DOI: 10.15625/vap.2016.00085

SỬ DỤNG DEEP NEURAL NETWORKS BIỂU DIỄN CÁC THUỘC TÍNH CHO BÀI TOÁN PHÁT HIỆN CẢNH BẠO LỰC TRONG VIDEO

Đỗ Văn Tiến [1], Lâm Quang Vũ [2], Phan Lê Sang [3], Ngô Đức Thành [1], Lê Đình Duy [1], Dương Anh Đức [1]

[1] Phòng Thí nghiệm Truyền thông Đa Phương tiện, Trường Đại học Công nghệ Thông tin, ĐHQG TP.HCM
[2] Khoa Công nghệ thông tin, Trường Đại học Khoa học Tự nhiên, ĐHQG TP.HCM
[3] Viện Tin học Quốc gia Nhật Bản (NII)

tiendv@uit.edu.vn, lqvu@fit.hcmus.edu.vn, plsang@nii.ac.jp, {thanhnd, lldduy,ducda}@uit.edu.vn

TÓM TẮT— *Deep Neural Networks (DNN) là một thuật toán máy học trong đó sử dụng mạng neural nhân tạo (Artificial Neural Networks) nhiều tầng để học, biểu diễn mô hình đối tượng. Với rất nhiều kết quả vượt trội so với các phương pháp trước đó, DNN đang được cộng đồng nghiên cứu thế giới sử dụng trong nhiều lĩnh vực như xử lý ảnh, xử lý âm thanh, xử lý ngôn ngữ tự nhiên...*

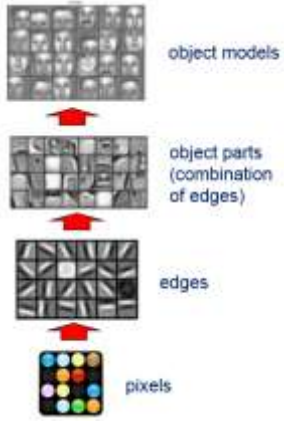
Trong bài báo này, chúng tôi đề xuất sử dụng DNN để biểu diễn các thuộc tính của khái niệm bạo lực như cảnh chứa máu, vũ khí, rượt đuổi xe, đánh nhau, cảnh chết chóc được sử dụng cho bài toán phát hiện cảnh bạo lực trong video (Violent Screne Detection -VSD). Đây là bài toán có tính thực tiễn và ứng dụng cao làm tiền đề để xây dựng các công cụ phân tích và kiểm duyệt nội dung video trên các kênh thông tin đa phương tiện trước khi tới người xem. Để đánh giá phương pháp đề xuất, chúng tôi xây dựng hệ thống trong đó sử dụng một số mô hình DNN phổ biến như Alexnet, UvANet, VGG để đánh giá độ chính xác trên tập dữ liệu chuẩn VSD[1] 2014. Kết quả thực nghiệm cho thấy, độ chính xác khi sử dụng DNN là 48,12% cao hơn so với phương pháp tốt nhất không sử dụng DNN 13%. Bên cạnh đó, bằng việc phân tích kết quả thực nghiệm chúng tôi sẽ đưa ra một số nhận xét trong việc lựa chọn thông tin từ các tầng phù hợp trong mô hình DNN cũng như cách thức biểu diễn video làm cơ sở cho các nhóm nghiên cứu có quan tâm đến bài toán này.

Từ khóa— *Violent scences detection, deep neural network, mid level feature.*

I. GIỚI THIỆU

Ngày nay, Internet đã trở nên rất phổ biến, mọi người ở mọi lứa tuổi đều có thể dễ dàng tiếp cận với các thông tin mà mình quan tâm dưới nhiều hình thức khác nhau như bằng văn bản, hình ảnh, âm thanh hoặc các đoạn video. Trong đó video là một những phương thức trực quan với lượng dữ liệu rất lớn, được chia sẻ trên nhiều kênh. Tuy nhiên, không phải tất cả các nội dung đều phù hợp với mọi lứa tuổi đặc biệt là trẻ em. Đã có nhiều nghiên cứu cũng như dẫn chứng đã chứng minh có sự ảnh hưởng giữa nội dung video đến hành vi của trẻ em đặc biệt là các nội dung bạo lực [1]. Từ thực tế này bài toán phát hiện cảnh bạo lực trong video được đề xuất và được mô tả như sau: đầu vào là video bất kì, đầu ra là các cảnh có chứa thông tin bạo lực. Trong đó, khái niệm cảnh bạo lực ở đây được định nghĩa như sau: cảnh bạo lực là cảnh chứa hình ảnh không phù hợp cho một đứa trẻ dưới 8 tuổi xem. Đây là một bài toán có tính ứng dụng cao, là tiền đề cho việc xây dựng các hệ thống tự động nhằm hỗ trợ phân tích và kiểm soát nội dung các video trước khi đến với người dùng, đặc biệt là trẻ em.

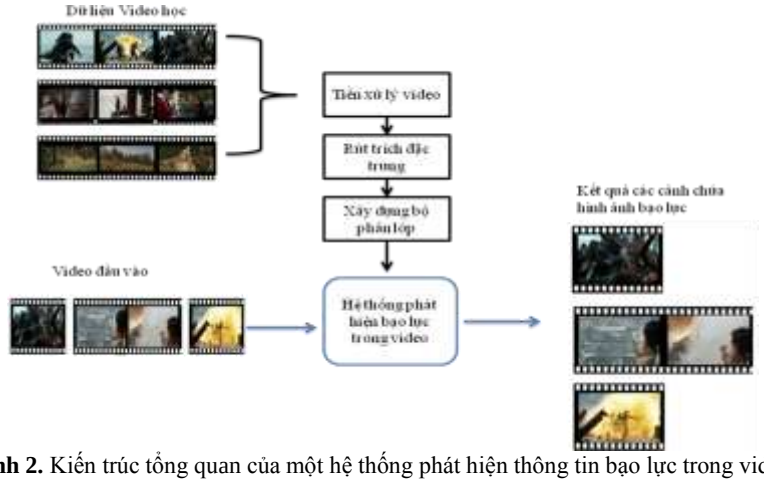
DNN là một khái niệm chỉ các thuật toán máy học để xây dựng mô hình đối tượng bằng cách học theo nhiều cấp biểu diễn từ các quan hệ phức tạp trong dữ liệu học [2]. Với các kết quả nổi bật trong bài toán nhận diện và phân lớp ảnh, trong đó độ chính xác tăng hơn 20% so với các thuật toán trước đây [3], cũng như được nhiều hãng công nghệ đầu tư áp dụng trong nhiều lĩnh vực khác nhau như: nhận dạng ảnh, xử lý tiếng nói, xử lý ngôn ngữ tự nhiên… DNN đang là một xu hướng mà cộng đồng nghiên cứu trên thế giới đặc biệt quan tâm.



Hình 1. Ý tưởng bài toán nhận diện mặt người sử dụng DNN [3].

[1] http://www.multimediaeval.org/

Ta có thể trình bày ý tưởng thuật toán DNN thực hiện trong bài toán nhận diện đối tượng như sau: để xây dựng mô hình biểu diễn được đối tượng cần học (trong bài toán này cụ thể là thông tin về mặt người – hình 1) thì thuật toán thực hiện học theo nhiều cấp. Trong đó, đầu ra của cấp thấp hơn sẽ là dữ liệu đầu vào của cấp cao hơn. Cụ thể đầu vào bài toán này là các bức ảnh mặt người cho quá trình học, thuật toán sử dụng các đơn vị điểm ảnh (pixel) trên bức ảnh để làm dữ liệu học cho tầng thứ nhất với kết quả học được là "khái niệm" edges (góc cạnh). Trong tầng tiếp theo bằng cách kết hợp các edges với nhau thuật toán sẽ học các "khái niệm" ở mức cao hơn như các phần của khuôn mặt (mắt, mũi…). Tương tự như vậy các tầng sau đó tiếp tục kết hợp các "khái niệm" để xây dựng mô hình khuôn mặt dùng cho việc nhận dạng.



Hình 2. Kiến trúc tổng quan của một hệ thống phát hiện thông tin bạo lực trong video.

Kiến trúc tổng quan của một hệ thống phát hiện cảnh bạo lực bao gồm các phần chính sau: (1) tiền xử lý video, (2) trích xuất đặc trưng, (3) sử dụng thuật toán máy học để xây dựng mô hình từ tập đặc trưng rút trích, (4) sử dụng mô hình đã học để phát hiện các cảnh bạo lực trong video đầu vào. Trong đó độ chính xác của hệ thống phụ thuộc nhiều vào việc trích chọn đặc trưng phù hợp ở bước (2) để biểu diễn cho thông tin bạo lực. Các nghiên cứu gần đây đã chỉ ra rằng việc sử dụng các đặc trưng cấp thấp như SIFT, HOG,... chưa thể hiện hết được ngữ nghĩa của khái niệm bạo lực [4]. Thay vào đó, các nghiên cứu này sử dụng tập các khái niệm và các thuộc tính liên quan đến hành vi, sự kiện, vật dụng liên quan đến bạo lực như: lửa (fire), vũ khí nóng (firearms), vật dụng gây sát thương (cold arms), đụng xe (car chases), cảnh chết chóc (gore), máu (blood), đánh nhau (fights) [5]. Bằng việc xây dựng các bộ phân lớp của các khái niệm và thuộc tính trên, cảnh bạo lực được xác định bằng cách tổng hợp điểm tương ứng của các bộ phân lớp. Tuy nhiên, các bộ phân lớp này vẫn sử dụng các đặc trưng cấp thấp.

Theo đó trong nghiên cứu này chúng tôi sẽ sử dụng DNN để xây dựng và biểu diễn các thuộc tính cho bài toán phát hiện cảnh bạo lực trong video, đây cũng là một nghiên cứu sơ khởi trong việc áp dụng DNN vào bào toán này. Chúng tôi sử dụng ba mô hình DNN được đánh giá là tốt nhất hiện nay bao gồm Alexnet [6] , UvANet [7], VGG [8] trên dữ liệu chuẩn VSD 2014 với gần 62,18 giờ video. Kết quả thực nghiệm cho thấy việc sử dụng DNN cho kết quả tốt hơn 13% so với việc sử dụng đặc trưng cấp thấp, trong đó với mô hình VGG 19 cho kết quả cao nhất là 48,12 %. Việc phân tích kết quả thực nghiệm lựa chọn và sử dụng thông tin được rút ra ở các tầng phù hợp nhất trong mô hình DNN cũng như cách thức biểu diễn thông tin một video làm cơ sở cho các nhóm nghiên cứu có liên quan đến việc áp dụng DNN cho bài toán này.

Bố cục của bài báo được trình bày như sau: phần II sẽ giới thiệu một số nghiên cứu liên quan đến bài toán phát hiện cảnh bạo lực trong video và sử dụng DNN trong các bài toán thị giác máy; phần III trình bày về hệ thống phát hiện cảnh bạo lực trong video sử dụng DNN để biểu diễn các thuộc tính để giải quyết bài toán; kết luận và hướng phát triển được trình bày trong phần IV.

II. MỘT SỐ NGHIÊN CỨU LIÊN QUAN

A. Một số nghiên cứu liên quan đến bài toán phát hiện cảnh bạo lực trong video

Độ lớn và phức tạp về mặt dữ liệu video cần xử lý cũng như sự nhập nhằng trong khái niệm bạo lực là những thách thức chính trong bài toán phát hiện cảnh bạo lực trong video. Đây cũng là bài toán được cộng đồng nghiên cứu trên thế giới đặc biệt quan tâm, trong đó các hướng nghiên cứu tập trung vào việc lựa chọn đặc trưng phù hợp để biểu diễn thông tin bạo lực. Các kết quả công bố đều sử dụng dữ liệu chuẩn trong cuộc thi VSD (MediaEval Affect Task: Violent Screnes Detection)[2]. Các nghiên cứu gần đây có thể được chia làm ba hướng nghiên cứu chính: hướng nghiên cứu sử dụng đặc trưng thị giác (visual feature) [9] [10] [11], hướng nghiên cứu sử dụng đặc trưng âm thanh [12] [13] , hướng nghiên cứu sử dụng kết hợp đa đặc trưng [14] [15] [16].

[2] http://www.multimediaeval.org/

Một số đặc trưng thị giác thường được sử dụng như Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HoG), Histograms of Optical Flow (HoF), ... trong đó một số nghiên cứu trước đây sử dụng những đặc trưng này để phát hiện các phân cảnh chứa lửa, máu, vụ nổ, ... từ đó làm cơ sở để phát hiện cảnh bảo lực. Nghiên cứu đầu tiên thuộc lĩnh vực này là của Jeho và cộng sự [9], nhóm tác giả đề xuất tiếp cận nhận dạng cảnh bạo lực bằng cách phát hiện các cảnh xuất hiện ngọn lửa, máu, phân tích mức độ chuyển động và sử dụng đặc trưng của hiệu ứng âm thanh. Trong khi đó Chen và cộng sự đã tách rời việc phát hiện cảnh bạo lực thành phát hiện cảnh hành động và cảnh đẫm máu [10]. Trong nghiên cứu của mình, Clarin và cộng sự giới thiệu hệ thống sử dụng lược đồ Kohonen để phát hiện cảnh có da người và máu kết hợp với phân tích cường độ các chuyển động để phát hiện các cảnh bạo lực [11].

Âm thanh cũng là một yếu tố quan trọng để phát hiện cảnh bạo lực trong video, Mel-frequency Cepstral Coefcient (MFCC) là đặc trưng âm thanh thường được các nhóm nghiên cứu sử dụng. Bằng việc sử dụng MFCC các nhóm nghiên cứu đã giành giải nhất cuộc thi về phát hiện sự kiện trong video (TRECVID Multimedia Event Detection) [12][13].

Hướng nghiên cứu giải quyết bài toán bằng cách kết hợp đa đặc trưng gần đây cũng được nhiều nhóm nghiên cứu quan tâm. Gong Yo và cộng sự đề xuất kết hợp đặc trưng âm thanh và đặc trưng thị giác [14]. Ngoài ra cách thức kết hợp các loại đặc trưng với nhau cũng được quan tâm nghiên cứu. Các nghiên cứu [15][16] chỉ ra rằng có hai hướng kết hợp đó là (1) Early Fusion: kết hợp các loại đặc trưng khác nhau thành đặc trưng chung để huấn luyện mô hình, (2) Late Fustion: kết quả được tổng hợp từ kết quả của các mô hình được học từ các đặc trưng riêng rẽ. Trong quá trình thực nghiệm, nhóm các tác giả này cũng đưa ra các kết quả thực nghiệm cho thấy độ chính xác của Late Fustion cao hơn so với Early Fusion.

Bên cạnh đó, các nghiên cứu gần đây sử dụng các thuộc tính để biểu diễn khái niệm bạo lực. Các thuộc tính ở đây liên quan đến hành vi, sự kiện, vật dụng liên quan đến bạo lực như: lửa (fire), vũ khí nóng (firearms), vật dụng gây sát thương (cold arms), rượt đuổi xe (car chases), cảnh chết chóc (gore), máu (blood), đánh nhau (fights) [5]. Trong nghiên cứu này nhóm cũng chỉ ra rằng việc sử dụng các thuộc tính sẽ cho kết quả nhận diện cảnh bạo lực tốt hơn so với việc sử dụng các đặc trưng thị giác thông thường. Tuy nhiên để biểu diễn các thuộc tính đề xuất nhóm nghiên cứu cũng chỉ sử dụng các đặc trưng thị giác như RGB-SIFT.

B. Một số nghiên cứu sử dụng DNN cho lĩnh vực thị giác máy

Một trong những lý do mà DNN được đặc biệt chú ý tới đó là khả năng học đặc trưng (learn feature representation). Khả năng này được cộng đồng nghiên cứu chú ý tới từ kết quả nghiên cứu của Andrew Ng [3] công bố trong việc nhận diện các đối tượng trong dữ liệu ImageNet[3]. Nhóm nghiên cứu đã sử dụng DNN để học mô hình của các đối tượng từ dữ liệu mà không sử dụng bất cứ đặc trưng thị giác nào, kết quả độ chính xác nhận dạng thu được cải thiện vượt trội so với phương pháp tốt nhất trước đó.

Tháng 10 năm 2012, trong cuộc thi về phân lớp ảnh (image classification) trên tập dữ liệu ImageNet (dữ liệu gồm 1,2 triệu ảnh của 1000 lớp) bằng cách sử dụng Deep Convolutional Neural Networks giáo sư Geoffrey Hinton và cộng sự đã thắng tuyệt đối với cách biệt lên đến 10 đến 15% so với đội đứng thứ hai [6]. Từ kết quả của nghiên cứu này, mô hình Alexet – kiến trúc mạng sử dụng trong quá trình huấn luyện mạng trên dữ liệu ImageNet ra đời, đây cũng là mô hình được rất nhiều nhóm nghiên cứu sử dụng cho các bài toán khác nhau. Mới đây nhất, bằng cách cải tiến kiến trúc mô hình Alexnet, nhóm nghiên cứu Zisserman đã đề xuất mô hình VGG, đây đang là mô hình cho kết quả tốt nhất đối với bài toán phân lớp ảnh trên dữ liệu ImageNet.

Trong nghiên cứu của nhóm Mettes [7] thay vì sử dụng một phần dữ liệu của ImageNet để huấn luyện mạng như Alexnet, thì nhóm sử dụng toàn bộ dữ liệu đã được tổ chức lại gồm 14 triệu ảnh với 21,814 lớp. Kết quả của quá trình huấn luyện là các mô hình UvANet, theo nhóm tác giả nghiên cứu đánh giá thì đây là mô hình cho kết quả tốt nhất cho bài toán phát hiện sự kiện trong video.

Ngoài ra, trong các lĩnh vực khác như xử lý tiếng nói, xử lý ngôn ngữ tự nhiên với việc áp dụng các thuật toán DNN đã đem lại các kết quả khả quan so với việc áp dụng các thuật toán trước đây. Đặc biệt các công ty lớn như Google, Facebook, Microsoft, Baidu đều thành lập các lab về DNN để nghiên cứu và áp dụng vào các sản phẩm của mình. Trong đó đã có một số ứng dụng được triển khai như dịch vụ tìm kiếm ảnh trong Google+, ứng dụng dịch của Microsoft Translator, hay chức năng nhận dạng tiếng nói trong Android.

Trong nghiên cứu này, chúng tôi sẽ sử dụng DNN để xây dựng và biểu diễn các thuộc tính được đề xuất trong nghiên cứu [5] cho bài toán phát hiện cảnh bạo lực trong video. Chúng tôi sử dụng ba mô hình DNN bao gồm Alexnet , UvANet, VGG đánh giá trên tập dữ liệu chuẩn VSD 2014 với gần 62,18 giờ video. Kết quả thực nghiệm cho thấy việc sử dụng DNN cho kết quả tốt hơn 13% so với việc sử dụng đặc trưng cấp thấp, trong đó với mô hình VGG 19 cho kết quả cao nhất là 48,12%.

[3] http://www.image-net.org/

III. DEEP NEURAL NETWORKS BIỂU DIỄN CÁC THUỘC TÍNH CHO BÀI TOÁN PHÁT HIỆN CẢNH BẠO LỰC TRONG VIDEO

C. Kiến trúc hệ thống

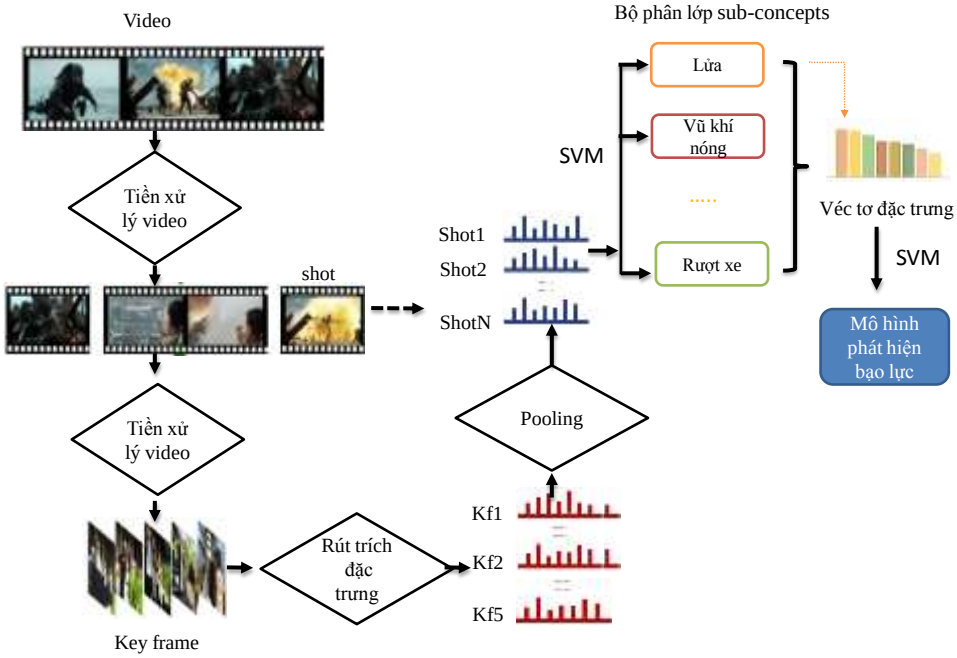
Chúng tôi xây dựng một hệ thống trong đó cho phép đánh giá việc sử dụng các mô hình DNN khác nhau biểu diễn các thuộc tính cho bài toán phát hiện cảnh bạo lực trong video. Hệ thống bao gồm các thành phần chính sau: tiền xử lý video, rút trích đặc trưng, xây dựng bộ phân lớp ứng với mỗi thuộc tính, huấn luyện mô hình.

1. Tiền xử lý video

Đầu vào của hệ thống là các video mà cụ thể ở đây trong dữ liệu mà chúng tôi sử dụng từ cuộc thi MediaEval Affect Task [17][18] là các bộ phim Hollywood. Các video sẽ được cắt thành các đoạn (shot) mỗi đoạn có thời lượng là 5 giây, trong mỗi đoạn chúng tôi sẽ lấy mẫu theo tần suất 5 cảnh (keyframe)/ giây được làm dữ liệu đầu vào cho quá trình rút trích đặc trưng tiếp theo. Việc lấy mẫu cũng như thông số về thời gian trong một đoạn được sử dụng theo nghiên cứu nhằm đảm bảo mức cân bằng giữa mặt thời gian và độ chính xác sau khi rút trích đặc trưng [19].

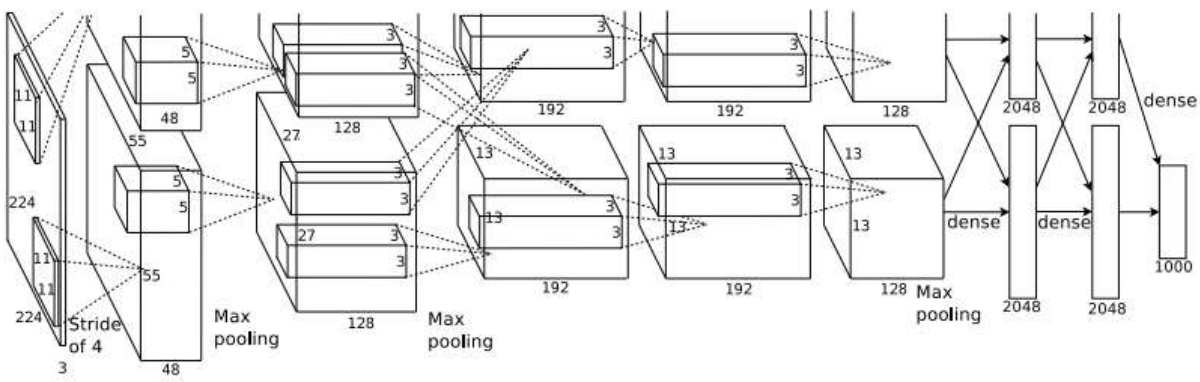
2. Rút trích đặc trưng

Chúng tôi sử dụng Caffe framework [20] để thực hiện quá trình rút trích đặc trưng từ ba mô hình bao gồm Alexnet , UvANet, VGG.



Hình 3. Kiến trúc hệ thống đề xuất phát hiện cảnh bạo lực trong video sử dụng DNN

Trong đó, Alexnet là mô hình được học từ dữ liệu của Imagnet với kiến trúc gồm 8 tầng (layer) trong đó có 5 layer đầu là convolutional layer và 3 layer còn lại là fully connected layer. Đầu ra lớp cuối cùng là 1000 chiều tương ứng với số lớp cần phân lớp. Chúng tôi sẽ thực nghiệm dựa trên việc rút trích đặc trưng ở 3 layer cuối trong kiến trúc mạng Alexnet (fc6, fc7, fully connected layer) dữ liệu rút trích lần lượt có số chiều tương ứng là 4096, 4096 và 1000.



Hình 4. Kiến trúc mạng Alexnet [6]

UvANet được giới thiệu là một mô hình được học trên toàn bộ dữ liệu của Imagnet trong đó nhóm nghiên cứu cung cấp bốn mô hình khác nhau căn cứ vào cách học mô hình và số đầu ra của lớp cuối cùng. Tên các mô hình tương ứng với số lớp bao gồm UvANet_Bottom-up-4k 4437, UvANet_Bottom-up-8k 8201, UvANet_Bottom-up-13k 12988.

VGG cung cấp 2 mô hình tương ứng với số layer khác nhau đó là VGG 16 và VGG 19. Trong mỗi mô hình đều có kiến trúc gồm có 3 layer cuối là fullyconnected layer, 2 lớp kế cuối có số chiều là 4096 và tầng cuối cùng có chiều là 1000 tương ứng với số lớp của bài toán phân lớp ảnh trên dữ liệu của ImagNet.

Ứng với mỗi shot chúng tôi sẽ rút trích đặc trưng của 5 keyframe tương ứng và tiến hành thực hiện tổng hợp để được đại diện đặc trưng cho shot. Trong quá trình thực nghiệm chúng tôi thực hiện tổng hợp và so sánh theo hai cách là lấy giá trị lớn nhất (max pooling) và lấy giá trị tổng từ dữ liệu véc-tơ đặc trưng của 5 keyframe/ giây để đại diện cho một shot.

3. *Xây dựng bộ phân lớp cho thuộc tính*

Từ tập đặc trưng được rút ra theo từng shot được chuẩn hóa về đoạn [0,1], chúng tôi sẽ sử dụng LibSVM [21] kết hợp với phương pháp k-fold cross validation với k=5 nhằm mục tiêu tối ưu các tham số (C,g) của thuật toán SVM với chi-square kenel để xây dựng 13 bộ phân lớp tương ứng với 13 thuộc tính biểu diễn cho khái niệm bạo lực được giới thiệu trong nghiên cứu [5]. Ứng với mỗi shot, chúng tôi sẽ xây dựng một véc-tơ đặc trưng biểu diễn thông tin 13 thuộc tính tương ứng với 13 chiều làm dữ liệu đầu vào cho quá trình huấn luyện xây dựng mô hình phát hiện cảnh bạo lực trong video. Trong đó giá trị tương ứng với mỗi chiều trong véc-tơ là điểm của từng bộ phân lớp của từng thuộc tính được xây dựng ở bước trên.

4. *Xây dựng mô hình phát hiện cảnh bạo lực trong video*

Đầu vào của quá trình huấn luyện mô hình để nhận diện cảnh bạo lực trong video là véc-tơ đặc trưng 13 chiều được đề cập ở bước trên, chúng tôi sử dụng SVM với cách thức tương tự như quá trình huấn luyện mô hình các thuộc tính. Kết quả là mô hình phân lớp được sử dụng cho bước đánh giá kết quả quá trình huấn luyện.

D. Thực nghiệm và đánh giá

Với mục tiêu đánh giá việc áp dụng DNN vào việc biểu diễn các thuộc tính cho bài toán phát hiện cảnh bạo lực trong video, đồng thời phân tích việc lựa chọn kiến trúc phù hợp và cách thức biểu diễn video cho bài toán này chúng tôi tiến hành thực nghiệm với các thông tin về dữ liệu, độ đo và kết quả như sau:

1. *Dữ liệu thực nghiệm*

Để đánh giá phương pháp đề xuất chúng tôi sử dụng dữ liệu từ cuộc thi MediaEval Affect Task 2014, dữ liệu được lấy từ 31 bộ phim Holllywood, đây cũng là dữ liệu chuẩn được sử dụng cho các nhóm nghiên cứu liên quan đến bài toán phát hiện cảnh bạo lực trong video. Đầu vào của bài toán là video và bài toán yêu cầu phát hiện ra các khung hình chứa cảnh bạo lực. Trong quá trình thực nghiệm chúng tôi chia tập dữ liệu ra làm hai phần dùng để học mô hình và kiểm tra mô hình xây dựng được. Tập học bao gồm 24 phim với tổng số giờ phim là 48,19 giờ tương ứng là 34.779 shot. Trong khi đó tập kiểm tra bao gồm 7 phim với tổng 13,89 giờ phim tương ứng là 10.006 shot.

Bảng 1. Thống kê dữ liệu trong tập xây dựng mô hình

STT	Tên phim	Thời gian (giây)	Số keyframe	Số shot
1	Armageddon	8681,05	217026	1737
2	BillyElliot	6349,36	158734	1270
3	Eragon	5985,57	149639	1198
4	Harry Potter 5	7954,72	198868	1591
5	I Am Legend	5780,58	144514	1157
6	Leon	6344,49	158612	1269
7	Midnight Express	6960,96	174024	1393
8	Pirates Of The Caribbean 1	8241,01	206025	1649
9	Reservoir Dogs	5721,98	142825	1143
10	Saving Private Ryan	9750,89	243772	1951
11	The Sixth Sense	6178,01	154450	1236
12	The Wicker Man	5870,89	146772	1175
13	The Bourne Identity	6816,29	170407	1364
14	The Wizard of Oz	5859,29	146482	1172
15	Dead Poets Society	7415,17	185379	1484
16	Fight Club	8006,34	200158	1602
17	Independence Day	8834,96	220874	1767
18	The Godfather	10194,96	254874	2039
19	Pulp Fiction	8887,97	222199	1778
20	Forrest Gump	8176,97	204424	1636
21	Fargo	5646,34	141158	1130
22	The Pianist	8567,10	241177	1714
23	Fantatic Four 1	6097,41	152360	1219
24	Legally Blond	5523,49	138087	1105
Tổng		173833,8	4345840	34779

Bảng 2. Thống kê dữ liệu trong tập đánh giá

STT	Tên phim	Thời gian (giây)	Số keyframe	Số shot
1	V for Vendetta	7626,49	190662	1526
2	Terminator 2	8831,37	220784	1767
3	Jumanji Collectors	5993,98	149849	1199
4	Ghost in the Shell	4966,00	124150	994
5	Desperado	6012,89	150322	1203
6	Brave Heart	10224,49	255612	2045
7	8 Mile	63655,53	158888	1272
Tổng		50010,75	1250267	10006

2. *Độ đo và phương pháp đánh giá*

Chúng tôi sử dụng độ đo MAP (Mean Average Precision) được ban tổ chức cuộc thi MediaEval VSD 2014 công bố ứng với tập dữ liệu mà nhóm đang sử dụng. Độ đo dựa trên thứ tự các shot được trả về từ hệ thống phát hiện cảnh bạo lực trong một video so với kết quả được đưa ra từ ban tổ chức. MAP được tính bằng công thức sau:

$$MAP = \frac{\sum_{v=1}^{V} AP(v)}{V}$$

Ở đây V là tổng số video và AP là độ chính xác trung bình cho từng video. Trong đó AP được tính theo công thức sau:

$$AP = \frac{\sum_{k=1}^{n}(P(k)\times rel(k))}{number\ \ of\ positive\ \ shots}$$

Trong đó P(k) là độ chính xác của top k phân đoạn có độ bạo lực cao nhất do hệ thống trả về và rel(k) sẽ bằng 1 nếu phân đoạn thứ k được gán nhãn là bạo lực (được ban tổ chức VSD cung cấp) hoặc 0 nếu đoạn đó không chứa cảnh bạo lực.

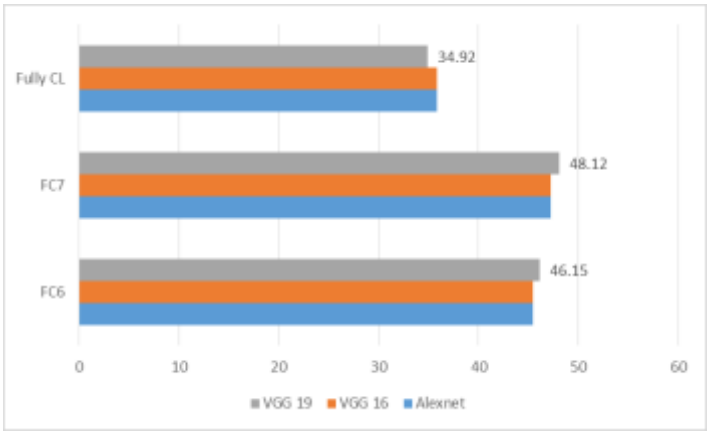
3. *Kết quả thực nghiệm*

Chúng tôi tiến hành đánh giá việc sử dụng DNN để biểu diễn các thuộc tính cho bài toán phát hiện cảnh bạo lực trong video. Bảng 3 là kết quả thực nghiệm trên 3 mô hình mà chúng tôi đã đề cập bên trên bao gồm Alexnet, UvANet và VGG trong đó gồm các kết quả rút trích đặc trưng ở các layer khác nhau. Hai phương thức kết hợp đặc trưng của 5 keyframe của một shot bao gồm lấy giá trị lớn nhất (max pooling) và lấy giá trị tổng (sum pooling) cũng được đánh giá.

Bảng 3. Kết quả thực nghiệm đánh giá mô hình DNN cho bài toán VSD

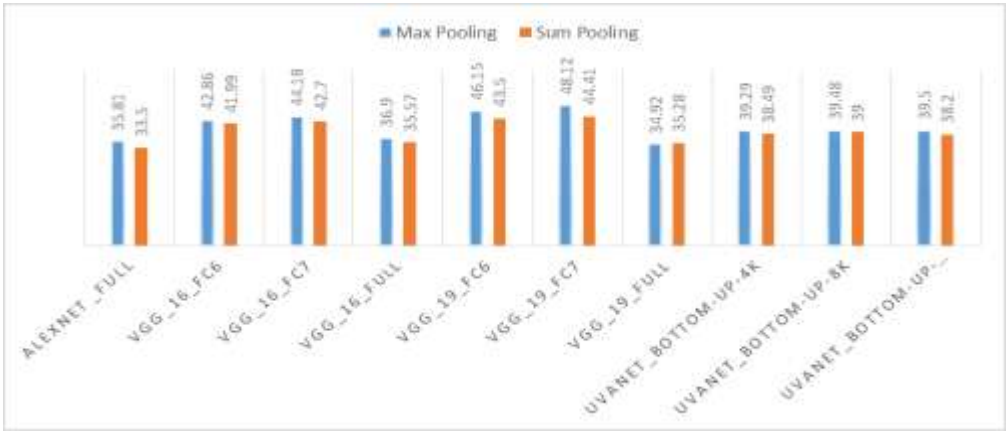
Mô hình	Cấu trúc mạng	MAP % (max pooling)	MAP % (sum pooling)
Alexnet	fc6	45,43	**44,89**
	fc7	47,21	44,8
	Fully connected layer	35.81	33.50
VGG 16	fc6	42.86	41.99
	fc7	44.18	42.7
	Fully connected layer	36.9	35.57
VGG 19	fc6	46.15	43.5
	fc7	**48.12**	44.41
	fullyconnected layer	34.92	35.28
UvANet	UvANet_Bottom-up-4k	39,29	38,49
	UvANet_Bottom-up-8k	39,48	39
	UvANet_Bottom-up-12k	39,5	38,2

Dựa vào bảng kết quả cho thấy trong ba mô hình đánh giá thì mô hình VGG cho kết quả cao nhất với độ chính xác là 48,12% ứng với đặc trưng được rút trích từ fc7 layer. Phân tích kết quả thực nghiệm trên phương diện cấu trúc mạng trên các mô hình cho phép rút trích đặc trưng biểu diễn ở các tầng khác nhau Alexnet và VGG – hình 4 , ta có thấy rằng việc sử dụng đặc trưng từ các lớp kế cuối luôn cho kết quả tốt hơn. Trong đó đặc trưng ở lớp fc7 luôn cho kết quả cao nhất, điều này cũng phù hợp với mô tả về DNN đã nhấn mạnh việc càng ở lớp kế sau thì mô hình càng mang tính tổng quát hóa.



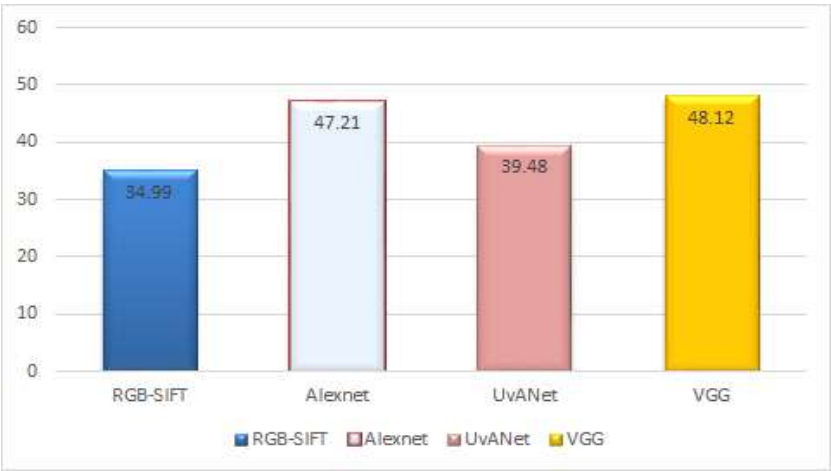
Hình 4. Phân tích kết quả lựa chọn kiến trúc DNN.

Trên phương diện biểu diễn thông tin video theo cách cắt video thành từng shot và lấy mẫu keyframe thì kết quả thực nghiệm cho thấy việc kết hợp các đặc trưng từ các keyframe bằng phương pháp max pooling thường cho kết quả tốt hơn so với sum pooling.



Hình 5. Phân tích kết quả lựa chọn cách biểu diễn video.

Ngoài ra, để so sánh việc sử dụng DNN với các đặc trưng thị giác thông thường, chúng tôi tiến hành thực nghiệm tương tự trong đó thay thế việc sử dụng DNN bằng đặc trưng RGB-SIFT – một đặc trưng mà rất nhiều nghiên cứu sử dụng trong xử lý ảnh. Theo đó độ chính xác đạt được khi sử dụng RGB-SIFT là 39.44%, dựa vào biểu đồ so sánh kết quả ở hình 6, so sánh kết quả tốt nhất của các mô hình cho ta thấy việc sử dụng DNN đem lại hiệu quả tốt hơn 13% so với việc sử dụng đặc trưng thị giác thông thường.



Hình 6. So sánh sử dụng DNN với phương pháp sử dụng đặc trưng thị giác thông thường

Như vậy, việc sử dụng DNN để biểu diễn các thuộc tính cho bài toán phát hiện cảnh bạo lực trong video sẽ mang lại hiệu quả tốt hơn so với sử dụng các đặc trưng thị giác thông thường. Trong đó, khi sử dụng các mô hình DNN thì thông tin ở các tầng kế cuối thường mang lại độ chính xác cao hơn. Ngoài ra, đối với bài toán phát hiện sự kiện

trong video nói chung cũng như phát hiện cảnh bạo lực trong video nói riêng thì phương pháp maxpooling các đặc trưng theo từng đoạn sẽ mang lại hiệu quả tốt hơn so với phương pháp sum pooling.

IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, chúng tôi đã đề xuất việc sử dụng DNN để biểu diễn các thuộc tính cho bài toán phát hiện cảnh bạo lực trong video. Đây là một bài toán có tính ứng dụng và thực tiễn cao, đặc biệt cần thiết cho việc kiểm soát các nội dung video bạo lực trước khi chúng tiếp cận với người dùng. Từ đó, chúng tôi đã xây dựng một hệ thống cho phép đánh giá việc sử dụng DNN mà cụ thể hơn là 3 mô hình đang được cộng đồng nghiên cứu về xử lý ảnh sử dụng đó là Alexnet, UvANet, VGG. Kết quả thực nghiệm trên tập dữ liệu và độ đo chuẩn của cuộc thi VSD 2014 cho thấy việc sử dụng DNN giúp tăng độ chính xác lên hơn 13% so với sử dụng đặc trưng thị giác thông thường. Đồng thời việc sử dụng các thông tin, đặc trưng ở các tầng kế cuối trong mô hình DNN và biểu diễn đoạn video bằng phương pháp maxpooling trên tập đặc trưng của keyframe sẽ mang lại hiệu quả.

Trong thời gian tới, chúng tôi sẽ tập trung vào việc xây dựng và bổ sung tập các thuộc tính mô tả khái niệm bạo lực để nâng cao độ chính xác của quá trình nhận diện. Ngoài ra, chúng tôi sẽ nghiên cứu và đề xuất mô hình DNN riêng cho bài toán phát hiện sự kiện nói chung và phát hiện cảnh bạo lực trong video nói chung.

V. LỜI CẢM ƠN

Nghiên cứu này là sản phẩm của đề tài "Nghiên cứu một số kĩ thuật deep learning cho các bài toán nhận dạng ảnh" mã số D2015-10, thuộc Trường Đại học Công nghệ thông tin - ĐHQG TP.HCM

TÀI LIỆU THAM KHẢO

[1] Http://www.ocd.pitt.edu/Files/PDF/Parenting/TvAndMovieViolence.pdf, "TV and Movie Violence: Why watching it is harmful to children," *Accessed 10 Jan 2015*.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[3] Q. V Le, "Building High-Level Features Using Large Scale Unsupervised Learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 8595–8598.

[4] C. Liang-Hua, H.-W. Hsu, L.-Y. Wang, and and Chih-Wen Su, "Violence Detection in Movies," in *Computer Graphics, Imaging and Visualization (CGIV)*, 2011, pp. 119–124.

[5] V. Lam, S. Phan, D. T. Ngo, D.-D. Le, D. A. Duong, and S. Satoh, "Violent Scene Detection Using Mid-level Feature," in *The Fourth Symposium on Information and Communication Technology (SoICT)*, 2013, pp. 198–205.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances In Neural Information Processing Systems*, 2012, pp. 1–9.

[7] P. Mettes, D. C. Koelma, and C. G. M. Snoek, "The ImageNet Shuffle," *Proc. 2016 ACM Int. Conf. Multimed. Retr. - ICMR '16*, pp. 175–182, 2016.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Prepr. arXiv1409.1556*, 2014.

[9] J. Nam, M. Alghoniemy, and a. H. Tewfik, "Audio-visual content-based violent scene characterization," *Proc. 1998 Int. Conf. Image Process. ICIP98 (Cat. No.98CB36269)*, vol. 1, pp. 353–357, 1998.

[10] C. Liang-Hua, H.-W. Hsu, L.-Y. Wang, and and Chih-Wen Su, "Violence Detection in Movies," *Comput. Graph. Imaging Vis.*, pp. 119–124, 2011.

[11] C. Clarin, J. Dionisio, M. Echavez, and P. Naval, "DOVE: Detection of movie violence using motion intensity analysis on skin and blood," *Pcsc*, pp. 150–156, 2005.

[12] R. Aly, R. Arandjelovic, K. Chatfield, M. Douze, B. Fernando, Z. Harchaoui, K. Mcguiness, N. O'Connor, D. Oneata, O. Parkhi, D. Potapov, J. Revaud, C. Schmid, J.-L. Schwenninger, D. Scott, T. Tuytelaars, J. Verbeek, H. Wang, and A. Zisserman, "The AXES submissions at TrecVid 2013," *TRECVID Work.*, 2013.

[13] D. Oneata, J. Verbeek, and C. Schmid, "The LEAR submission at Thumos 2014," *ECCV2014 THUMOS Chall.*, 2014.

[14] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5353 LNCS, pp. 317–326, 2008.

[15] C. Penet, C. H. Demarty, G. Gravier, and P. Gros, "Technicolor and INRIA/IRISA at MediaEval 2011: Learning temporal modality integration with Bayesian Networks," *CEUR Workshop Proc.*, vol. 807, 2011.

[16] X.-F. Liu and X.-X. Zhu, "Parallel Feature Extraction through Preserving Global and Discriminative Property for Kernel-Based Image Classification," *J. Inf. Hiding Multimed. Signal Process.*, vol. 6, no. 5, pp. 977–986.

[17] C. H. Demarty, C. Penet, M. Soleymani, and G. Gravier, "VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation," *Multimed. Tools Appl.*, vol. 74, no. 17, pp. 7379–7404, 2015.

[18] C. H. Demarty, B. Ionescu, Y. G. Jiang, V. L. Quang, M. Schedl, and C. Penet, "Benchmarking violent scenes detection in movies," in *Proceedings - International Workshop on Content-Based Multimedia Indexing*, 2014, pp. 1–6.

[19] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Trans. Multimed.*, vol. 14, no. 1, pp. 88–101, 2012.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe," in *Proceedings of the ACM International Conference on Multimedia - MM '14*, 2014, pp. 675–678.

[21] C.-C. Chang and C.-J. Lin, "Libsvm," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

ATTRIBUTES REPRESENTATION USING DEEP NEURAL NETWORKS FOR VIOLENT SCENES DETECTION

Do Van Tien, Lam Quang Vu, Phan Le Sang, Ngo Duc Thanh, Le Dinh Duy, Duong Anh Duc

ABSTRACT— *Deep Neural Networks (DNN) is a subfield of machine learning algorithms that is based on Artificial Neural Networks for learning multiple levels of representation in order to model complex relationships among data. With so many outstanding results compared with the previous method, several research groups use DNN in many different areas such as image processing, audio processing, natural language processing.*

In this paper, we propose using DNN to represent attributes for violent scenes detection. This is a problem not only highly practical but also the basis to build analytical tools and video content moderated. To evaluate the proposed method, we use some common pre-train model such as Alexnet, UvANet, VGG and experiments conducted on VSD 2014. The experimental results showed that the accuracy when using DNN is 48.12% higher than the best method does not use DNN 13%.

Keywords— *Violent scences detection, deep neural networks, mid level feature.*