ۱$ و $\Gamma = ۲$ رئوس $v_{۱۱}$، $v_{۱۲}$، $v_{۲۱}$ و $v_{۲۲}$ همسایه هستند. هم‌چنین دقت شود که رابطه‌ی همسایگی یک رابطه‌ی تعدی نیست.

۳_۳ مدل توجه

مدل توجه استفاده شده در این پروژه، الهام گرفته از مدل‌های [۱۹] و [۲۰] است. استفاده از این روش بر دو ایده‌ی کلی استوار است:

- اگر درست بعد از اتمام لایه‌های پیچشی، خروجی را به‌صورت یک بردار درآورده و به لایه‌ی کاملا متصل بدهیم، پارامترهای شبکه بسیار زیاد شده و یادگیری را مشکل می‌کند. به همین دلیل بهتر است که قبل این کار، به‌گونه‌ای اندازه‌ی خروجی را کاهش دهیم و بعد از آن به یک لایه‌ی کاملا متصل بدهیم.[۲۰]
- روشی که برای کاهش اندازه اتخاذ می‌کنیم بهتر است به‌ازای ورودی‌های مختلف، پاسخ‌های متفاوتی داشته باشد. چرا که همان‌گونه که در فصل پیش به آن اشاره شد، برای برخی از ورودی‌ها احتیاجی به کل داده نیست و بهتر است پارامترها بسته به نوع ورودی، ضرایب متفاوتی داشته باشند.

برای دستیابی به این دو مورد، [۱۹] و [۲۰] لایه‌ی ادغام توجه را معرفی کرده‌اند. بردار ساخته شده توسط این لایه با رابطه‌ی

$$score(X) = Tr(X^T X W) \tag{۳-۴}$$

محاسبه می‌شود.[۱۹]

در رابطه‌ی ۳-۴، X ورودی لایه‌ی ادغام به اندازه‌ی $n \times f$ و W پارامتر این لایه به اندازه‌ی $f \times f$ است. برای این‌که رابطه‌ی بالا قادر به جذب خاصیت توجه باشد، کافی است پارامتر W را به صورت ضرب دو بردار $f \times ۱$ بنویسیم.

$$W = ab^T \rightarrow W^T = ba^T \quad a, b \in R^{f \times ۱} \tag{۵-۳}$$

$$score(X) = Tr(X^T X b a^T) \tag{۶-۳}$$

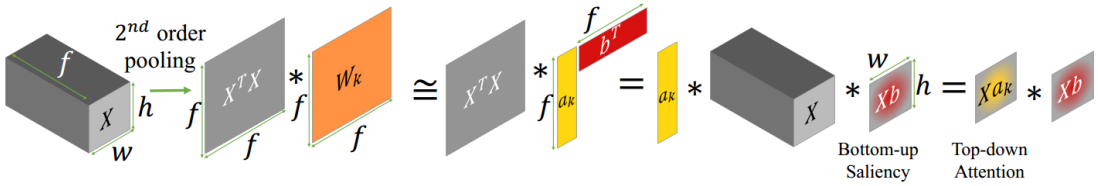
چون برای ماتریس‌ها داریم $Tr(ABC) = Tr(CAB)$، در نتیجه رابطه‌ی ۳-۶ را می‌توان به شکل

$$score(X) = Tr(a^T X^T X b) \tag{۷-۳}$$

نوشت.

هم‌چنین چون برای یک بردار مانند u داریم $Tr(u) = u$ در نتیجه رابطه‌ی ۳-۷ را نیز می‌توان به شکل

$$\begin{aligned} score(X) &= a^T X^T X b \\ &= (Xa)^T (Xb) \end{aligned} \tag{۸-۳}$$



شکل ۳ـ۳: نموداری از مدل توجه با استفاده از لایه‌ی ادغام [۱۹]

نوشت.[۱۹]

حال بردار $score$ آماده است تا به‌عنوان ورودی به لایه‌های کاملا متصل و یا حتی به یک لایه‌ی دسته‌بند بیشینه‌ی هموار داده شود. تصویر ۳ـ۳ جزئیات این مدل را نمایش می‌دهد.

فصل ۴

نتایج تجربی

۴ـ۱ مجموعه داده‌ی مورداستفاده

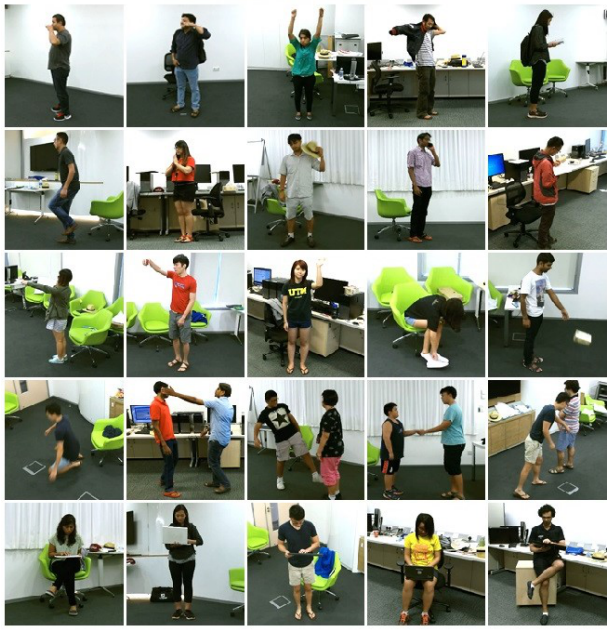
مجموعه داده‌ی NTU-RGB+D بزرگ‌ترین مجموعه داده‌ی شامل اطلاعات ۳بعدی اسکلت بدن است.[۵] تصویر ۴ـ۱ برخی از قاب‌های موجود در این مجموعه داده را نمایش می‌دهد. این مجموعه داده شامل ۵۶۰۰۰ کلیپ و ۶۰ دسته کنش است که از A1 (آشامیدن آب) تا A60 (از هم فاصله‌گرفتن) برچسب[۱] گذاری شده‌اند. این ویدیوها از ۳ زاویه‌ی مختلف و با استفاده از سنجش‌گرهای کینکت [۲] ضبط شده‌اند تا مختصات ۳بعدی مفاصل به دست آیند. این مجموعه داده به‌صورت کلی به دو دسته سنجه[۳] تقسیم شده است.[۱۳]

۱. X-sub: در این سنجه، بازیگران برای مجموعه‌ی آموزش [۴] و مجموعه‌ی آزمون [۵] متفاوت هستند.

۲. X-view: در این سنجه، برای مجموعه‌ی آموزش از زاویه‌های ۱ و ۲ دوربین و برای مجموعه‌ی آزمون از زاویه‌ی ۳ دوربین استفاده شده است.

در این پروژه نیز از هردوی این سنجه‌ها به‌صورت جداگانه استفاده شده است.

[۱] Label
[۲] Kinect Sensors
[۳] Benchmark
[۴] Train Set
[۵] Test Set



شکل ۴ـ۱: برخی از قاب‌های مجموعه داده‌ی NTU RGB+D[۲۱]

۴ـ۲ معیار تابع هزینه‌ی آنتروپی متقابل

در اکثر مسائل دسته‌بندی، از این تابع به‌عنوان معیار درستی تخمین استفاده می‌کنند.[۲۲] [۵] در مسائل دسته‌بندی دودویی این تابع با رابطه‌ی

$$L(y,\hat{y}) = -(y\log(\hat{y}) + (۱-y)\log(۱-\hat{y})) \qquad (۱ـ۴)$$

تعریف می‌شود.[۲۲]

برای مسائل دسته‌بندی که تعداد دسته در آن‌ها از دو بیشتر است، این تابع از رابطه‌ی

$$L(y,\hat{y}) = -\sum_{i} y_i \log \hat{y}_i \qquad (۲ـ۴)$$

محاسبه می‌شود.[۲۲]

در هر دوی روابط ۴ـ۱ و ۴ـ۲، y برچسب واقعی است که دو مقدار صفر یا یک را می‌گیرد و $\hat{y}$ مقدار مشاهده‌شده است که هر مقداری بین صفر تا یک می‌تواند بگیرد.

فصل ۵

جمع‌بندی و راه‌کارهای آتی

در مطالعات آینده، می‌توان اندازه‌ی هسته‌ی استفاده‌شده (چه در حوزه‌ی مکان و چه در زمان) و هم‌چنین تعداد لایه‌های ST-GCN را افزایش داد و تاثیر هر کدام از این تغییرات را بر شبکه‌ی نهایی مشاهده کرد. هم‌چنین می‌توان به‌جای مدل توجه استفاده شده در این پروژه، از مدل‌های دیگری استفاده کرد به‌عنوان مثال با تغییراتی در لایه‌های پیچشی، آن‌ها را حساس به ورودی کرد تا بر روی ورودی‌های مختلف، پردازش‌های متفاوتی صورت گیرد.

مراجع

[1] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles. Advances in human action recognition: A survey. *Dept. of Computer Science and Engineering, University of North Texas*, 2015.

[2] F. Han, B. Reily, W. Hoff, and H. Zhang. Space-Time representation of people based on 3D skeletal data: A review. *Division of Computer Science, Colorado School of Mines, Golden, CO 80401, USA*, 2017.

[3] Activity recognition. https://en.wikipedia.org/wiki/Activity_recognition. Retrieved: 2019-07-29.

[4] A. Ben Tamou, L. Ballihi, and D. ABOUTAJDINE. Automatic learning of articulated skeletons based on mean of 3D joints for efficient action recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 2016.

[5] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[6] T. N. Kipf and M. Welling. Semi-Supervised classification with graph convolutional networks. *ICLR*, 2017.

[7] A. B. Sargano, P. Angelov, and Z. Habib. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Applied Sciences*, 2017.

[8] M. A. Aghbolaghi. A robust and compressed descriptor for action recognition from 4d data". Master's thesis, Sharif University of Technology, 2018.

[9] Dive into deep learning. https://www.d2l.ai/chapter_recurrent-neural-networks/lstm.html. Retrieved: 2019-07-23.

[10] J. Liuy, A. Shahroudyy, D. Xuz, , and G. Wangy. Spatio-Temporal LSTM with trust gates for 3D human action recognition. *School of Electrical and Electronic Engineering, Nanyang Technological University*, 2016.

[11] J. Liu, G. Wang, L. Duan, K. Abdiyeva, and A. C. Kot. Skeleton-Based human action recognition with global Context-Aware attention LSTM networks. *IEEE*.

[12] S. Hochreiter and J. Schmidhuber. Long Short-Term memory. *Neural Computation*, 1997.

[13] A. Shahroudy, J. Liu, T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[14] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. *Signals & Systems*. Prentice-Hall International, Inc., 1997.

[15] S. Kim. Applications of convolution in image processing with MATLAB, 2013.

[16] Convolutional neural network. https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529. Retrieved: 2019-07-23.

[17] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Massachusetts Institute of Technology, 2016.

[18] Graph convolutional networks. https://tkipf.github.io/graph-convolutional-networks/. Retrieved: 2019-07-25.

[19] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. *Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[20] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai. An attention pooling based representation learning method for speech emotion recognition. *Interspeech*, 2018.

[21] Action recognition datasets: "NTU RGB+D" dataset and "NTU RGB+D 120" dataset. http://rose1.ntu.edu.sg/datasets/actionrecognition.asp. Retrieved: 2019-07-23.

[22] Loss functions. https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html. Retrieved: 2019-07-29.

واژه‌نامه

الف

ادغام pooling

اشیا objects

انتها به انتها end to end

انفجار explosion

آشکارسازی detection

آموزش train

ب

بازشناسی recognition

بلادرنگ real time

بیشینه‌ی هموار softmax

پ

پیچش convolution

پیچشی convolutional

ت

توجه attention

ح

حافظه‌ی کوتاه بلندمدت long short-term memory

د

دروازه gate

دسته‌بند classifier

دسته‌بندی classification

ر

رگرسیون خطی linear regression

ز

زمینه context

ژ

ژرفا depth

س

سراسری ..... global

سیستم‌های هوشمند ..... intelligent systems

ش

شبکه‌های عصبی ..... neural networks

شبکه‌های عصبی بازگشتی ..... recurrent neural networks

شبکه‌های پیچشی ..... convolutional networks

شبکه‌های پیچشی گرافی ..... graph convolutional networks

شبکه‌های پیچشی گرافی زمانی‌ـ‌مکانی ..... spatial tempooral graph convolutional networks

ک

کاملا متصل ..... fully connected

کانال ..... channel

کنش ..... action

م

مرتبه ..... itteration

معماری ..... architecture

میانگین ..... average

میرایی ..... vanishing

و

وزن‌های اهمیت مفصل .. edge importance weights

ویژگی ..... feature

ی

یادگیری عمیق ..... deep learning

Abstract

Recently, human action recognition have become one of the most studied topics in the world of computer science and engineering. With the enormous increasing of available datasets and the advent of neural networks, new horizons have been opened to the concept of human action recognition. Representation of human body data is one of the fundamentals of this concept. There are two major methods that's been introduced to optimize this representation. These two methods largely include using RGB-D data and using 3D skeleton information. Nowadays, lots of research have been taken place on representation of 3D skeleton information, due to its flexibility and lower data size.

In this project studies towards representation of 3D skeleton information and its use in human action recognition have been continued. Type of network used in this project is spatial temporal graph convolutional network which is a modification of graph convolutional network which is also a modification of convolutional networks. Also, by introducing a beneficial attention model and using filan criteria, current networks have been improved. Future works could introduce more efficient attention models and study their impact on these networks.

Keywords: Human Action Recognition, 3D Skeleton Data, Graph Convolutional Networks, Attention Model

Sharif University of Technology

Department of Computer Engineering

B.Sc. Thesis

Skeleton Based Human Action Recognition Using Spatial Temporal Graph Convolutional Networks With an Attention Model

By:

Reza Rahimi Azghan

Supervisor:

Prof. Kasaei

August 2019