BỘ CÔNG THƯƠNG

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**NGHIÊN CỨU, XÂY DỰNG HỆ THỐNG NHẬN DIỆN CẢNH BÁO HÀNH VI NGUY HIỂM**

**Nhóm sinh viên thực hiện:**

1. **Nguyễn Nhật Minh**
2. Đỗ Viết Thịnh
3. Phạm Văn Tiến
4. Trần Quang Vinh
5. Vũ Trọng Tuấn

Khoa : Công nghệ thông tin

Người hướng dẫn : Th.S Chu Thị Quyên

**Hà Nội, tháng 5/2022**

BỘ CÔNG THƯƠNG

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**NGHIÊN CỨU, XÂY DỰNG HỆ THỐNG NHẬN DIỆN CẢNH BÁO HÀNH VI NGUY HIỂM**

Nhóm sinh viên thực hiện:

1. **Nguyễn Nhật Minh**
2. Đỗ Viết Thịnh
3. Phạm Văn Tiến
4. Trần Quang Vinh
5. Vũ Trọng Tuấn

Khoa : Công nghệ thông tin

Người hướng dẫn : Th.S Chu Thị Quyên

**Hà Nội, tháng 5/2022**

**MỤC LỤC**

[DANH MỤC BẢNG BIỂU 5](#_Toc103845470)

[DANH MỤC HÌNH ẢNH 6](#_Toc103845471)

[DANH MỤC TỪ VIẾT TẮT 7](#_Toc103845472)

[PHẦN 1: MỞ ĐẦU 8](#_Toc103845473)

[1) Lý do lựa chọn đề tài: 8](#_Toc103845474)

[2) Mục tiêu đề tài: 8](#_Toc103845475)

[3) Phương Pháp Nghiên Cứu: 8](#_Toc103845476)

[4) Đối Tượng Và Phạm Vi Nghiên Cứu: 8](#_Toc103845477)

[PHẦN 2: KẾT QUẢ NGHIÊN CỨU 9](#_Toc103845478)

[CHƯƠNG I: TỔNG QUAN VỀ ĐỀ TÀI 9](#_Toc103845479)

[CHƯƠNG 2: PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN 10](#_Toc103845480)

[2.1) Giới Thiệu Phương Pháp: 10](#_Toc103845481)

[2.2) Các phương pháp Tiền Xử Lý dữ liệu Video thành dạng Khung Xương: 10](#_Toc103845482)

[2.3) Lựa Chọn Kiến Trúc Cho Bài Toán: 12](#_Toc103845483)

[2.4) Mạng Đồ Thị Tích Chập (Graph Convolutional Network): 13](#_Toc103845484)

[2.5) ResGCN: 13](#_Toc103845485)

[2.6) Đa nhánh đầu vào (Multiple Input Branches): 16](#_Toc103845486)

[2.7) Khối chú ý khôn ngoan (Part-wise Attention): 17](#_Toc103845487)

[2.8) Xây dựng mô hình cho bài toán Nhận Dạng: 19](#_Toc103845488)

[CHƯƠNG 3: THỰC NGHIỆM PHƯƠNG PHÁP 21](#_Toc103845489)

[3.1) Phân lớp cho bài toán Nhận Diện Hành Vi Nguy Hiểm 21](#_Toc103845490)

[3.2) Cài đặt tham số huấn luyện mô hình: 27](#_Toc103845491)

[CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 28](#_Toc103845492)

[Danh Mục Tài Liệu Tham Khảo 29](#_Toc103845493)

# DANH MỤC BẢNG BIỂU

Bảng 3.1: Các hành động thuộc nhóm hoạt động hàng ngày

Bảng 3.2: Các hành động thuộc nhóm tình trạng y tế

Bảng 3.3: Các hành động thuộc nhóm tương tác giữa hai người

Bảng 3.4: Các hành động thuộc nhóm hành động nguy hiểm

Bảng 3.5: Các hành động thuộc nhóm hành động có thể nguy hiểm

Bảng 3.6: Các hành động thuộc nhóm hành động an toàn

# DANH MỤC HÌNH ẢNH

Hình 2.1 Phương pháp Thu Thập Chuyển Động Thông Qua Cảm Biến Trên Cơ Thể Người

Hình 2.2: So sánh độ chính xác một số Neural Network

Hình 2.3 Mô tả các lớp Thời Gian (Temporal) và Không Gian (Spatial) trong các khối Basic và Bottleneck của mô hình ResGCN.

Hình 2.4 Mô phỏng việc nhảy block của mạng ResNet

Hình 2.5: Mô phỏng các điểm khớp trên cơ thể người

Hình 2.6: Tổng quan về khối PartAtt được đề xuất

Hình 2.7 Mô tả khung xương con người qua mạng GCN

Hình 2.8: Chi tiết về mô hình ResGCN với cấu trúc nút cổ chai (Bottleneck)

# DANH MỤC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ Viết Tắt** | **Ý Nghĩa** |
| SOTA | State-Of-The-Art: Mô tả các kiến trúc mạng hiện đại nhất |
| NLP | Natural Language Processing: Xử Lý Ngôn Ngữ Tự Nhiên |
| CNN | Convolutional Neural Network: Mạng Nơ-Ron Tích Chập |
| RNN | Recurrent Neural Network: Mạng Nơ-Ron Hồi Quy |
| LSTM | Long Short-Term Memory: Bộ Nhớ Dài Ngắn Hạn |
| DB-LSTM | Deep Bidirectional – Long Short-Term Memory |
| FCL | Fully-Connected Layer: Mạng kết nối đầy đủ |
| GPU | Graphics Processor Units: Bộ xử lý đồ họa |
| GCN | Graph Convolutional Networks: Mạng Tích Chập Đồ Thị |
| MIB | Multiple Input Branches: Đa nhánh đầu vào |
| ResGCN | Residual GCN : Mạng GCN dư |
| PartAtt | Part-wise Attention: Khối Chú ý Khôn Ngoan theo từng phần |
| DGNN | Directed Graph Neural Networks: Mạng Nơ Ron đồ thị có hướng |

# PHẦN 1: MỞ ĐẦU

### 1) Lý do lựa chọn đề tài:

Hành vi, hành động nguy hiểm có thể được coi là những việc làm gây ảnh hưởng đến sức khỏe, tính mạng, tinh thần và của cải vật chất lên con người. Những hành vi này có thể là cố tình (như xảy ra xô xát, v.v) hoặc vô tình (như không để ý con trẻ dẫn đến hậu quả đáng tiếc xảy ra, v.v) . Nhìn chung, hành vi nguy hiểm luôn để lại các hậu quả đáng tiếc xảy ra. Bài toán nhận diện hành vi nguy hiểm xuất hiện cách đây khá lâu và vẫn luôn thu hút được nhiều sự quan tâm, nghiên cứu của các nhà khoa học trên toàn thế giới. Đã xuất hiện nhiều ứng dụng thực hiện nhận diện hành động, hành vi nguy hiểm của con người tùy theo yêu cầu cụ thể, có thể kể ra một số sản phẩm tiêu biểu như: The SWEET-HOME project của Cơ quan Nghiên cứu Quốc gia Pháp,v.v. Việc nghiên cứu, xây dựng các hệ thống nhận diện hành động, hành vi nguy hiểm của con người cũng được bắt đầu từ rất sớm tại Việt Nam. Chúng em lựa chọn đề tài này với mục đích phát triển mô hình trí tuệ nhân tạo giúp hạn chế đi các tình huống đáng tiếc xảy ra trong nhà, trường học, v.v.

### 2) Mục tiêu đề tài:

Mục tiêu đề tài là nghiên cứu, xây dựng hệ thống giúp phát hiện ra các hành động, hành vi nguy hiểm từ đó nghiên cứu và cài đặt các thuật toán trong Nhận Diện Hành Động Con Người. Và trong tương lai, có thể áp dụng vào việc xây dựng các ứng dụng điện thoại, hệ thống cảnh báo tích hợp mô hình học máy.

### 3) Phương Pháp Nghiên Cứu:

Phương pháp nghiên cứu của chúng em là nghiên cứu thực nghiệm, quan sát. Đây là một nhóm các phương pháp gây tác động trực tiếp lên các đối tượng có trong thực tiễn để có thể quan sát, nhìn nhận, đánh giá làm bộc lộ bản chất của như quy luật vận động của đối tượng ,mà chúng em đang nghiên cứu.

### 4) Đối Tượng Và Phạm Vi Nghiên Cứu:

Phạm vi nghiên cứu của chúng em là các công sở, trường học, nhà, v.v. Với đối tượng nghiên cứu là hành động của con người diễn ra trong thường ngày, bao gồm cả tình huống có nguy hiểm và không nguy hiểm.

# PHẦN 2: KẾT QUẢ NGHIÊN CỨU

## CHƯƠNG I: TỔNG QUAN VỀ ĐỀ TÀI

Hành vi, hành động nguy hiểm có thể được coi là những việc làm gây ảnh hưởng đến sức khỏe, tính mạng, tinh thần và của cải vật chất lên con người. Những hành vi này có thể là cố tình (như xảy ra xô xát, v.v) hoặc vô tình (như không để ý con trẻ dẫn đến hậu quả đáng tiếc xảy ra, v.v) . Nhìn chung, hành vi nguy hiểm luôn để lại các hậu quả đáng tiếc xảy ra.Để có thể huấn luyện một mô hình học máy phục vụ cho việc nhận diện hành vi nguy hiểm, bước đệm đầu tiên chính là bài toán Nhận Diện Hành Động Con Người. Hành Động Con Người là một chuỗi các thao tác liên tiếp có thể được mô phỏng bởi một bộ phận trên cơ thể hoặc toàn thân kết hợp.

Bài Toán Nhận Diện Hành Động Con Người (Human Action Recognition) được đưa ra nhằm mục đích nhận diện các hành động và mục tiêu hành động của một hoặc nhiều tác nhân từ một loạt các quan sát về hành động của tác nhân và bối cảnh tác nhân hành động. Nhờ sự phát triển mạnh mẽ của Học Sâu (Deep Learning), cụ thể ở những mảng như: Xử Lý Ngôn Ngữ Tự Nhiên (Natural Language Processing), Thị Giác Máy Tính (Computer Vision) đã có hàng loạt các phương pháp ra đời nhằm giải quyết được bài toán Nhận Diện Hành Động tiêu biểu như : CNN, RNN và LSTM, GCN. Hàng loạt các tập dữ liệu ra đời nhằm phục vụ cho nhu cầu thực nghiệm các bài toán như : UCF-101 chứa 13320 video với 101 thể loại hành động khác nhau, Kinetics-400 chứa ít nhất 400 video mỗi thể loại trong 400 thể loại hành động khác nhau hay NTU RGB+ 60 và NTU RGB+ 120 có 60/120 thể loại hành động khác nhau, video được mô tả thành các điểm khớp (Joint) trên cơ thể.

Dữ liệu cho bài toán Nhận Diện Hành Động Con Người thường được chia thành 2 dạng: dạng Tĩnh và dạng Động. Những hành động dạng tĩnh thường có thể dễ dàng phân loại bởi hành động dạng tĩnh thường cố định, không di chuyển, có thể kể đến một số hành động: Đứng, Nằm, Ngồi, v.v. Những hành động dạng Động như: Đi, Chạy, Nhảy, v.v cần được mô phỏng thông qua một chuỗi các thao tác, dịch chuyển của từng bộ phận cơ thể. Bài toán thực tế yêu cầu các hành động dạng Động bởi một hành động để xác định cần nhận diện được chuỗi hành động liên tiếp thông qua các video.

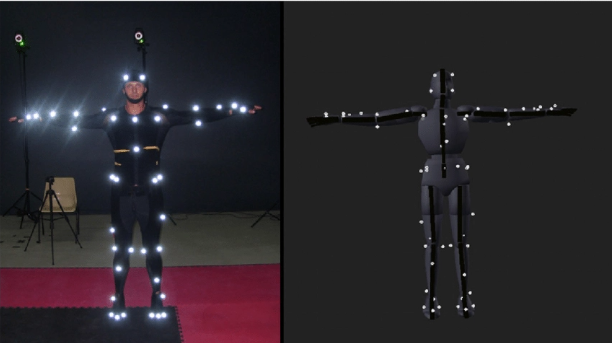
## CHƯƠNG 2: PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN

### 2.1) Giới Thiệu Phương Pháp:

Một vấn đề thiết yếu trong nhận dạng hành động dựa trên Khung Xương Con Người là làm thế nào để trích xuất các đặc điểm riêng biệt trên tất cả các khớp xương của một con người. Tuy nhiên, sự phức tạp của các mô hình State-Of-The-Art (SOTA) càng làm cho nhiệm vụ này có xu hướng phức tạp hóa và mô hình được tham số hóa quá mức, dẫn đến hiệu quả thấp trong việc huấn luyện mô hình, đặc biệt là đối với bộ dữ liệu hành động có quy mô lớn. Vì vậy, chúng em đề xuất thêm một trong phương pháp mới với cơ sở dựa trên Mạng tích chập đồ thị (Graph Convolutional Network), trong đó ba cải tiến chính được tổng hợp, đó là Đa nhánh đầu vào (Multiple Input Branches), ResGCN với kiến trúc nút cổ chai và khối Chú ý từng phần (PartAtt). Thứ nhất, MIB được thiết kế để làm phong phú thêm các đặc trưng về Khung Xương Sau đó, được truyền cảm hứng bởi sự thành công của kiến trúc ResNet trong Convolutional Neural Network (CNN), một mô-đun ResGCN được giới thiệu trong GCN để giảm chi phí tính toán và giảm khó khăn trong việc học tập trong mô hình đào tạo trong khi duy trì độ chính xác của mô hình. Cuối cùng, một khối PartAtt được đề xuất để khám phá các bộ phận cơ thể quan trọng nhất trên toàn bộ chuỗi hành động và có được nhiều đại diện có thể giải thích hơn cho các chuỗi hành động khác nhau của khung xương. Các thử nghiệm mở rộng trên hai bộ dữ liệu quy mô lớn, tức là NTU RGB + D 60 và 120, xác thực rằng đường cơ sở được đề xuất hơi tốt hơn so với các mô hình SOTA khác và trong khi đó yêu cầu ít tham số hơn nhiều trong quá trình huấn luyện, ví dụ, ít hơn DGNN nhiều nhất 34 lần, là một trong những phương pháp SOTA tốt nhất.

### 2.2) Các phương pháp Tiền Xử Lý dữ liệu Video thành dạng Khung Xương:

Dữ liệu Video phục vụ mục đích huấn luyện có nhiều phương pháp để khai thác như: Phương pháp thụ thập chuyển động qua các cảm biến trên cơ thể con người (Motion-Capture) hay Bài Toán Tiên Đoán Tư Thế (Pose Estimation).



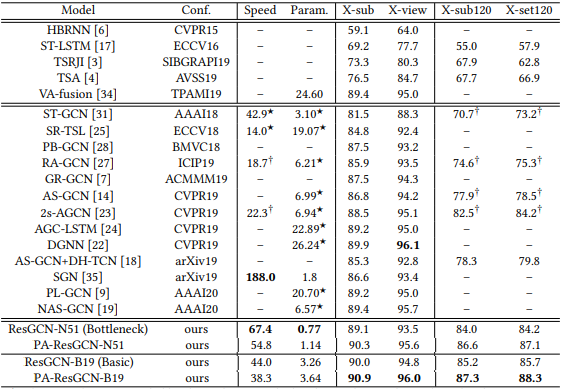
*Hình 2.1 Phương pháp Thu Thập Chuyển Động Thông Qua Cảm Biến Trên Cơ Thể Người (Nguồn: researchgate.com)*

Phương pháp Thu Thập Chuyển Động Thông Qua Cảm Biến Trên Cơ Thể Người có ưu điểm là thu thập dữ liệu chuẩn xác, phù hợp nhất đối với bài toán Nhận Diện Hành Động dựa trên Bộ Xương bởi các cảm biến có thể coi là các “khớp” (Joint) trên Bộ Xương từ đó ta thu được các thông tin về hành động con người dựa trên sự thay đổi về tọa độ của mỗi điểm khớp trên các khung hình, sự thay đổi về góc độ của các “Xương” (biểu diễn bằng cạnh kề trên đồ thị). Phương pháp thường được sử dụng thực tế khi dựng các kỹ xảo phim, game, vì vậy chi phí đầu tư để xây dựng hệ thống Thu Thập Chuyển Động qua cảm biến là không hề nhỏ.

Ta có thể lựa chọn một giải pháp là Bài toán Tiên Đoán Tư Thế (Pose Estimation), đây là một bài toán dự đoán, phân tích tư thế con người trong Video. Với mô hình trên bộ xương, bao gồm một tập hợp các khớp (điểm chính) như mắt cá chân, đầu gối, vai, khuỷu tay, cổ tay và các hướng chi bao gồm cấu trúc xương của cơ thể người. Mô hình này được sử dụng cả trong kỹ thuật ước tính tư thế người 2D và 3D vì tính linh hoạt của nó. Hiện nay có nhiều phương pháp để giải quyết bài toán này, tiêu biểu như VideoPose3D do nhóm Facebook phát triển hay OpenPose do CMU phát triển.

Ta có thể sử dụng camera cảm biến, tiêu biểu như các loại camera Kinect do Microsoft phát triển. Cảm biến Kinect cho phép máy tính thu thập các dữ liệu Video dạng 3D của tác nhân (con người trong video) và môi trường xung quanh. Camera cũng nhận diện và có thể diễn giải chuyển động của con người và chuyển dữ liệu sang định dạng trích ra các thông tin như Tọa Độ mà nhà phát triển có thể sử dụng để xây dựng trải nghiệm mới.

### 2.3) Lựa Chọn Kiến Trúc Cho Bài Toán:



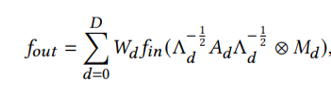
*Hình 2.2: So sánh độ chính xác một số Neural Network*

*(Nguồn: Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition)*

Một số thông tin được chỉ ra tại hình 2.2 cho mô hình có thể được biết như Tốc độ dự đoán là số chuỗi (hay dãy hành động) được mô hình dự đoán thành công trong một giây với một GPU, hay param là số tham số mô hình có thể “huấn luyện” được. Các thông số còn lại mô tả độ chính xác của một số phương pháp. Dựa trên một số kết quả từ hình 2.2, chúng em có nhiều mô hình để lựa chọn cho bài toán nhận diện hành động, có thể kể đến là ST-GCN với tốc độ dự đoán là 42,9 và độ chính xác trên Xsub-120 là 70,7% hoặc ta có thể lưa chọn PA-ResGCN-N51 với tốc đọ dự đoán là 54.8 và độ chính xác trên Xsub-120 là 86,6%, ResGCN-B19 với tốc độ dự đoán là 44.0 và độ chính xác trên Xsub-120 là 85,2%. Trong bài toán này, chúng em chọn mô hình PA-ResGCN-N51 bởi đây là mô hình tối ưu nhất với độ chính xác cao và hiệu quả của tốc độ dự đoán cao. PA là viết tắt của Partwise Attention (khối chú ý từng phần) sẽ được đề cập trong mục sau của báo cáo, N51 có nghĩa là có 51 lớp tích chập hoặc lớp kết nối đầy đủ (Fully Connected Layer) trong Model, ResGCN được mô tả trong mục sau của báo cáo.

### 2.4) Mạng Đồ Thị Tích Chập (Graph Convolutional Network):

Mạng đồ thị tích chập được tính theo công thức sau:



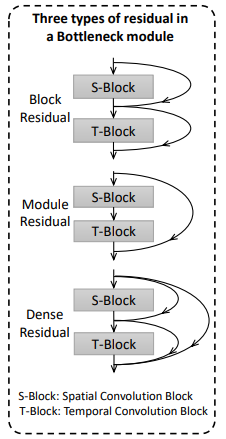
Trong đó 𝐷 là khoảng cách đồ thị tối đa được xác định trước, 𝑓𝑖𝑛 và 𝑓𝑜𝑢𝑡 biểu thị các đầu vào và đầu ra của các Feature Map, ⊗ có nghĩa là phép nhân Element-wise ,𝐴*d* đại diện cho ma trận kề bậc đánh dấu các cặp khớp với khoảng cách đồ thị 𝑑, Λ𝑑 được sử dụng để chuẩn hóa Ad. 𝑊𝑑 và 𝑀𝑑 đều là các tham số được cập nhật được, được sử dụng để thực hiện phép toán tích chập và điều chỉnh mức độ quan trọng của mỗi cạnh tương ứng.

Để trích xuất đặc trưng về sự thay đổi về tọa độ của khớp, góc độ của xương theo thời gian, một lớp tích chập có kích thước 𝐿 × 1 được thiết kế để tổng hợp các đặc trưng theo ngữ cảnh được nhúng trong các khung hình liền kề nhau. Với 𝐿 là một siêu tham số được xác định trước, xác định độ dài của các cửa sổ tạm thời. Cả hai lớp chập không gian và thời gian đều được theo sau bởi một lớp Batch-Norm và một lớp ReLU, và hoàn toàn xây dựng một khối cơ bản.

### 2.5) ResGCN:

Nút thắt cổ chai (Bottleneck) của ResGCN chèn hai lớp chập (convolutional layers) 1 × 1 trước và sau convolution layer chung, để giảm số lượng kênh đặc trưng (feature channels ) với tỷ lệ giảm 𝑟 trong phép tính tích chập. Trong báo cáo này, chúng em thay thế các khối cơ bản không gian và thời gian bằng cấu trúc nút cổ chai, đồng thời có được việc triển khai đào tạo và suy luận mô hình nhanh hơn đáng kể.

Giả sử rằng các kênh đầu vào và đầu ra đều là 256 và tốc độ giảm kênh 𝑟 là 4, kích thước cửa sổ tạm thời 𝐿 là 9. Khi đó, khối cơ bản chứa 256 × 256 × 9 = 589824 tham số, trong khi khối nút cổ chai chỉ chứa 256 × 64 + 64 × 64 × 9 + 64 × 256 = 69632 tham số, ít hơn gần 8,5 lần so với khối cơ bản.



*Hình 2.3 Mô tả các lớp Thời Gian (Temporal) và Không Gian (Spatial) trong các khối Basic và Bottleneck của mô hình ResGCN.*

*(Nguồn: Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition)*

Việc nối tắt giúp mô hình giảm các tham số có thể học được, thông qua việc nối tắt theo 3 cách : Block Residual (nhảy qua theo từng khối), Module Residual (nhảy qua theo từng modul Spatial-Temporal) và Dense Residual (sự kết hợp của 2 cách Block Residual và Module Residual). Việc giảm các tham số có thể học được trong mô hình, giúp cải thiện tốc độ huấn luyện của mô hình, đồng thời chống lại hiện tượng Vanishing Gradient (triệt dốc) gây ảnh hưởng đến việc học ảnh hưởng do đạo hàm thường có xu hướng về 0.

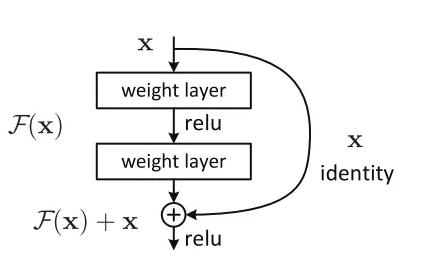
#### 2.5.1) Thuật toán Lan Truyền Ngược và hiện tượng Vanishing Gradient:

Thuật toán lan truyền ngược là một kĩ thuật thường được sử dụng trong trong quá trình huấn luyện mô hình mạng Nơ Ron. Ý tưởng chung của thuật toán là sẽ đi từ output layer đến input layer và tính toán gradient (đạo hàm) của cost function tương ứng cho từng parameter (weight) của network. Gradient Descent, sau đó, sẽ được sử dụng để cập nhật các parameter đó.

Quá trình trên sẽ được lặp lại cho tới khi các parameter của mạng hội tụ. Thông thường chúng ta sẽ có số Epoch (hay số lần mà training set được duyệt qua một lần và weights được cập nhật). Nếu số lượng vòng lặp quá nhỏ, mô hình có thể sẽ không cho ra kết quả tốt, và ngược lại thì thời gian training sẽ quá dài nếu số lượng vòng lặp quá lớn. Ở đây ta có một đánh đổi giữa độ chính xác và thời gian huấn luyện mô hình.

Tuy nhiên trên thực tế gradients (đạo hàm) thường sẽ có giá trị nhỏ dần khi đi xuống các layer thấp hơn(thậm chí về gần giá trị 0). Kết quả là các cập nhật thực hiện bởi Gradient Descent không làm thay đổi nhiều weights của các layer đó, khiến chúng không thể hội tụ và mô hình sẽ không thu được kết quả tốt. Hiện tượng này được gọi là Vanishing Gradients.

#### 2.5.2) Mạng ResNet:



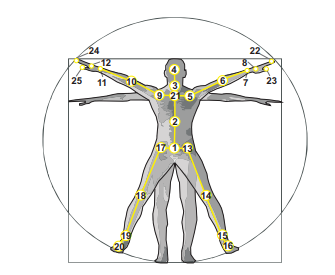
*Hình 2.4 Mô phỏng việc nhảy block của mạng ResNet*

*(Nguồn: Giới thiệu mạng ResNet – Tô Đức Thắng – viblo.asia)*

ResNet gần như tương tự với các mạng CNN khác, đều có convolution, pooling, activation và fully-connected layer. Ảnh bên trên hiển thị khối dư được sử dụng trong mạng. Xuất hiện một mũi tên cong xuất phát từ đầu và kết thúc tại cuối khối dư. Hay nói cách khác là sẽ bổ sung Input X vào đầu ra của layer, hay chính là phép cộng mà ta thấy trong hình minh họa, việc này sẽ chống lại việc đạo hàm bằng 0, do vẫn còn cộng thêm X. Với H(x) là giá trị dự đoán, F(x) là giá trị thật (nhãn), chúng ta muốn H(x) bằng hoặc xấp xỉ F(x). Việc F(x) có được từ x như sau: X->weight1-> ReLU -> weight2. Giá trị H(x) có được bằng cách: F(x) + x -> ReLU. Như chúng ta đã biết việc tăng số lượng các lớp trong mạng làm giảm độ chính xác, nhưng muốn có một kiến trúc mạng sâu hơn có thể hoạt động tốt.

### 2.6) Đa nhánh đầu vào (Multiple Input Branches):

Dữ liệu được mô phỏng thành đồ thị dạng 3 chiều với 25 điểm khớp (25 đỉnh của đồ thị). Cụ thể về nhãn của các điểm khớp và hình sẽ được mô tả như sau:



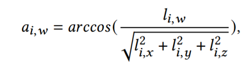
*Hình 2.5: Mô phỏng các điểm khớp trên cơ thể người*

*(Nguồn: researchgate.net )*

Các nhãn của điểm khớp được mô phỏng như sau:

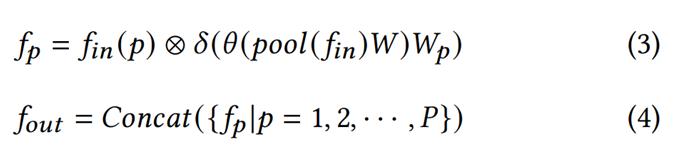
|  |  |  |  |
| --- | --- | --- | --- |
| **Điểm Khớp** | **Nhãn** | **Điểm Khớp** | **Nhãn** |
| 1 | Giữa Hông | 14 | Đầu Gối Phải |
| 2 | Giữa Lưng | 15 | Mắt Cá Chân Phải |
| 3 | Gáy Cổ | 16 | Bàn Chân Phải |
| 4 | Đầu | 17 | Hông Bên Trái |
| 5 | Bả Vai Phải | 18 | Đầu Gối Trái |
| 6 | Khuỷu Tay Phải | 19 | Mắt Cá Chân Trái |
| 7 | Cổ Tay Phải | 20 | Bàn Chân Trái |
| 8 | Bàn Tay Phải | 21 | Giữa vai |
| 9 | Bả Vai Trái | 22 | Cổ Tay Phải |
| 10 | Khuỷu Tay Trái | 23 | Đầu Ngón Tay Phải |
| 11 | Cổ Tay Trái | 24 | Cổ Tay Trái |
| 12 | Bàn Tay Trái | 25 | Đầu Ngón Tay Trái |
| 13 | Hông Bên Phải |  |  |

Dữ liệu được chia thành 3 nhánh: Vị Trí Khớp, Vận Tốc Di Chuyển, Đặc Điểm Xương. Giả sử rằng tập tọa độ 3D ban đầu của một chuỗi hành động là X = {𝑥 ∈ 𝐶 × 𝑇 × 𝑉}, trong đó 𝐶, 𝑇, 𝑉 lần lượt biểu thị tọa độ, khung và khớp. Khi đó tập vị trí tương đối R = {𝑟𝑖 | 𝑖 = 1, 2, · · ·, 𝑉}, trong đó 𝑟𝑖 = 𝑥 [:,:, 𝑖] - 𝑥 [:,:, 𝑐], 𝑥 [:,:, 𝑐] đại diện cho khớp trung tâm của một bộ xương. Hai tập hợp này được nối thành một chuỗi duy nhất và được gửi đến nhánh đầu tiên như là đầu vào của các vị trí khớp. Hơn nữa, dễ dàng có được hai bộ vận tốc chuyển động F = {𝑓𝑡 | 𝑡 = 1, 2, · · ·, 𝑇} và S = {S𝑡 | 𝑡 = 1, 2, · · ·, 𝑇} với các định nghĩa sau: 𝑓𝑡 = 𝑥 [:, 𝑡 +2,:] −𝑥 [:, 𝑡,:] và S*t* = 𝑥 [:, 𝑡 + 1,:] - 𝑥 [:, 𝑡,:]. Và đầu vào của vận tốc chuyển động có được bằng cách ghép F và S cho mỗi khớp để thu được vectơ đặc trưng 6-d tại mỗi thời điểm. Cuối cùng, đầu vào của các đặc điểm xương bao gồm chiều dài xương L = {𝑙𝑖 | 𝑖 = 1, 2, · · ·, 𝑉} và các góc xương A = {𝑎𝑖 | 𝑖 = 1, 2, · · ·, 𝑉} . Để có được hai bộ này, độ dài mỗi xương được tính bằng 𝑙𝑖 = 𝑥 [:,:, 𝑖] - 𝑥 [:,:,i𝑎𝑑𝑗], trong đó 𝑖𝑎𝑑𝑗 có nghĩa là khớp liền kề của khớp thứ i. Tiếp theo, góc của mỗi xương được tính bằng công thức dưới đây trong đó 𝑤 ∈ {𝑥, 𝑦, 𝑧} biểu thị các tọa độ 3D.

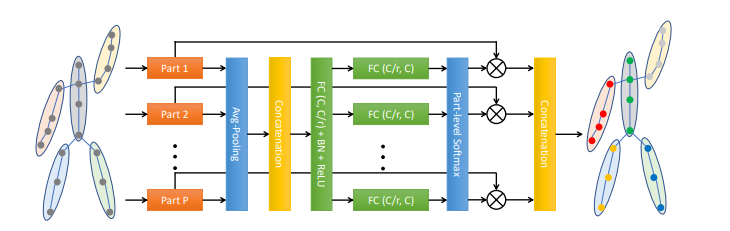


### 2.7) Khối chú ý khôn ngoan (Part-wise Attention):

Mục đích trích xuất đặc trưng từ chuyển động của các bộ phận cơ thể riêng lẻ theo thời gian. Các đặc trưng của tất cả các bộ phận được nối và gộp trung bình (Average Pooling) theo thứ nguyên thời gian, rồi chuyển qua một lớp được kết nối đầy đủ với một lớp BatchNorm và một hàm kích hoạt ReLU. Sau đó, năm lớp được kết nối đầy đủ được thông qua để tính toán ma trận chú ý và hàm Softmax được sử dụng để xác định các bộ phận cơ thể quan trọng nhất. Cuối cùng, các tính năng của năm phần được nối với nhau như một biểu diễn bộ xương tích hợp với các trọng số chú ý khác nhau. Khối PartAtt này có thể được xây dựng dưới dạng:



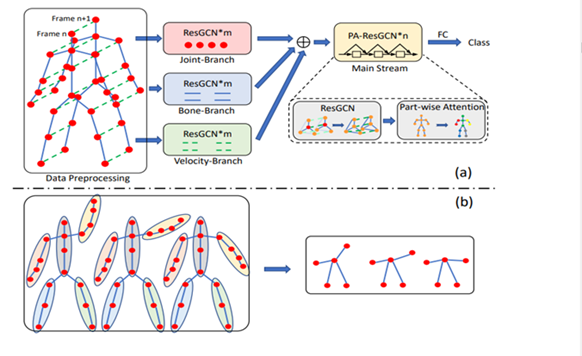
Trong đó 𝑓𝑖𝑛 và 𝑓𝑜𝑢𝑡 biểu thị bản đồ tính năng đầu vào và đầu ra, ⊗ có nghĩa là phép nhân khôn ngoan phần tử, 𝑝𝑜𝑜𝑙 (·) biểu thị hoạt động nhóm trung bình theo thời gian và nhóm gộp, 𝛿 (·) và 𝜃 (·) đại diện cho kích hoạt Softmax và ReLU cấp một phần chức năng. Và 𝑊 ∈ R 𝐶 × 𝐶/𝑟, 𝑊𝑝 ∈ R 𝐶 /𝑟 × 𝐶 đều là các tham số có thể học được, trong đó 𝑊 được dùng chung cho tất cả các phần để giảm kích thước và 𝑊𝑝 là cụ thể cho từng phần để tính trọng số chú ý cuối cùng.



*Hình 2.6: Tổng quan về khối PartAtt được đề xuất, trong đó 𝐶 biểu thị số lượng kênh đầu vào, 𝑟 = 4 được sử dụng để giảm chi phí tính toán, 𝑃 = 5 đại diện cho năm bộ phận cơ thể riêng lẻ, ⊗ có nghĩa là phép nhân thông minh phần tử và cấp Phần Softmax có nghĩa là tính toán Softmax ở cấp bộ phận.*

*(Nguồn: Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition)*

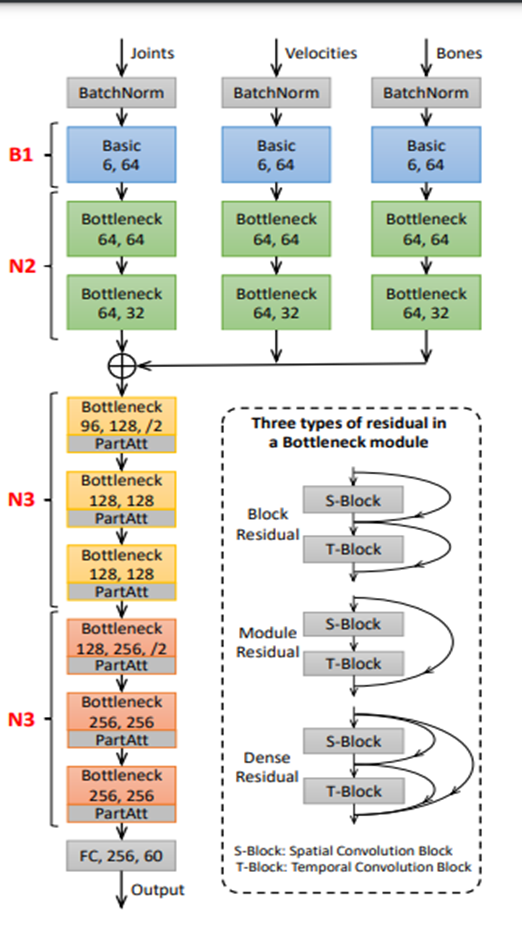
### 2.8) Xây dựng mô hình cho bài toán Nhận Dạng:



*Hình 2.7 Mô tả khung xương con người qua mạng GCN*

*(Nguồn: Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition)*

Dựa trên hình 2.7, Ba nhánh thông tin đầu vào (Khớp, Vận tốc và Xương) ban đầu được trích xuất từ trình tự khung xương ban đầu. Tiếp theo, mỗi chuỗi đầu vào được gửi đến một nhánh đầu vào bao gồm một số mô-đun ResGCN. Sau đó, ba nhánh sẽ được nối và gửi đến một số mô-đun PartAtt -ResGCN, trong đó mỗi mô-đun PartAtt -ResGCN sẽ thực thi tuần tự từ mô-đun ResGCN đến một khối PartAtt. Chi tiết về cài đặt mô hình sẽ được mô tả bởi hình sau đây:



Hình 2.8: Chi tiết về mô hình ResGCN với cấu trúc nút cổ chai (Bottleneck). Các thông số cấu trúc của ví dụ này là [B1, N2, N3, N3] tương ứng với loại và số lượng mô-đun trong các bộ phận

*(Nguồn: Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition)*

Từ video đầu vào, ta trích xuất ra 3 thông tin của video bao gồm: Joint-Branch (khớp nối), Bone-Branch (khung xương) và Velocity-Branch(vận tốc). Các thông tin sẽ được xử lý Batch-Norm và đưa vào khối B1, B1 biểu thị một mô-đun ResGCN với các khối (Block) cơ bản. Tiếp theo thông tin sẽ đi qua khối N2 bao gồm 2 lớp Bottleneck (Nút thắt cổ chai) nhân hội tụ thông qua phép nhân khôn ngoan phần tử (Element-wise). Tiếp theo, sẽ lần lượt đi qua 2 khối N3 bao gồm 3 lớp Bottleneck kèm theo PartAtt kèm theo sau. Sau đó để thu kết quả của Video, ta xử lý lớp kết nối đầy đủ (FCL) để đưa ra đầu ra hành động cho bài toán.

## CHƯƠNG 3: THỰC NGHIỆM PHƯƠNG PHÁP

3.1) Phân lớp cho bài toán Nhận Diện Hành Vi Nguy Hiểm**:**

Chúng em thực nghiệm bài toán trên tập dữ liệu NTU RGB+ D và NTU RGB+ D 120. Bộ data “NTU RGB + D” chứa 60 lớp hành động và 56.880 mẫu video. Bộ data “NTU RGB + D 120” là một bộ mở rộng cảu bộ “NTU RGB + D” bằng cách bổ sung thêm 60 lớp hành động và 57.600 mẫu video khác. Vậy “NTU RGB + D 120 ”có tổng cộng 120 lớp hành động và 114.480 mẫu video. Hai bộ dữ liệu này đều chứa video RGB, depth map sequences, dữ liệu khung 3D và video hồng ngoại (IR) cho mỗi mẫu. Mỗi tập dữ liệu được chụp đồng thời bởi ba camera Kinect V2. Độ phân giải của video RGB là 1920x1080, bản đồ độ sâu và video IR đều có kích thước 512x424 và dữ liệu khung 3D chứa tọa độ 3D của 25 khớp cơ thể tại mỗi khung hình.

Các hành động trong hai tập dữ liệu này thuộc ba danh mục chính: hành động hàng ngày, hành động tương hỗ và tình trạng y tế, như được hiển thị trong bảng bên dưới. Trong đó bộ “NTU RGB + D ” gồm các hành động được gán nhãn từ A1 đến A60. Các hành động A1 đến A120 đều nằm trong bộ “NTU RGB + D 120”:

***\*) Nhóm hành động hàng ngày:***

|  |  |  |  |
| --- | --- | --- | --- |
| A1: uống nước | A2: ăn bữa ăn | A3: đánh răng | A4: chải tóc |
| A5: thả | A6: nhận | A7: ném | A8: ngồi xuống |
| A9: đứng lên | A10: vỗ tay | A11: đọc | A12: đang viết |
| A13: xé giấy | A14: mặc áo khoác vào | A15: cởi áo khoác | A16: xỏ giày vào |
| A17: cởi giày | A18: đeo kính vào | A19: bỏ kính | A20: đội mũ / mũ lưỡi trai |
| A21: cởi mũ / mũ lưỡi trai | A22: vui lên | A23: vẫy tay | A24: đá cái gì đó |
| A25: đưa tay vào túi | A26: nhảy | A27: nhảy lên | A28: cuộc gọi điện thoại |
| A29: chơi với điện thoại / máy tính bảng | A30: gõ trên bàn phím | A31: chỉ vào một cái gì đó | A32: chụp ảnh tự sướng |
| A33: kiểm tra thời gian (từ đồng hồ) | A34: xoa hai tay | A35: gật đầu / cúi đầu | A36: lắc đầu |
| A37: lau mặt | A38: chào | A39: đặt hai lòng bàn tay vào nhau | A40: bắt chéo tay ở phía trước |
| A61: đeo tai nghe | A62: tháo tai nghe | A63: bắn vào rổ | A64: bóng nảy |
| A65: swing gậy tennis | A66: bóng bàn tung hứng | A67: im lặng | A68: vuốt tóc |
| A69: giơ ngón tay cái lên | A70: ngón tay cái xuống | A71: bàn tay OK | A72: tạo dấu hiệu chiến thắng |
| A73: gập sách | A74: đếm tiền | A75: cắt móng tay | A76: cắt giấy |
| A77: búng tay | A78: mở chai | A79: ngửi | A80: ngồi xổm xuống |
| A81: tung đồng xu | A82: gấp giấy | A83: bóng lên giấy | A84: chơi khối lập phương ma thuật |
| A85: thoa kem lên mặt | A86: thoa kem lên tay | A87: đeo vào túi | A88: cởi túi |
| A89: bỏ đồ vật vào túi | A90: Lấy đồ vật ra khỏi túi | A91: mở hộp | A92: di chuyển vật nặng |
| A93: bắt tay | A94: Ném mũ / nón lên | A95: đầu hàng | A96: khoanh tay |
| A97: vòng tròn cánh tay | A98: vung tay | A99: chạy tại chỗ | A100: đá mông |
| A101: chạm chéo ngón chân | A102: đá bên |  |  |

*Bảng 3.1: Các hành động thuộc nhóm hoạt động hàng ngày*

***\*) Nhóm tình trạng y tế:***

|  |  |  |  |
| --- | --- | --- | --- |
| A41: hắt hơi / ho | A42: đáng kinh ngạc | A43: rơi xuống | A44: đau đầu |
| A45: đau ngực | A46: đau lưng | A47: đau cổ | A48: buồn nôn / nôn |
| A49: người hâm mộ tự | A103: ngáp | A104: vươn mình | A105: xì mũi |

*Bảng 3.2: Các hành động thuộc nhóm tình trạng y tế*

***\*) Nhóm hành động tương tác giữa hai người:***

|  |  |  |  |
| --- | --- | --- | --- |
| A50: cú đấm / cái tát | A51: đá | A52: đẩy | A53: vỗ nhẹ vào lưng |
| A54: ngón tay trỏ | A55: ôm | A56: cho đối tượng | A57: túi cảm ứng |
| A58: bắt tay | A59: đi về phía | A60: cách xa nhau | A106: đánh vật thể |
| A107: dao cầm | A108: xô ngã | A109: lấy đồ | A110: bắn bằng súng |
| A111: bước chân lên | A112: đập tay | A113: cổ vũ và uống | A114: mang đối tượng |
| A115: chụp ảnh | A116: theo dõi | A117: thì thầm | A118: Trao đổi mọi thứ |
| A119: hỗ trợ ai đó | A120: oẳn tù tì | - | - |

*Bảng 3.3: Các hành động thuộc nhóm tương tác giữa hai người*

Trong đề tài này chúng em chọn bộ “NTU RGB + D 120” làm bộ data chính.Trong đó để phù hợp với bài toán nhận diện hành động nguy hiểm chúng em chia lại bộ data trên thành 3 nhóm hành động chính mới là: nhóm hành động nguy hiêm, nhóm hành động có thể nguy hiểm và nhóm hành động an toàn.

***\*) Nhóm hành động nguy hiểm:***

Các hành động trong nhóm này là những hành động có tính nguy hiểm cao cần được đưa ra thông báo ngay cho người dùng, thậm chí có thể là cảnh báo ở cường dộ cao.

|  |  |  |  |
| --- | --- | --- | --- |
| A43: rơi xuống | A45: đau ngực | A52: đẩy | A116: theo dõi |
| A114:mang đối tượng | A107:cầm dao | A108: xô ngã | A109: lấy đồ |
| A110: bắn súng |  |  |  |

*Bảng 3.4: Các hành động thuộc nhóm hành động nguy hiểm*

***\*) Nhóm hành động có thể nguy hiểm:***

Các hành động trong nhóm này có thể sẽ gây nguy hiểm tùy thuộc vào ngoại cảnh của hành động đó. Ví dụ khi ném các đồ vật bình thường như giấy bút là bình thường nhưng khi các vật bị ném là bóng đèn, đất đá có thể gây nguy hiểm nhất khi đối tượng ném là trẻ con.

|  |  |  |  |
| --- | --- | --- | --- |
| A5: thả | A7: ném | A24: đá cái gì đó | A92:di chuyển vật nặng |
| A95:đầu hàng | A41:hắt hơi/ho | A44 đau đầu | A46:đau lưng |
| A47: đau cổ | A48: buồn nôn/nôn | A50:đấm/tát | A51:đá |
| A106:đánh vật thể | A42: kinh ngạc | A98:vung tay |  |

*Bảng 3.5: Các hành động thuộc nhóm hành động có thể nguy hiểm*

***\*) Nhóm hành động an toàn:***

Các hành động trong nhóm này đươc xem là những hoạt đông an toàn và gây ra nguy hiểm gì

|  |  |  |  |
| --- | --- | --- | --- |
| A1: uống nước | A2: ăn bữa ăn | A3: đánh răng | A4: chải tóc |
| A49: fan self | A6: nhận | A103: ngáp | A8: ngồi xuống |
| A9: đứng lên | A10: vỗ tay | A11: đọc | A12: đang viết |
| A13: xé giấy | A14: mặc áo khoác vào | A15: cởi áo khoác | A16: xỏ giày vào |
| A17: cởi giày | A18: đeo kính vào | A19: bỏ kính | A20: đội mũ / mũ lưỡi trai |
| A21: cởi mũ / mũ lưỡi trai | A22: vui lên | A23: vẫy tay | A104: vươn mình |
| A25: đưa tay vào túi | A26: nhảy | A27: nhảy lên | A28: cuộc gọi điện thoại |
| A29: chơi với điện thoại / máy tính bảng | A30: gõ trên bàn phím | A31: chỉ vào một cái gì đó | A32: chụp ảnh tự sướng |
| A33: kiểm tra thời gian (từ đồng hồ) | A34: xoa hai tay | A35: gật đầu / cúi đầu | A36: lắc đầu |
| A37: lau mặt | A38: chào | A39: đặt hai lòng bàn tay vào nhau | A40: bắt chéo tay ở phía trước |
| A61: đeo tai nghe | A62: tháo tai nghe | A63: bắn vào rổ | A64: bóng nảy |
| A65: swing gậy tennis | A66: bóng bàn tung hứng | A67: im lặng | A68: vuốt tóc |
| A69: giơ ngón tay cái lên | A70: ngón tay cái xuống | A71: ký OK | A72: tạo dấu hiệu chiến thắng |
| A73: gập sách | A74: đếm tiền | A75: cắt móng tay | A76: cắt giấy |
| A77: búng tay | A78: mở chai | A79: ngửi | A80: ngồi xổm xuống |
| A81: tung đồng xu | A82: gấp giấy | A83: bóng lên giấy | A84: chơi khối lập phương ma thuật |
| A85: thoa kem lên mặt | A86: thoa kem lên tay | A87: đeo vào túi | A88: cởi túi |
| A89: bỏ đồ vật vào túi | A90: Lấy đồ vật ra khỏi túi | A91: mở hộp | A53: vỗ nhẹ vào lưng |
| A93: bắt tay | A94: Ném mũ / nón lên | A54: ngón tay trỏ | A96: khoanh tay |
| A97: vòng tròn cánh tay | A55: ôm | A99: chạy tại chỗ | A100: đá mông |
| A101: chạm chéo ngón chân | A102: đá bên | A105:xì mũi | A56: cho đối tượng |
| A57: túi cảm ứng | A58: bắt tay | A59: đi về phía | A60: cách xa nhau |
| A111: bước chân lên | A112: high-five | A113: cổ vũ và uống | A115: chụp ảnh |
| A117: thì thầm | A118: Trao đổi mọi thứ | A119: hỗ trợ ai đó | A120: oẳn tù tì |

*Bảng 3.6: Các hành động thuộc nhóm hành động an toàn*

### 3.2) Cài đặt tham số huấn luyện mô hình:

Trong các thử nghiệm của chúng em, khoảng cách đồ thị lớn nhất 𝐷 và kích thước cửa sổ tạm thời 𝐿 được đề cập trong Phần 2.3 được đặt lần lượt là 2 và 9. Số lượng epochs tối đa là 70. Tốc độ học tập ban đầu (Learning Rate) được đặt thành 0,1 và dropout ở các epochs thứ 20 và thứ 50. Hơn nữa, chiến lược tối ưu được sử dụng ở 10 epochs đầu tiên để làm cho quy trình huấn luyện ổn định hơn. Xuống dốc ngẫu nhiên (Stohastic Gradient Descent) được đặt làm thuật toán tối ưu hóa mô hình với động lượng Nesterov là 0,9 và phân rã trọng lượng (Weight) là 0,0001 được sử dụng để điều chỉnh các tham số. Các thông số cấu trúc khác được xác định như Hình 2.15.

## CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Báo cáo đã tìm hiểu được phương pháp hiện đại trong bài toán Nhận Diện Hành Vi Nguy Hiểm dựa trên phương pháp Nhận dạng hành động dựa trên Khung Xương Con Người sử dụng mạng GCN. Đã mô tả, phân biệt dữ liệu hành động dạng động và dạng tĩnh, từ đó đưa ra loại dữ liệu phù hợp cho bài toán. Đồng thời trong các phương pháp đã đề xuất cải thiện độ chính xác cho mô hình trong việc nhận diện. Kết quả đã đáp ứng được các nhu cầu đề ra. Với bộ dữ liệu lớn, ta hoàn toàn có thể mang mô hình ra áp dụng vào các ứng dụng thực tế.

Tuy nhiên,vẫn còn một số vấn đề tồn đọng chưa được giải quyết, việc hạn chế khá nhiều về mặt phần cứng phục vụ cho quá trình huấn luyện mô hình, cũng như không có đủ trang thiết bị cần thiết phục vụ cho quá trình triển khai thực tế bài toán càng làm cho quá trình ứng dụng thực tế bài toán trở nên khó khăn hơn.

Trong tương lai, chúng em dự kiến sẽ áp dụng vào các ứng dụng điện thoại, triển khai một hệ thống cảnh báo dựa trên mô hình đã được nghiên cứu để đưa ra cảnh báo cho người dùng.

# Danh Mục Tài Liệu Tham Khảo

**Tiếng Việt**

1. Phạm Minh Hoàng, “Giới thiệu sơ qua bài toán nhận diện hành động của người trong video”, 2021.
2. Báo Tuổi Trẻ, 'Người hùng' Nguyễn Ngọc Mạnh: Lan tỏa nghĩa cử cao đẹp, 2021
3. Trần Thanh Tuấn, “Nhận diện hành động con người trong Video sử dụng Deep Learning”, 2017
4. Nguyễn Tuấn, “Transfer learning và data augmentation”, Deep Learning Cơ Bản, 2019
5. Vinh Nguyen , “Vanishing & Exploding Gradients Problems in Deep Neural Networks (Part 1)”, 2017
6. Tô Đức Thắng, “Giới thiệu mạng ResNet”,2020

**Tiếng Anh**

1. Yi-Fan Song, Caifeng Shan, Zhang Zhang, Liang Wang,“Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition”, 2010
2. Sijie Yan, Yuanjun Xiong, Dahua Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition”, 2018
3. Yi Zhu, Xinyu Li và các đồng tác giả, “A Comprehensive Study of Deep Video Action Recognition”, 2020.
4. Wenjug Zang, “Microsoft Kinect Sensor and Its Effect”, 2021
5. Jiuxiang Gu, Zhenhua Wang và các đồng tác giả, “Recent Advances in Convolutional Neural Networks”
6. Will Kay, Joao Carreira, và các đồng tác giả, “The Kinetics Human Action Video Dataset”, 2017
7. Taha Anwar, “Introduction to Video Classification and Human Activity Recognition”, 2021
8. “3D Pose Estimation”, Wikipedia, 2021
9. “Graph Theory”, Wikipedia, 2021