**HỌC VIỆN KỸ THUẬT QUÂN SỰ**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO THỰC TẬP CÔNG NGHỆ THÔNG TIN**

NGÀNH KHOA HỌC MÁY TÍNH

**ĐỀ TÀI: ĐẾM SỐ LẦN LẶP LẠI HÀNH ĐỘNG SỬ DỤNG MÔ HÌNH MẠNG REPNET**

Sinh viên thực hiện: Nguyễn Việt Đức

Mã sinh viên: 18150008

Giáo viên hướng dẫn: Phan Thị Hải Hồng

Hà Nội, 2022

**Lời nói đầu**

**\*\*\***

Đây là phương pháp để ước tính khoảng thời gian mà một hành động cũng như đếm số lần được lặp lại trong video. Điểm mấu chốt của phương pháp này nằm ở việc dự đoán chu kỳ bằng cách sử dụng tính tương tự theo thời gian như một nút thắt biểu diễn trung gian cho việc tổng quát hóa lặp lại trong các video tự nhiên. Mô hình này được gọi là RepNet, với tập dữ liệu tổng hợp được tạo từ bộ sưu tập video lớn không gắn nhãn bằng cách lấy mẫu các clip ngắn có độ dài khác nhau và lặp lại với các khoảng thời gian và số lượng khác nhau. Sự kết hợp giữa dữ liệu tổng hợp và một mô hình mạnh mẽ cho phép dự đoán các chu kỳ bất khả tri (chưa biết rõ).

Bài viết này có sự tham khảo phần lớn từ bài báo gốc. Link bài báo gốc: https://arxiv.org/pdf/2006.15418.pdf

Mục lục

[1. Giới thiệu 3](#_Toc107062158)

[2. Ứng dụng tiêu biểu 6](#_Toc107062159)

[3. Mô hình mạng Repnet 6](#_Toc107062160)

[3.1. Bộ mã hóa 7](#_Toc107062161)

[3.2. Ma trận tự tương tự theo thời gian (TSM-Temporal Self similarity Matrix) 7](#_Toc107062162)

[3.3. Dự đoán chu kỳ 8](#_Toc107062163)

[3.4. Hàm Loss 9](#_Toc107062164)

[3.5. Suy luận 9](#_Toc107062165)

[4. Huấn luyện mô hình với các lần lặp lại tổng hợp 9](#_Toc107062166)

[4.1. Tổng hợp các video lặp lại 10](#_Toc107062167)

[4.2. Tăng cường chuyển động camera 11](#_Toc107062168)

[5. Bộ dữ liệu Countix 12](#_Toc107062169)

[6. Đánh giá 13](#_Toc107062170)

[7. Xây dựng ứng dụng web bằng Streamlit 15](#_Toc107062171)

[8. Tổng kết 19](#_Toc107062172)

# **1. Giới thiệu**

Hình dung những cảnh đời thường nhất - một người đang ăn trong quán cà phê. Họ có thể khuấy đường trong cà phê trong khi nhai thức ăn và gõ chân theo nhạc. Người này đang thực hiện song song ít nhất ba hoạt động định kỳ. Các hành động và quy trình lặp đi lặp lại rất phổ biến trong cuộc sống hàng ngày của chúng ta. Chúng bao gồm các chu kỳ hữu cơ, chẳng hạn như nhịp tim và nhịp thở, đến các chu kỳ hành tinh như chu kỳ ngày-đêm và các mùa. Do đó, nhu cầu nhận dạng các lần lặp lại trong video là phổ biến và một hệ thống có thể xác định và đếm số lần lặp lại trong video sẽ mang lại lợi ích cho bất kỳ hệ thống tri giác nào nhằm mục đích quan sát và hiểu thế giới của chúng ta trong một khoảng thời gian dài.

Sự lặp lại cũng thú vị vì những lý do sau:

(1) Thường có một ý định hoặc một nguyên nhân thúc đẩy đằng sau một cái gì đó xảy ra nhiều lần.

(2) Cùng một sự kiện có thể được quan sát lại nhưng có sự khác biệt nhỏ.

(3) Có thể có những thay đổi dần dần trong cảnh do kết quả của những lần lặp lại này;

(4) Chúng cung cấp cho chúng ta các đơn vị hành động rõ ràng, một chuỗi con trong hành động có thể được phân đoạn kịp thời (ví dụ: nếu bạn đang cắt một củ hành, đơn vị hành động là hành động thao tác được lặp lại để tạo ra các lát bổ sung).

Do những lý do trên, bất kỳ tác nhân nào tương tác với thế giới sẽ được hưởng lợi rất nhiều từ một hệ thống như vậy. Hơn nữa, việc đếm số lần lặp lại là thích hợp cho nhiều ứng dụng thị giác máy tính; chẳng hạn như đếm số lần một bài tập thể dục đã được thực hiện, đo lường các sự kiện sinh học (như nhịp tim), v.v.

Tuy nhiên, nghiên cứu về video định kỳ còn hạn chế, có thể do thiếu tập dữ liệu lặp lại video được gắn nhãn quy mô lớn. Ngược lại, để nhận dạng hành động, có những bộ dữ liệu quy mô lớn, như Kinetics, nhưng bộ dữ liệu này ở quy mô lớn, hữu dụng nhờ sự sẵn có của các từ khóa/văn bản được liên kết với video. Thật không may, hiếm khi video được gắn nhãn chú thích liên quan đến hoạt động lặp lại vì văn bản có nhiều khả năng mô tả nội dung ngữ nghĩa hơn. Vì lý do này, nên sử dụng tập dữ liệu với các nhãn hành động ngữ nghĩa thường được sử dụng để nhận dạng hành động (Kinetics) và chọn thủ công video về các lớp đó có chuyển động tuần hoàn (bóng nảy, vỗ tay, v.v.). Bước tiếp theo tiến hành gắn nhãn các video đã chọn với số lần lặp lại có trong mỗi clip.

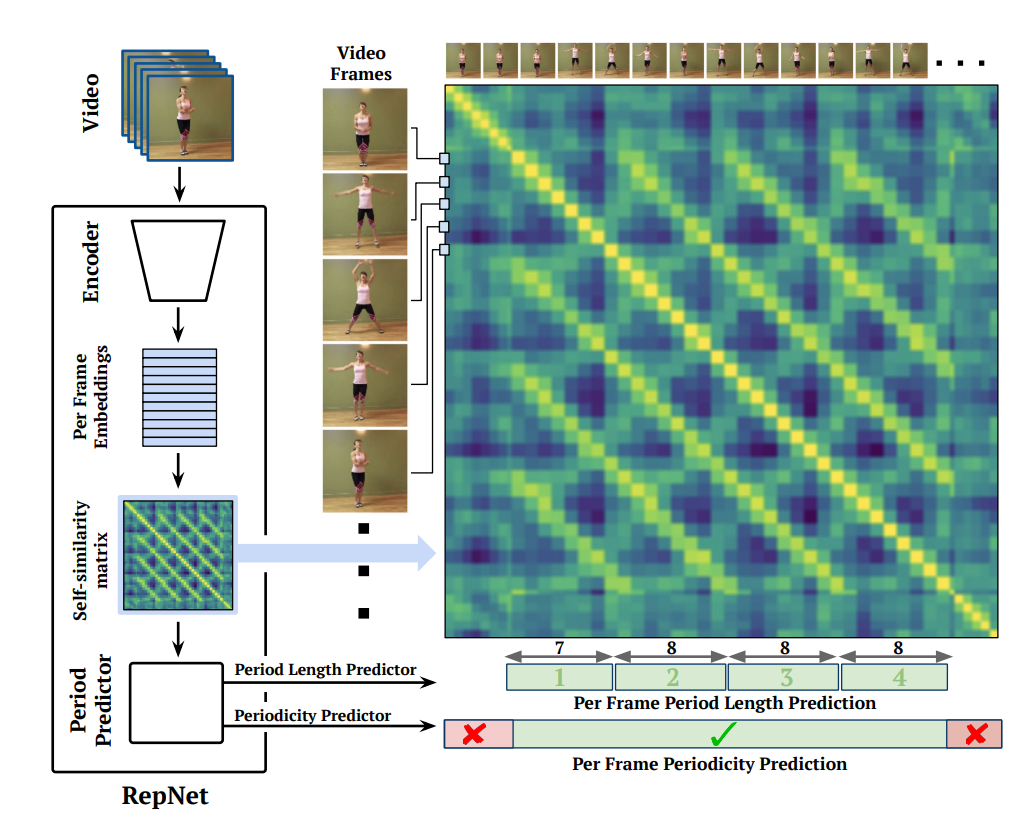
Việc gắn nhãn thủ công giới hạn số lượng video có thể được chú thích - việc gắn nhãn rất tẻ nhạt và tốn kém do tính chất chi tiết về thời gian của nhiệm vụ. Để giảm bớt lượng dữ liệu đào tạo, đề xuất một phương pháp tạo video lặp lại tổng hợp bằng cách lặp lại các clip từ các video hiện có với các khoảng thời gian khác nhau. Vì chúng ta đang tổng hợp định cỡ các video này nên cũng có chú thích chính xác cho khoảng thời gian và số lần lặp lại trong video, có thể được sử dụng cho các mô hình đào tạo sử dụng phương pháp học có giám sát. Như đã thấy trong công việc của mình, những video tổng hợp như vậy không thể nắm bắt được tất cả các sắc thái của các video lặp lại thực sự và dễ bị các mô hình học sâu dung lượng cao làm cho quá khớp. Để giải quyết vấn đề này, đề xuất chiến lược tăng dữ liệu cho video tổng hợp để các mô hình được đào tạo về chúng chuyển sang video thực có lặp lại. Sử dụng kết hợp dữ liệu thực và dữ liệu tổng hợp để phát triển mô hình.

Mục tiêu là một mô hình duy nhất hoạt động cho nhiều lớp video định kỳ và trên thực tế, cho cả các lớp video không được nhìn thấy trong quá trình đào tạo. Để đạt được điều này sử dụng một đại diện trung gian khuyến khích tạo thành các lớp không nhìn thấy. Ma trận tự tương tự về thời gian- được sử dụng để dự đoán khoảng thời gian mà một hành động được lặp lại trong video. Biểu diễn chung này được sử dụng trên các loại video lặp lại khác nhau cho phép tổng quát hóa mong muốn. Ví dụ: cho dù một người đang chống đẩy hay một đứa trẻ đang đu trong sân chơi, thì ma trận tự tương tự là phép tham số hóa được chia sẻ từ đó suy ra số lần lặp lại. Nút thắt cổ chai cực độ này (số lượng kênh trong bản đồ tính năng giảm từ 512 xuống còn 1) cũng hỗ trợ việc tổng hợp hóa từ dữ liệu tổng hợp thành dữ liệu thực. Ưu điểm khác của cách biểu diễn này là khả năng diễn giải của mô hình được đưa vào kiến ​​trúc mạng khi chúng ta buộc mạng chỉ dự đoán khoảng thời gian từ ma trận tự tương tự, trái ngược với việc suy ra khoảng thời gian từ các đặc trưng chiều cao tiềm ẩn.

Tập trung vào hai nhiệm vụ:

(i) Đếm số lần lặp lại và xác định số lần lặp lại trong video. Diễn đạt lại vấn đề này như ước tính độ dài trên mỗi chu kỳ khung hình, sau đó chuyển đổi chúng thành số lần lặp lại;

(ii) Dự đoán chu kỳ, xác định xem khung hiện tại có phải là một phần của mô hình thời gian lặp lại hay không. Tiếp cận điều này như một vấn đề phân loại nhị phân trên mỗi khung. Giải thích trực quan về các nhiệm vụ này và tổng quan về giải pháp được hiển thị trong Hình 1.



*Hình 1: RepNet sử dụng ma trận tự tương tự thời gian làm lớp trung gian để dự đoán độ dài chu kỳ và tính chu kỳ của mỗi khung hình trong video.*

Những ý chính của bài viết:

(i) RepNet, một kiến trúc mạng nơ-ron được thiết kế để đếm số lặp lại trong video một cách tự nhiên.

(ii) Một phương pháp để tạo và tăng cường các video lặp lại tổng hợp từ các video không được gắn nhãn.

(iii) Bằng cách huấn luyện RepNet trên tập dữ liệu tổng hợp, mô hình làm tốt hơn các phương pháp SOTA về cả các nhiệm vụ đếm lặp lại và dự đoán chu kỳ so với các điểm chuẩn hiện có một cách đáng kể.

(iv) Tập dữ liệu đếm lặp lại video mới, Countix, lớn hơn ∼ 90 lần so với tập dữ liệu lớn nhất trước đó.

# **2. Ứng dụng tiêu biểu**

*Dự đoán thay đổi tốc độ của các lần lặp lại*: đưa vào một video clip và dự đoán khoảng thời gian của bất kỳ hành động lặp lại nào. Sự khác biệt liên tiếp của tỷ lệ dự đoán mã hóa tốc độ thay đổi tốc độ của các lần lặp lại. Theo dõi các thay đổi về tốc độ rất hữu ích cho các ứng dụng theo dõi tập thể dục, trong đó điều quan trọng là phải biết ai đó đang tăng tốc hay chạy chậm lại.

*Ước tính tần suất của các quy trình*: dự đoán số lượng và tần suất của các hiện tượng lặp lại từ video, ví dụ: quá trình sinh học (nhịp tim).

*Truy xuất chi tiết xuyên thời kỳ*: nắm bắt những điểm tương đồng hiện diện trong các thời điểm khác nhau trong khi vẫn mã hóa sự khác biệt nhỏ nhất giữa các khung hình trông giống nhau gọt hoa quả (hoa quả nhỏ dần sau từng lần gọt) và cậu bé chơi xích đu (người nhỏ dần khi tiến ra xa, và to dần khi tiến lại gần).

*Các lần lặp lại với thời gian dài*: Nhiều hiện tượng lặp lại xảy ra trên quy mô thời gian dài hơn (theo thứ tự ngày hoặc năm) ví dụ như dự đoán khoảng thời gian của một ngày từ các video về trái đất được vệ tinh thu lại.

*Hỗ trợ học có giám sát từ video*. Phương pháp học tập tự giám sát để nhúng video, ví dụ: Các mạng Shuffle and Learn, Odd-One-Out, DPC, TCC và TCN không được thiết kế để xử lý các lần lặp lại theo trình tự. RepNet có thể xác định các phần lặp lại và có thể giúp đào tạo về các video có phần lặp lại mà không cần sửa đổi các mục tiêu đã đề xuất.

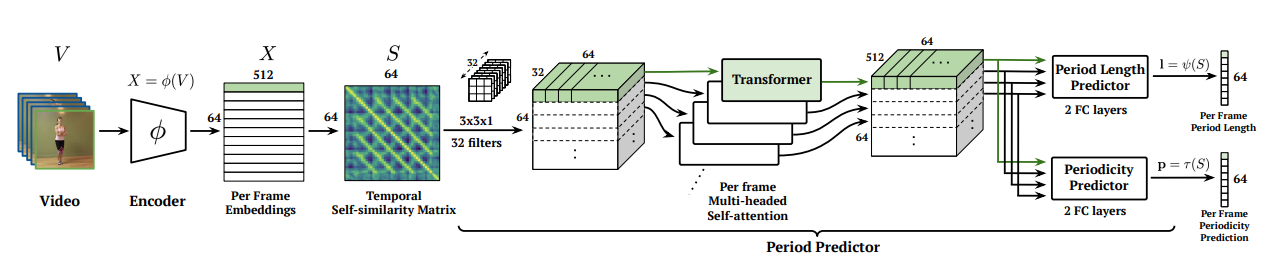
# **3. Mô hình mạng Repnet**

Mô hình mạng Repnet bao gồm 2 thành phần: bộ mã hóa và bộ dự đoán chu kỳ, với một lớp tự tương tự theo thời gian ở giữa chúng.

Giả sử chúng ta được cho một video V = [v1, v2, ..., vN] là một chuỗi N khung. Đầu tiên, cung cấp video V vào bộ mã hóa hình ảnh φ là X = φ(V) để tạo ra các nhúng trên mỗi khung hình X = [x1, x2, ..., xN]T. Sau đó, sử dụng phép nhúng X chúng ta thu được ma trận tự tương tự S bằng cách tính toán các điểm tương đồng theo cặp Sij giữa tất cả các cặp nhúng.

Cuối cùng, S được đưa vào mô-đun dự đoán chu kỳ, mô-đun này xuất ra hai phần tử cho mỗi khung: ước tính độ dài chu kỳ Ɩ = ψ(S) và điểm tuần hoàn

p = τ(S). Độ dài khoảng thời gian là tốc độ lặp lại diễn ra trong khi điểm số chu kỳ cho biết liệu khung hình có nằm trong một phần định kỳ của video hay không. Kiến trúc tổng thể có thể được xem trong Hình 1 và một phiên bản chi tiết hơn có thể được xem trong Hình 2.



*Hình 2: Kiến trúc RepNet. Các tính năng được tạo ra bởi một khung video duy nhất được đánh dấu bằng màu xanh lá cây trên toàn mạng.*

## 3.1. Bộ mã hóa

Bộ mã hóa φ bao gồm ba thành phần chính:

*Bộ giải nén tính năng(Convolutional feature extractor):* sử dụng kiến ​​trúc ResNet-50 làm mạng (CNN) cơ sở để trích xuất các tính năng phức hợp 2D từ các khung hình vi riêng lẻ của video đầu vào. Các khung này có kích thước 112 × 112 × 3. Sử dụng đầu ra của lớp conv4\_block3 để có một bản đồ tính năng 2D không gian lớn hơn. Các tính năng trên mỗi khung hình thu được có kích thước 7 × 7 × 1024.

*Bối cảnh thời gian*: chuyển các tính năng phức hợp này qua một lớp phức hợp 3D để thêm thông tin thời gian cục bộ vào các tính năng trên mỗi khung hình. Sử dụng 512 bộ lọc có kích thước 3 × 3 × 3 với kích hoạt ReLU với tỷ lệ giãn nở là 3. Bối cảnh thời gian giúp lập mô hình chuyển động ngắn hạn và cho phép mô hình phân biệt giữa các khung trông giống nhau nhưng khác chuyển động (ví dụ: tay di chuyển lên hoặc xuống trong khi tập thể dục).

*Giảm kích thước*: giảm kích thước của các đối tượng không gian-thời gian được trích xuất bằng cách sử dụng Global 2D Max-pooling và tạo ra các vectơ nhúng xi tương ứng với mỗi khung hình vi trong video.

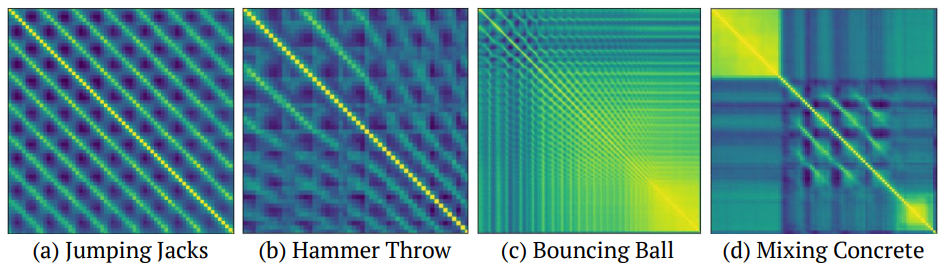
## 3.2. Ma trận tự tương tự theo thời gian (TSM-Temporal Self similarity Matrix)

Sau khi thu được các phép nhúng tiềm ẩn xi cho mỗi khung vi, xây dựng ma trận tự tương tự S bằng cách tính toán tất cả các điểm tương tự theo cặp

Sij = f(xi, xj) giữa các cặp nhúng xi và xj, trong đó f() là hàm tương tự. Sử dụng phủ định của khoảng cách euclide bình phương làm hàm tương tự:

f (a, b) = -|| a - b ||2, theo sau là phép toán softmax theo hàng.

Vì TSM chỉ có một kênh, nó hoạt động như một nút thắt cổ chai thông tin ở giữa mạng và cung cấp chính quy hóa. TSM cũng làm cho mô hình có thể diễn giải được theo thời gian, điều này mang lại hiểu biết sâu hơn về các dự đoán được thực hiện bởi mô hình. Một số ví dụ có thể được xem trong Hình 3.



*Hình 3: Sự đa dạng của ma trận tự tương tự theo thời gian được tìm thấy trong các video lặp lại trong thế giới thực (màu vàng có nghĩa là độ tương đồng cao, màu xanh lam có nghĩa là độ tương đồng thấp).*

(a) Chuyển động tuần hoàn lặp lại đồng đều (nhảy dây)

(b) Lặp lại với gia tốc (vận động viên thực hiện ném búa – lúc đầu chậm và nhanh dần về sau)

(c) Lặp lại với chu kỳ giảm dần (một quả bóng nảy mất tốc độ do bị đập xuống đất nhiều lần)

(d) Chuyển động lặp lại trước đó và không chuyển động tiếp theo (chờ trộn bê tông, trộn bê tông, dừng trộn).

Một mô hình phức tạp là cần thiết để dự đoán chu kỳ và tính tuần hoàn từ các ma trận tự tương tự đa dạng như vậy.

## 3.3. Dự đoán chu kỳ

Mô-đun cuối cùng của RepNet là công cụ dự đoán thời gian. Mô-đun này chấp nhận ma trận tự tương tự S = [s1, s2, ..., sN]T trong đó mỗi hàng si là biểu diễn tự tương tự trên mỗi khung và tạo ra hai đầu ra: ước tính độ dài chu kỳ trên mỗi khung hình Ɩ = ψ( S), và phân loại tuần hoàn nhị phân trên mỗi khung p = τ(S). Lưu ý rằng cả Ɩ và p đều là vectơ và các phần tử của chúng là các dự đoán trên mỗi khung (tức là Ɩi là độ dài chu kỳ dự đoán cho khung thứ i).

Kiến trúc của mô-đun dự báo chu kỳ có thể được xem trong Hình 2. Lưu ý rằng các mô-đun dự báo ψ và τ có chung một kiến ​​trúc và trọng số cho đến giai đoạn phân loại cuối cùng. Quy trình xử lý được chia sẻ bắt đầu với 32 bộ lọc tích chập 2D có kích thước 3 × 3, tiếp theo là lớp Transformer sử dụng sự chú ý từ nhiều phía (multi-headed attention) với các phép nhúng vị trí có thể huấn luyện dưới dạng một biến độ dài 64 được học qua đào tạo. Sử dụng 4 đầu với 512 chiều trong Transformer với mỗi đầu có kích thước 128 chiều. Sau đường ống được chia sẻ, chúng ta có hai bộ phân loại, bộ phân loại độ dài chu kỳ ψ và bộ phân loại tuần hoàn τ. Mỗi lớp bao gồm hai lớp được kết nối đầy đủ có kích thước 512.

## 3.4. Hàm Loss

Bộ phân loại tuần hoàn τ xuất ra mỗi khung phân loại định kỳ pi và sử dụng hàm Loss phân loại nhị phân (binary cross-entropy) để tối ưu hóa. Công cụ ước tính độ dài chu kỳ của ψ đầu ra cho mỗi ước tính độ dài chu kỳ khung Ɩi ∈ L trong đó các lớp là độ dài chu kỳ rời rạc L = {2, 3, ..., N/2} trong đó N là số khung đầu vào. Sử dụng mục tiêu phân loại nhiều lớp (softmax cross-entropy) để tối ưu hóa mô hình. Đối với tất cả các thử nghiệm sử dụng N = 64. Lấy mẫu video đầu vào với các tốc độ khung hình khác nhau như mô tả bên dưới để dự đoán độ dài khoảng thời gian lớn hơn.

## 3.5. Suy luận

Việc suy ra số lần lặp lại cho một video nhất định yêu cầu hai thao tác chính:

Đếm từ bước dự đoán độ dài chu kỳ: lấy mẫu các cửa sổ không chồng chéo liên tiếp của N khung và cung cấp nó làm đầu vào cho RepNet, xuất ra số pi theo chu kỳ trên mỗi khung và độ dài chu kỳ Ɩi. Xác định số lượng mỗi khung hình làpi/Ɩi. Tổng Số lần lặp lại được tính bằng tổng của tất cả các lần đếm trên mỗi khung hình: . Bộ dữ liệu đánh giá để đếm số lần lặp lại chỉ có các phân đoạn định kỳ. Do đó, đặt pi thành 1 làm giá trị mặc định cho các thử nghiệm đếm.

Đánh giá đa tốc độ: Vì mô hình có thể dự đoán độ dài khoảng thời gian lên đến 32, để bao gồm độ dài khoảng thời gian dài hơn nhiều, lấy mẫu video đầu vào với các tốc độ khung hình khác nhau. (tức là phát video ở tốc độ 1 ×, 2 ×, 3 × và 4 ×). Chọn tốc độ khung hình có điểm cao nhất trong khoảng thời gian dự đoán.

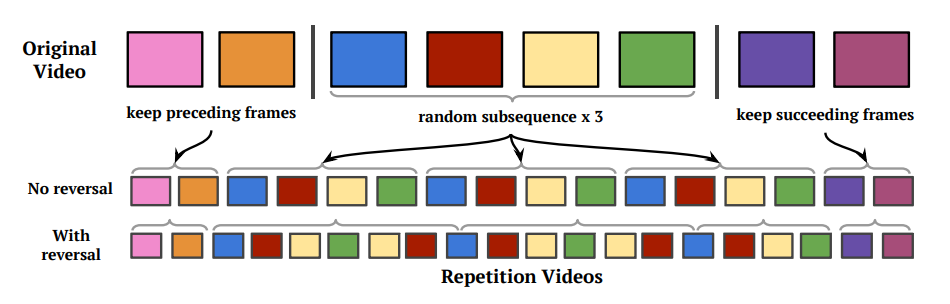
# **4. Huấn luyện mô hình với các lần lặp lại tổng hợp**

Một phương pháp tiếp cận có giám sát tiềm năng để ước tính khoảng thời gian sẽ là thu thập một tập hợp lớn các video đào tạo định kỳ và chú thích phần đầu và phần cuối của mỗi chu kỳ trong tất cả các hành động lặp lại. Tuy nhiên, việc thu thập một tập dữ liệu như vậy rất tốn kém do tính chất chi tiết của nhiệm vụ.

Là một giải pháp thay thế rẻ hơn và có thể mở rộng hơn, đề xuất một chiến lược đào tạo sử dụng các lần lặp lại được tạo tổng hợp bằng cách sử dụng các video không được gắn nhãn trong tự nhiên (ví dụ: YouTube). Tạo các video định kỳ tổng hợp bằng cách sử dụng các video được chọn ngẫu nhiên và dự đoán độ dài chu kỳ và chu kỳ trên mỗi khung hình. Tiếp theo, sẽ giải thích cách tạo ra số lần lặp lại tổng hợp và giới thiệu tính năng nâng cao chuyển động của máy ảnh rất quan trọng để đào tạo các mô hình đếm hiệu quả từ video tổng hợp.

## 4.1. Tổng hợp các video lặp lại

Với một tập hợp lớn các video không được gắn nhãn, đề xuất một cách tiếp cận đơn giản nhưng hiệu quả để tạo các video lặp lại tổng hợp từ chúng (thể hiện trong Hình 4). Lợi thế của việc sử dụng video thực để tạo dữ liệu tổng hợp là dữ liệu đào tạo gần với video lặp lại thực hơn nhiều so với việc sử dụng các mẫu tổng hợp. Một lợi thế khác của việc sử dụng video thực là việc sử dụng một bộ dữ liệu lớn như Kinetics đảm bảo rằng tính đa dạng của dữ liệu mà mô hình nhìn thấy là rất lớn. Điều này cho phép đào tạo các mô hình phức tạp lớn có thể hoạt động trên các video lặp lại thực tế.



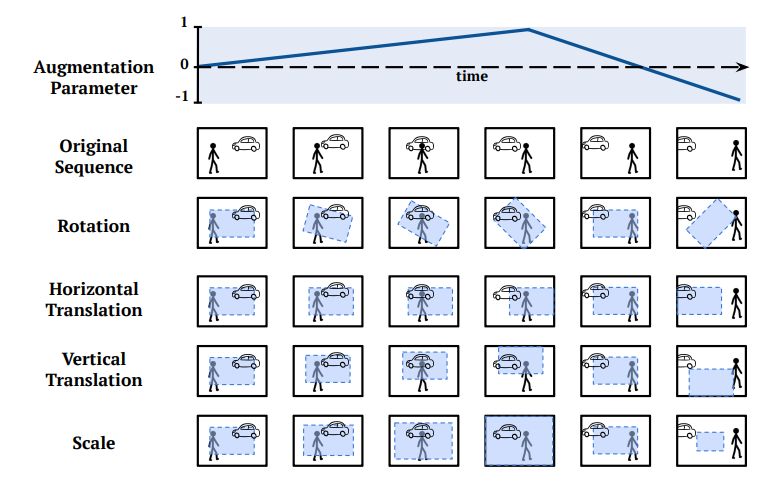
*Hình 4: Quy trình tạo dữ liệu tổng hợp tạo ra các video có sự lặp lại từ bất kỳ video nào. Lấy mẫu ngẫu nhiên một phần của video lặp lại N lần để tạo ra các video lặp lại tổng hợp.*

Quy trình bắt đầu bằng việc lấy mẫu một video V ngẫu nhiên từ tập dữ liệu các video. Sử dụng tập huấn luyện của Kinetics mà không có bất kỳ nhãn nào. Sau đó, lấy mẫu một clip C có độ dài ngẫu nhiên P khung hình từ V. Đoạn clip C này được lặp lại K lần (trong đó K> 1) để mô phỏng các video có số lần lặp lại. Ghép ngẫu nhiên clip đã đảo ngược trước khi lặp lại để mô phỏng các hành động trong đó chuyển động được thực hiện ngược lại trong khoảng thời gian (như nhảy dây). Sau đó, thêm vào đầu và nối các khung lặp lại với các đoạn không lặp lại khác từ V, các đoạn này ngay trước và sau C, tương ứng. Độ dài của các đoạn không theo chu kỳ này được chọn ngẫu nhiên và cũng có thể bằng 0.

Thao tác này đảm bảo rằng có cả phân đoạn định kỳ và không định kỳ trong video được tạo. Cuối cùng, mỗi khung hình trong phần lặp lại của video đã tạo được gán nhãn độ dài chu kỳ P. Một nhãn định kỳ cũng được tạo cho biết khung hình nằm bên trong hay bên ngoài phần lặp lại của video đã tạo.

## 4.2. Tăng cường chuyển động camera

Nhận thấy việc nâng cao chuyển động của máy ảnh là một phần quan trọng trong việc đào tạo các mô hình hiệu quả bằng video tổng hợp. Một bước quan trọng trong quá trình tạo video tổng hợp là tăng cường chuyển động của camera (thể hiện trong Hình 5).



*Hình 5: Chuyển động của camera. Thay đổi các thông số nâng cao cho từng loại máy ảnh di chuyển trơn tru theo thời gian thay vì lấy mẫu ngẫu nhiên chúng một cách độc lập cho từng khung hình. Điều này đảm bảo rằng trình tự tăng cường vẫn giữ được sự mạch lạc theo thời gian hiện diện tự nhiên trong video.*

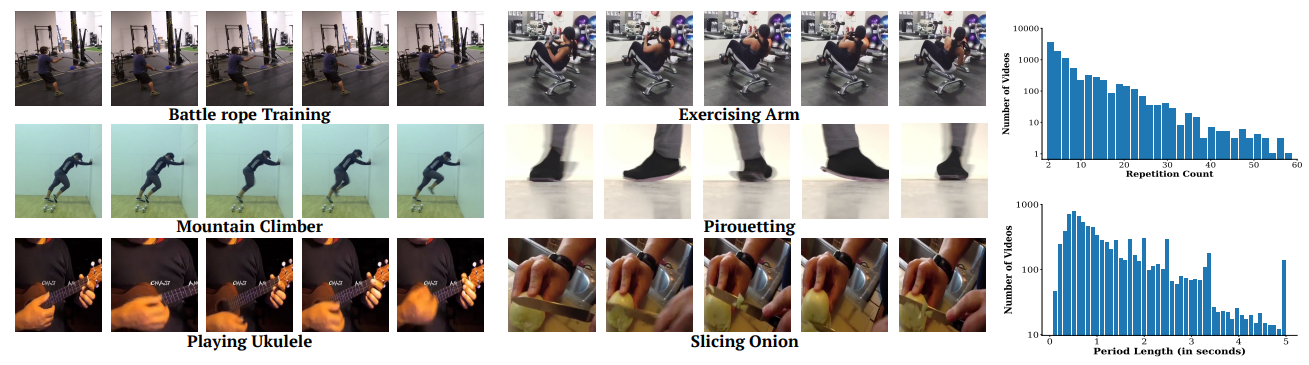
Để tăng cường chuyển động của máy ảnh, tạm thời thay đổi các thông số cho các loại chuyển động khác nhau một cách liên tục khi video tiếp tục. Ví dụ như thay đổi góc quay trơn tru theo thời gian. Điều này đảm bảo rằng video nhất quán về mặt thời gian ngay cả sau khi nâng cấp. Hình 5 minh họa cách tham số tăng thời gian điều khiển khung xem (được hiển thị trong hình chữ nhật màu xanh lam) cho từng loại chuyển động. Điều này dẫn đến các video có ít bản sao gần giống trên các phân đoạn lặp lại.

# **5. Bộ dữ liệu Countix**

Các bộ dữ liệu hiện có để đếm số lần lặp lại video hầu hết được sử dụng cho mục đích thử nghiệm, chủ yếu là do kích thước hạn chế của chúng. Tiêu chuẩn thử thách và gần đây nhất cho nhiệm vụ này là tập dữ liệu lặp lại QUVA bao gồm các video lặp lại thực tế với hiện tượng khớp, chuyển động của camera và thay đổi về tốc độ của các hành động lặp lại. Nó bao gồm 100 video kiểm tra bất khả tri (chưa biết rõ) theo lớp, được chú thích với số lượng các hành động lặp lại. Mặc dù đầy thách thức, nhưng kích thước hạn chế của nó làm cho nó khó có thể bao gồm các loại lặp lại ngữ nghĩa đa dạng. Ngoài ra, đào tạo mô hình sâu có giám sát với quy mô dữ liệu này là không khả thi.

Để tăng tính đa dạng về ngữ nghĩa và mở rộng quy mô của bộ dữ liệu đếm, giới thiệu bộ dữ liệu Countix: bộ dữ liệu trong thế giới thực gồm các video lặp lại được thu thập trong môi trường hoang dã (tức là YouTube) bao gồm một loạt các cài đặt ngữ nghĩa với những thách thức đáng kể như máy ảnh và chuyển động của đối tượng, tập hợp các khoảng thời gian và số đếm đa dạng, và những thay đổi về tốc độ của các hành động lặp lại.

Countix bao gồm các video lặp đi lặp lại về các hoạt động tập luyện (ngồi xổm, kéo xà, luyện dây chiến đấu, tập thể dục cánh tay), các động tác nhảy (đánh đàn, nắm đấm), chơi nhạc cụ (chơi ukulele), sử dụng các dụng cụ nhiều lần (búa đập vào đồ vật, cưa máy cắt gỗ, cắt hành tây), biểu diễn nghệ thuật (nhảy hula hooping, tung hứng bóng đá), thể thao (chơi bóng bàn và quần vợt) và nhiều môn khác. Hình 6 minh họa một số ví dụ từ tập dữ liệu cũng như sự phân bố số lần lặp lại và độ dài chu kỳ. Tập hợp tập dữ liệu: Tập dữ liệu Countix là một tập con của tập dữ liệu Kinetics được chú thích bằng các phân đoạn của các hành động được duyệt lại và số lượng tương ứng. Trong quá trình thu thập, trước tiên, chọn một cách thủ công một tập hợp con các lớp từ Kinetics có khả năng lặp lại cao hơn trong chúng, ví dụ: nhảy dây, thái hành tây, v.v., chứ không phải là các lớp học như đấu vật đối đầu hoặc đấu vật cá sấu.



*Hình 6: Tập dữ liệu Countix. Trong hai cột bên trái, trình bày các ví dụ về video lặp lại từ tập dữ liệu Countix. Cột cuối cùng hiển thị phân phối số lượng video trong tập dữ liệu liên quan đến số lượng và nhãn thời lượng.*

Cộng gộp các nhãn cho các phân đoạn lặp lại và số lượng cho các lớp đã chọn. Giao diện được sử dụng tương tự như giao diện thường được sử dụng để đánh dấu các phân đoạn thời gian để nhận dạng hành động chi tiết. Trước tiên, người chú thích được yêu cầu phân đoạn phần video chứa các đoạn lặp lại hợp lệ với số lượng rõ ràng. Các ký hiệu viên sau đó tiến hành đếm số lần lặp lại trong mỗi phân đoạn. Số lượng này đóng vai trò là nhãn cho toàn bộ clip. Từ chối các phân đoạn có sự trùng lặp không đáng kể trong phạm vi thời gian được đánh dấu bởi 3 chú thích khác nhau. Đối với các phân đoạn còn lại, sử dụng giá trị trung bình của số lượng chú thích và phạm vi phân đoạn làm chân lý cơ bản. Tập dữ liệu Countix lớn hơn khoảng 90 lần so với tập dữ liệu đếm số lần lặp lại lớn nhất trước đó (Bộ dữ liệu lặp lại QUVA).

Lưu ý rằng giữ lại phần tách train/val/test từ tập dữ liệu Kinetics. Do đó, các mô hình được đào tạo trước với Kinetics có thể được sử dụng để đào tạo các mô hình đếm mà không bị rò rỉ dữ liệu.

# **6. Đánh giá**

Thảo luận về hai bộ dữ liệu điểm chuẩn được thiết lập để dự đoán chu kỳ và đếm số lần lặp lại cùng với các số liệu đánh giá thường được sử dụng.

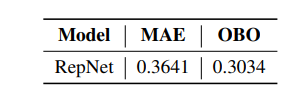
Dự đoán chu kỳ: Tập dữ liệu chuẩn cho tác vụ này là tập dữ liệu PERTUBE, có các nhãn trên mỗi khung xác định tính chu kỳ, nếu khung là một phần của hành động lặp lại hay không. Xử lý vấn đề dưới dạng một nhiệm vụ phân loại nhị phân trên mỗi khung hình và báo cáo độ chính xác, thu hồi, điểm F1 và chồng chéo (overlap).

Đếm số lần lặp lại: Như đã thảo luận trong Phần 5, bộ dữ liệu QUVA là bộ dữ liệu lớn nhất hiện có để đếm số lần lặp lại. Tài liệu hiện có sử dụng hai chỉ số chính để đánh giá số lần lặp lại trong video:

Lỗi đếm từng lỗi một (OBO - Off-By-One). Nếu số lượng dự đoán nằm trong một số đếm của giá trị sự thật cơ bản, thì video đó được coi là được phân loại chính xác, nếu không, đó là phân loại sai. Lỗi OBO là tỷ lệ phân loại sai trên toàn bộ tập dữ liệu.

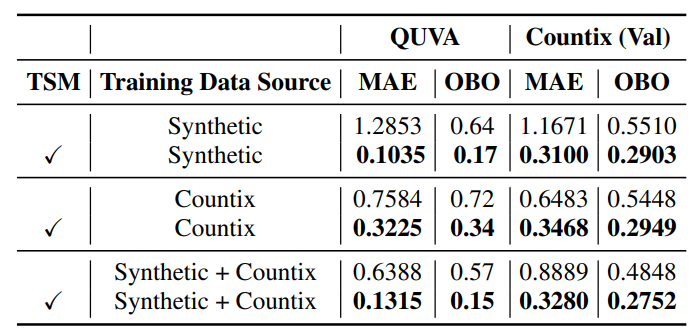
Sai số tuyệt đối trung bình (MAE - Mean Absolute Error) của số lượng. Chỉ số này đo lường sự khác biệt tuyệt đối giữa số lượng sự thật cơ bản và số lượng dự đoán, sau đó chuẩn hóa nó bằng cách chia cho số lượng sự thật cơ bản. Lỗi MAE được báo cáo là giá trị trung bình của sự khác biệt tuyệt đối được chuẩn hóa trên toàn bộ tập dữ liệu.

Cả trong các thử nghiệm cắt bỏ và so sánh SOTA, báo cáo lỗi OBO và MAE qua bộ xác thực QUVA và Countix. Cung cấp điểm số cuối cùng của bài kiểm tra Countix trong Bảng 7.



*Bảng 7: Kết quả đếm trên bộ thử nghiệm Countix*

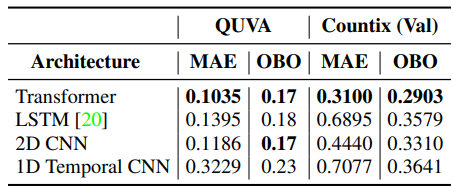
Ma trận tự tương tự theo thời gian (TSM): Trong Bảng 2, so sánh tác động của việc thêm TSM vào mô hình. Ma trận tự tương tự theo thời gian cải thiện đáng kể hiệu suất trên tất cả các chỉ số và bộ dữ liệu xác thực cho dù đào tạo mô hình bằng cách sử dụng video lặp lại tổng hợp, Countix hay kết hợp cả hai. Hơn nữa, lớp TSM giúp tổng quát hóa các video lặp lại thực sự ngay cả khi mô hình chỉ xem các video lặp lại tổng hợp (hàng 1 và 2 trong Bảng 2).



*Bảng 2: Kết quả đánh giá có hoặc không có ma trận tự tương tự theo thời gian (TSM) với các nguồn dữ liệu đào tạo khác nhau.*

Nguồn dữ liệu đào tạo: thay đổi các nguồn dữ liệu đào tạo trong Bảng 2 trong khi so sánh các video lặp lại tổng hợp với các video thực từ tập dữ liệu Countix. Nhận thấy rằng RepNet đạt được hiệu suất tương tự trên tập dữ liệu Countix khi được đào tạo với các video tổng hợp hoặc với các video lặp lại thực của tập dữ liệu Countix. Nhưng mô hình được đào tạo trên tập dữ liệu Countix kém hơn trên tập dữ liệu QUVA so với mô hình đào tạo về video lặp lại tổng hợp. Điều này cho thấy việc sử dụng tập dữ liệu lặp lại tổng hợp dẫn đến một mô hình cũng hoạt động cạnh tranh trên các lớp không nhìn thấy. Hiệu suất tốt nhất về lỗi OBO đạt được khi mô hình được đào tạo với cả hai bộ dữ liệu.

So sánh với các kiến trúc khác: Trong Bảng 3, so sánh kiến ​​trúc Transformer với các mô hình trình tự hiện đại khác như LSTM và CNNs tạm thời và so sánh nó với một mô hình sử dụng CNN 2D trên chính ma trận tự tương tự. Bhận thấy rằng kiến ​​trúc Transformer hoạt động tốt hơn các lựa chọn thay thế này.



*Bảng 3: Hiệu suất của các kiến trúc khác nhau khi được huấn luyện với dữ liệu tổng hợp*

# **7. Xây dựng ứng dụng web bằng Streamlit**

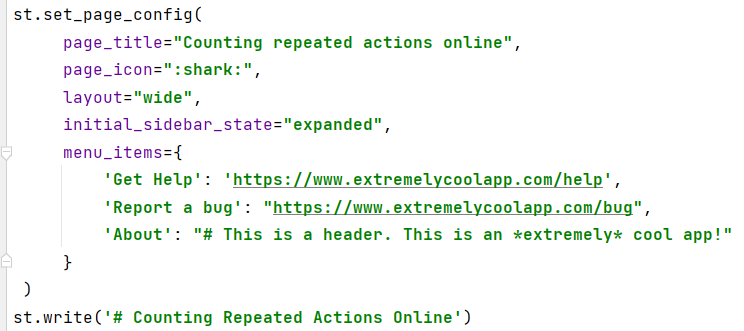
Streamlit là một thư viện Python mã nguồn mở giúp dễ dàng tạo và chia sẻ các ứng dụng web đẹp, tùy chỉnh cho máy học và khoa học dữ liệu.

*Cài đặt thư viện streamlit:*

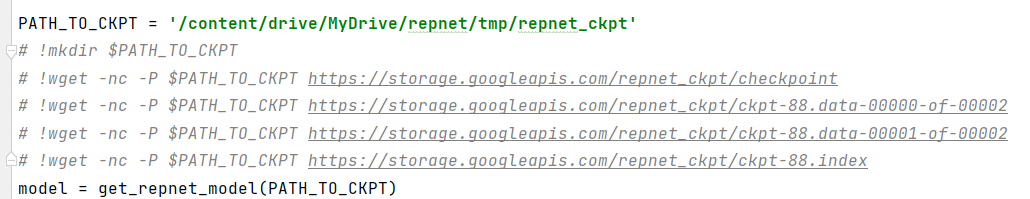
Pip install streamlit

Import streamlit as st

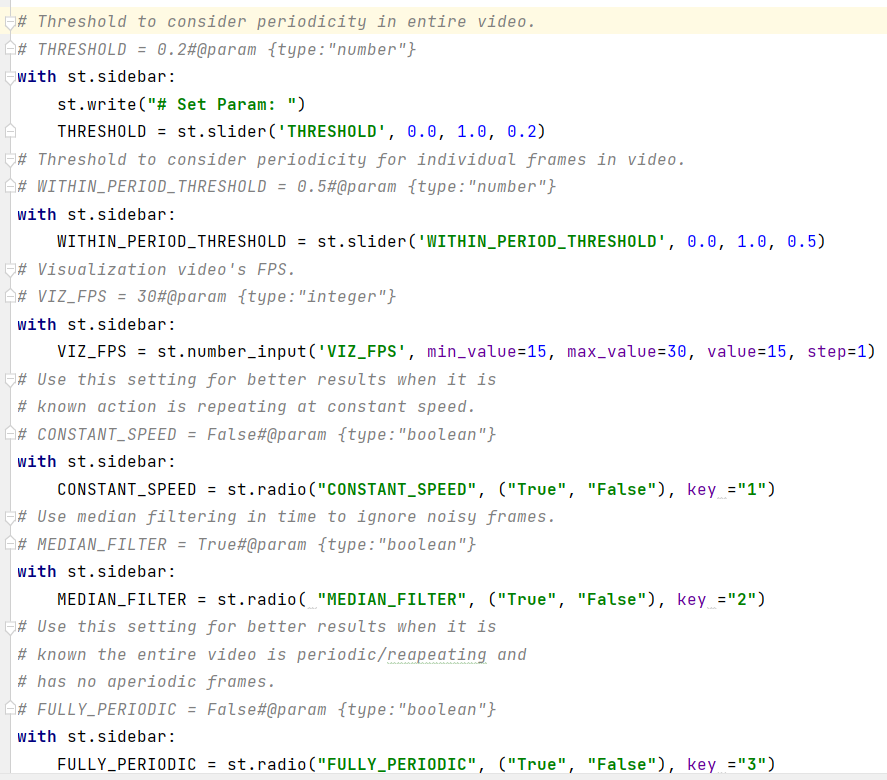
*Cài đặt các thông số cho trang web:*



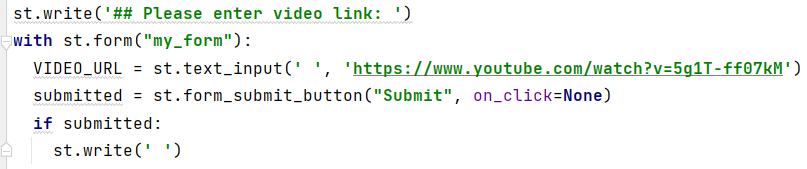
*Load pretrained model:*



*Thêm widget chỉnh sửa tham số mô hình:*



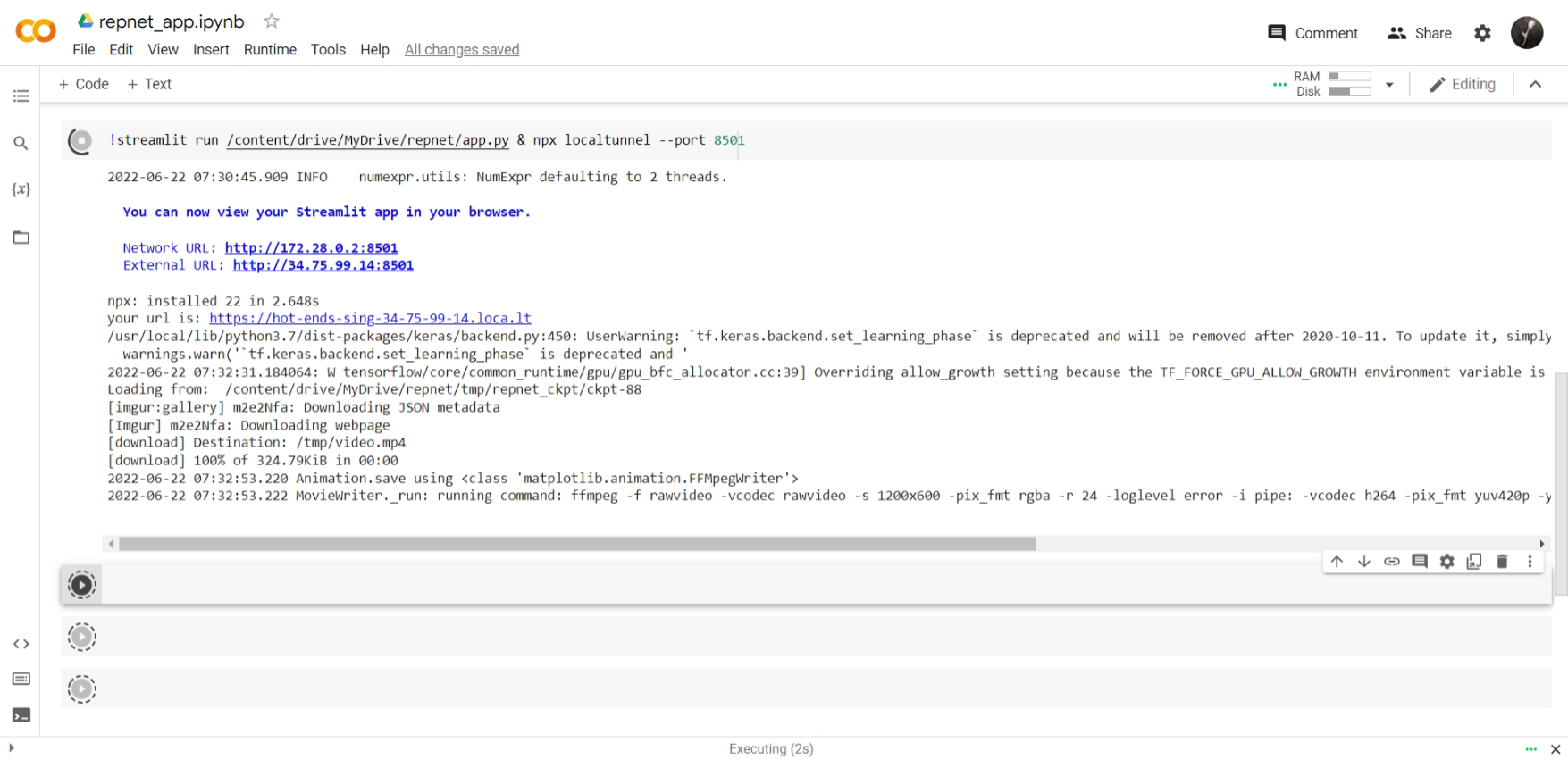
*Lấy link video đầu vào từ người dùng:*



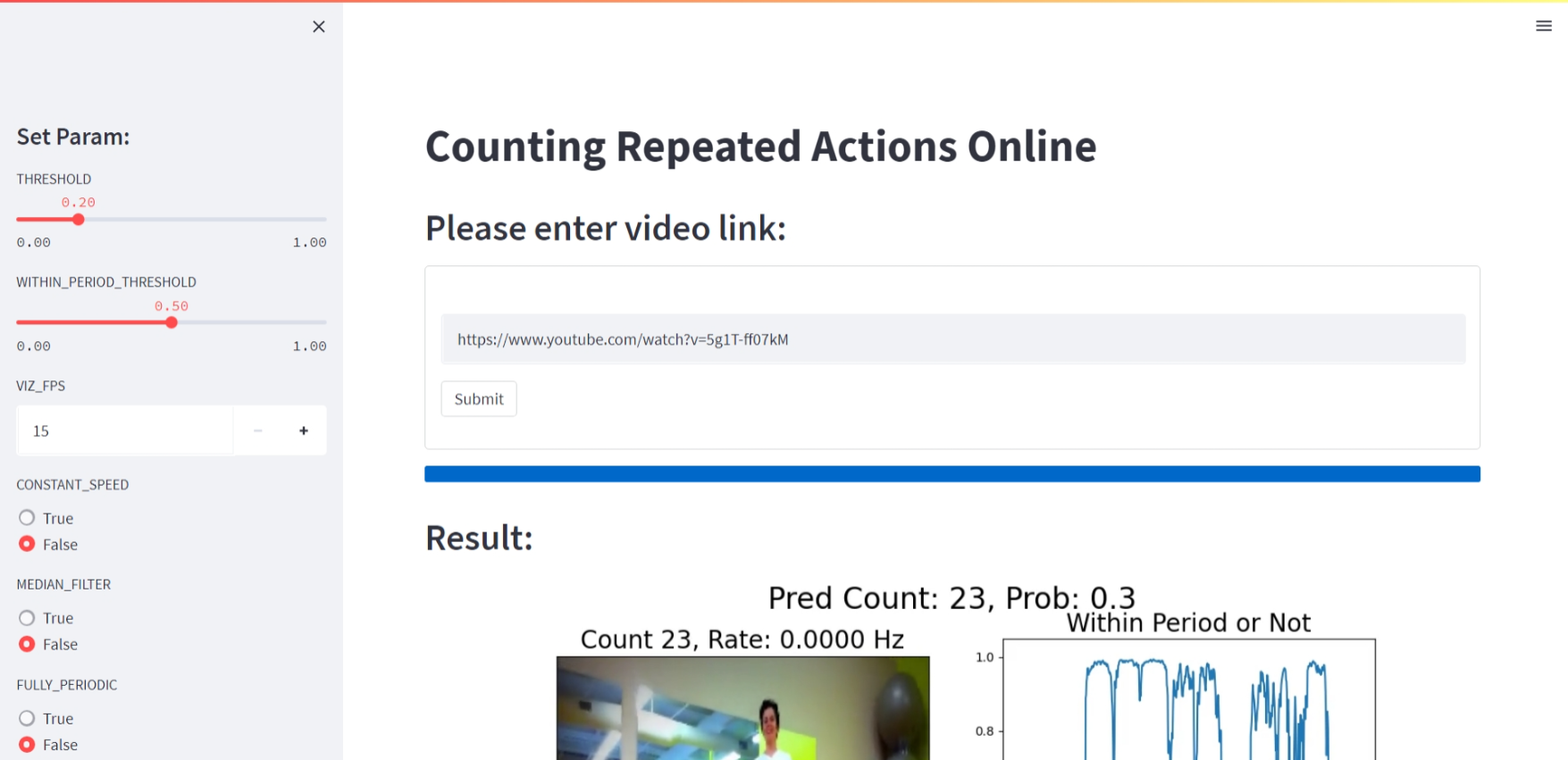
*In kết quả:*



*Chạy ứng dụng Streamlit trên Google Colab thay vì trên Localhost để tận dụng sức mạnh của GPU bằng cách sử dụng LocalTunnel.*



*Điền link video sau đó bấm submit để bắt đầu đếm lặp lại.*





*Giải thích các thông số trên menu Set param:*

THRESHOLD: Ngưỡng để xem xét tính chu kỳ trong toàn bộ video.

WITHIN\_PERIOD\_THRESHOLD: Ngưỡng để xem xét tính chu kỳ cho các khung hình riêng lẻ trong video.

VIZ\_FPS: FPS của video trực quan hóa.

CONSTANT\_SPEED: Sử dụng cài đặt này để có kết quả tốt hơn khi biết rằng hành động đang lặp lại với tốc độ không đổi.

MEDIAN\_FILTER: Sử dụng tính năng lọc trung vị kịp thời để bỏ qua các khung nhiễu.

FULLY\_PERIODIC: Sử dụng cài đặt này để có kết quả tốt hơn khi biết toàn bộ video là định kỳ / lặp lại và không có khung hình theo chu kỳ.

# **8. Tổng kết**

Một sự kết hợp đơn giản giữa dữ liệu huấn luyện tổng hợp, cùng với kiến trúc sử dụng ma trận tự tương tự theo thời gian, dẫn đến một mô hình đếm lặp lại bất khả tri theo lớp mạnh mẽ. Mô hình này dự đoán thành công tính chu kỳ và dự đoán số lượng trên một loạt các tác nhân (vật thể, con người, động vật, trái đất) và cảm biến (máy ảnh tiêu chuẩn, siêu âm, kính hiển vi laze) và đã được đánh giá trên một bộ sưu tập lớn các video. Tóm lại phương pháp đã giải quyết trường hợp lặp lại đơn giản và bước tiếp theo là xem xét các trường hợp phức tạp hơn như nhiều tín hiệu lặp lại đồng thời và sắp xếp thời gian của các phần lặp lại như trong các bước nhảy và âm nhạc.

Đã triển khai được mô hình qua ứng dụng web và cần nhiều cải tiến để ứng dụng trở nên tốt hơn.