

NHẬN DẠNG CẢM XÚC TRONG VIDEO SỬ DỤNG MẠNG NƠ RON TÍCH CHẬP

Nguyễn Thị Thu Hiền^{1*}, Nguyễn Thị Phương Nhung²

¹Trường Đại học Sư phạm – ĐH Thái Nguyên,

²Trường Đại học Kỹ thuật Công nghiệp – ĐH Thái Nguyên

TÓM TẮT

Phân tích và nhận dạng cảm xúc trong video là dựa trên các đặc trưng hình ảnh trong đoạn video để nhận dạng và phân loại cảm xúc thể hiện trong video đó. Việc phân tích đòi hỏi phải trích xuất được các đặc trưng phù hợp từ dữ liệu hình ảnh. Dữ liệu video thường có kích thước lớn đòi hỏi khả năng xử lý tính toán mạnh cũng như thuật toán phù hợp và hiệu quả, đặc biệt khi yêu cầu đòi hỏi tính toán trong thời gian thực. Trong nghiên cứu này, chúng tôi đề xuất một phương pháp trích xuất dữ liệu nhận dạng mới phù hợp với yêu cầu phân lớp và nhận dạng cảm xúc trong video. Đồng thời, chúng tôi cũng đề xuất một mô hình phân lớp dựa trên ứng dụng và cải tiến mô hình học sâu tiên tiến hiện nay là mạng nơ ron tích chập. Hiệu quả của các đề xuất mới được kiểm nghiệm bằng thực nghiệm và cho thấy kết quả tốt hơn so với các mô hình học máy truyền thống.

Từ khóa: nhận dạng cảm xúc; phân loại cảm xúc; trích chọn đặc trưng; mạng nơ ron tích chập; học sâu

MỞ ĐẦU

Ngày nay, cùng với sự phát triển vượt bậc của mạng Internet và các mạng xã hội như Facebook, Flickr, YouTube, .v.v, người dùng tải lên rất nhiều dữ liệu hình ảnh như các bức ảnh, các đoạn video. Các dữ liệu này không chỉ chứa đựng các thông tin quan điểm cụ thể của người dùng mà còn thể hiện trạng thái cảm xúc của họ trước các đối tượng cụ thể. Những thông tin cảm xúc này có ý nghĩa rất quan trọng cho những nhà sản xuất, kinh doanh và chính phủ, giúp tối đa hóa lợi ích cho cả hai phía, người dùng và nhà sản xuất. Các nhà sản xuất nội dung, nhà quảng cáo cũng muốn biết hiệu quả tác động đến cảm xúc của người dùng của những đặc trưng hình ảnh trên các bức ảnh, video, qua đó giúp tạo ra những sản phẩm hình ảnh có hiệu quả tác động lớn hơn.

Trong vài năm trở lại đây, với sự bùng nổ của lượng dữ liệu cũng như sự phát triển mạnh mẽ của hiệu năng tính toán, việc tích hợp các thông tin hình ảnh trong các nghiên cứu nhận dạng cảm xúc đã trở nên khả thi [1][2]. Một số nghiên cứu đã bắt đầu thực hiện dự đoán trạng thái cảm xúc dựa trên những đặc trưng

hình ảnh của các bức ảnh [2],[3],[4] và video [5]. Phân tích cảm xúc dựa trên hình ảnh gặp phải thách thức lớn hơn so với nhận dạng đối tượng trên hình ảnh. Công việc này đòi hỏi mức độ cao hơn về trừu tượng hóa và khái quát hóa các đặc trưng [6], qua đó giúp mô hình nhận dạng có thể thích nghi với bất kỳ bức ảnh hoặc video nào mà không phụ thuộc vào nội dung của chúng. Bên cạnh đó, những đặc trưng liên quan đến cảm xúc đôi khi là bất định và không rõ ràng do bản chất tự nhiên của cảm xúc.

Mạng nơ ron tích chập (Convolutional Neural Network - CNN) là mô hình học sâu được phát triển mạnh gần đây. Mạng CNN đã đạt được những kết quả vượt bậc trong hầu hết các lĩnh vực nhận dạng và phân lớp, đặc biệt là trong xử lý ảnh, xử lý tiếng nói và xử lý ngôn ngữ tự nhiên [7]. Mạng học sâu CNN có khả năng trích xuất những đặc trưng bậc cao từ dữ liệu đầu vào nhờ tính năng khái quát hóa và trừu tượng hóa của các lớp tích chập bên trong mạng. Tuy nhiên, mạng CNN có nhược điểm là đòi hỏi lượng dữ liệu huấn luyện rất lớn để tránh hiện tượng overfitting. Đồng thời CNN cũng đòi hỏi khả năng tính toán lớn của hệ thống tính toán do số lượng rất lớn các tham số của mạng. Một số các

*Tel: 0982 203129, Email: thuhiensptn@gmail.com

nghiên cứu đã áp dụng các mạng CNN cho bài toán nhận dạng cảm xúc trong ảnh [3], [8], [9] cũng như trong video [10], [11]. Các nghiên cứu này đều dựa trên nội dung của các hình ảnh để xác định cảm xúc của chúng. Để nhận dạng được nội dung hình ảnh, các mạng học sâu có cấu trúc rất phức tạp đã được sử dụng. Các mô hình này đều được huấn luyện trước bằng các tập huấn luyện rất lớn sẵn có. Cách tiếp cận này đòi hỏi hệ thống tính toán có hiệu năng cao cũng như thời gian huấn luyện mạng lớn.

Trong nghiên cứu [12], các tác giả đã đề xuất một phương pháp mới sử dụng các đặc trưng về màu sắc và hướng trong video, kết hợp với bộ nhận dạng sử dụng mạng nơ ron thích nghi mờ (ANFIS) cho kết quả nhận dạng tương đối tốt.

Trong nghiên cứu này, chúng tôi đề xuất một cách tiếp cận mới trong nhận dạng cảm xúc dựa trên hình ảnh. Đầu tiên, chúng tôi cải tiến phương pháp trích chọn đặc trưng dựa trên màu sắc và hướng đã áp dụng trong [12], qua đó có thể áp dụng mô hình mạng nơ ron tích chập để nhận dạng. Sau đó, chúng tôi đề xuất một mô hình mạng nơ ron tích chập phù hợp với tập đặc trưng đã trích chọn. Kết quả thực nghiệm cho thấy mô hình nhận dạng được đề xuất kết hợp với tập đặc trưng cải tiến đã cho độ chính xác nhận dạng cao hơn.

Bài báo được cấu trúc theo các phần như sau: Phần tiếp theo sẽ trình bày về phương pháp trích chọn đặc trưng hình ảnh từ tập dữ liệu video huấn luyện. Phần sau đó sẽ trình bày về mô hình mạng CNN và biến thể sử dụng trong nghiên cứu này. Phần kết quả thực nghiệm và thảo luận sẽ đưa ra so sánh kết quả của mô hình với các kết quả đã có. Cuối cùng sẽ là phần kết luận.

TRÍCH CHỌN ĐẶC TRƯNG HÌNH ẢNH

Thông tin hình ảnh trong video hay các bức ảnh liên tiếp có thể tác động đến cảm xúc của người xem. Dựa trên nghiên cứu [13], có mối liên quan mật thiết giữa sự thay đổi cảm xúc và hướng của các đường thẳng trong ảnh.

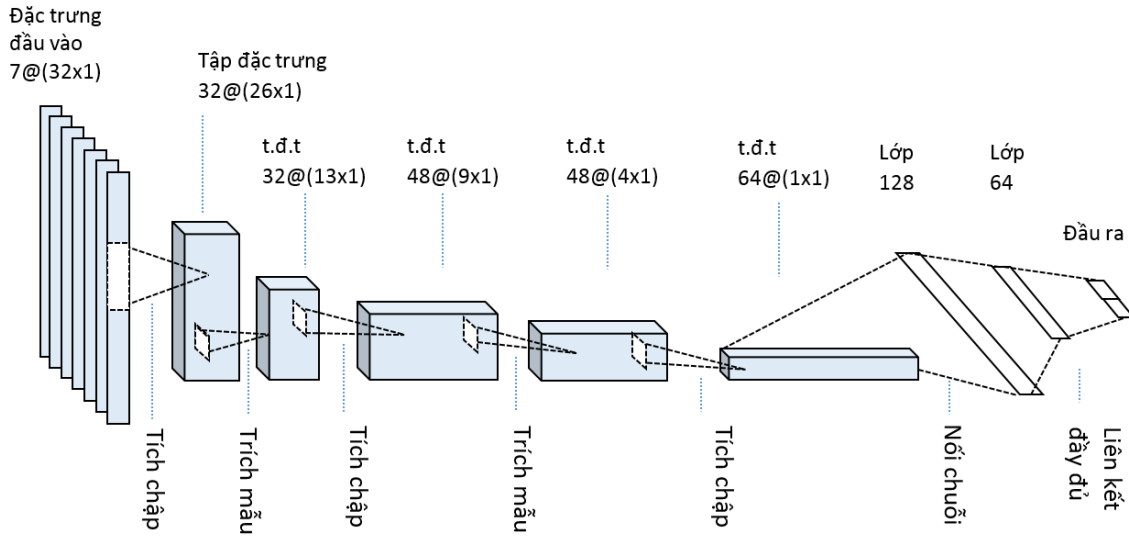
Thông tin màu sắc và thông tin hình ảnh đều là các yếu tố chính được sử dụng để tạo nên một tensor đầu vào từ một đoạn phim. Một tensor chính là một vector có số chiều lớn hơn 2. Để trích xuất thông tin định hướng chúng tôi sử dụng hàm Gabor [14] được mô tả bởi phương trình (1). Độ lớn đầu ra của các bộ lọc định hướng đa cấp là các tensor đặc trưng theo hướng.

$$F(x, y) = e^{-\left(\frac{u_0^2}{\sigma^2} + \frac{v^2 y_0^2}{2\sigma^2}\right)} \cos\left(2\pi \frac{u_0}{\lambda} + \varphi\right) \quad (1)$$

Không gian màu H*S*I* (viết tắt của Hue, Saturation và Intensity) [15] được chọn cho đặc trưng màu sắc. Để giảm chi phí tính toán chúng tôi áp dụng việc trích mẫu cho cả hai tính năng màu sắc và định hướng. Sau khi thu được ba hình ảnh của không gian màu H, S, I và bốn ma trận về hướng, đối với mỗi khung hình của một đoạn video chúng ta có bảy ma trận. Nghiên cứu [12] áp dụng phương pháp phân cụm mờ (FCM) để trích xuất các đặc trưng phù hợp với bộ phân lớp. Tuy nhiên, trong nghiên cứu này, để phát huy hiệu quả nhất khả năng trích xuất đặc trưng bậc cao của mạng CNN, chúng tôi cải tiến bằng cách áp dụng biến đổi histogram để chuyển các ma trận đầu vào sang miền tần số. Do đó, bằng cách tính biểu đồ histogram của mỗi ma trận, chúng tôi xây dựng bảy vector đặc trưng đầu vào cho mỗi khung hình của đoạn video. Chi tiết các tham số của quá trình trích chọn đặc trưng đầu vào sẽ được trình bày trong phần thực nghiệm.

MÔ HÌNH BỘ NHẬN DẠNG DÙNG CNN

Mô hình mạng CNN sử dụng trong nghiên cứu được thể hiện ở hình 1, trong đó bao gồm 5 lớp chính. Đầu tiên, dữ liệu đầu vào được đưa qua các lớp tích chập (Convolutional layer). Các bộ lọc W ở các lớp tích chập sẽ trích xuất các đặc trưng (feature) bậc cao z từ dữ liệu thô đầu vào x theo phương trình (1). Trong đó m và n là kích thước của bộ lọc W .



Hình 1. Mô hình mạng nơ ron tích chập nhận dạng cảm xúc trong video

$$z_{ij} = \sum_{a=0}^{m-1} \sum_{b=0}^{n-1} W x_{(i+a)(j+b)} \quad (2)$$

Đầu ra của lớp tích chập được phi tuyến hóa sử dụng các hàm kích hoạt $\sigma(\cdot)$. Các hàm kích hoạt thường được sử dụng là hàm Sigmoid, hàm Tanh hoặc ReLU. Trong mô hình này hàm ReLU (Rectifier Linear Unit) được sử dụng. Phương trình (2) mô tả hoạt động của hàm ReLU.

$$f(x) = \begin{cases} x & (\text{if } x > 0) \\ 0 & (\text{if } x \leq 0) \end{cases} \quad (3)$$

Sau mỗi lớp tích chập là một lớp pooling (trích mẫu) như mô tả bởi phương trình (3). Lớp pooling có vai trò làm giảm kích thước của các tập đặc trưng nhưng vẫn giữ lại những đặc trưng quan trọng nhất.

$$y_{ij} = \max_{|k| < \tau_i, |l| < \tau_j} x_{i-k, j-l} \quad (4)$$

Trong mô hình này chúng tôi sử dụng ba lớp tích chập và hai lớp max-pooling. Cuối cùng, đầu ra của mô hình là một lớp liên kết đầy đủ (Fully Connected - FC). Lớp FC hoạt động như một bộ phân lớp có cấu trúc như một mạng nơ ron truyền thẳng. Bộ phân lớp này sẽ dựa trên tập các đặc trưng đã được kết xuất từ các lớp tích chập trước đó để phân lớp dữ liệu đầu vào thành các lớp đầu ra tương ứng.

THỰC NGHIỆM VÀ KẾT QUẢ

Tiền xử lý dữ liệu

Tập dữ liệu trong thử nghiệm bao gồm các đoạn video được sử dụng làm tác nhân kích thích trong nghiên cứu [12]. Tập dữ liệu này bao gồm 16 đoạn phim có độ dài ba mươi giây. Mỗi đoạn phim được lựa chọn từ các bộ phim và phim tài liệu. Các clip được chia thành hai nhóm video tích cực và video tiêu cực dựa trên điểm số ý kiến trung bình (MOS) thu được từ các đối tượng không tham gia thử nghiệm.

Chúng tôi sử dụng một nhóm các bộ lọc định hướng đa mức để xây dựng các đặc trưng thông tin định hướng. Do đó, có bốn giá trị định hướng khác nhau, cho $0^\circ, 45^\circ, 90^\circ$ và 135° , thu được. Sau đó, tất cả những giá trị định hướng này đã được giảm kích thước từ 320×240 thành 32×32 để có được một tensor đặc trưng có kích thước $32 \times 32 \times 4$. Các đặc trưng thông tin màu sắc của các khung có thể thu được bằng cách chuyển đổi từng khung màu RGB thành không gian màu HSI. Tiếp đó, ba kênh của HSI đã được thay đổi kích thước từ 320×240 xuống 100×100 pixel. Chúng tôi xây dựng tính năng đầu vào cuối cùng bằng cách tính biểu đồ histogram 32 lớp của mỗi kênh trong số bảy kênh, trong đó có bốn kênh là thông tin định hướng, và ba kênh là thông tin màu HSI. Cuối cùng, chúng tôi thu được một tensor đầu vào kích thước

(7×32) cho mỗi khung hình của đoạn video. Với 16 đoạn video, mỗi đoạn có độ dài 30 giây với tỷ lệ mẫu là 24 khung hình/giây, chúng tôi đã thu được các tập các đặc trưng có kích thước $11520 \times (7 \times 32)$. Giá trị của các tensor đầu vào sau đó được chuẩn hóa về dạng số thực trong giải $[0, 255]$.

Các thiết lập thực nghiệm

Thực nghiệm được thực hiện trên hệ thống máy tính có cấu hình cho bởi Bảng 1. Máy tính sử dụng hệ điều hành Ubuntu 16.04 LTS. Chương trình được cài đặt bằng ngôn ngữ lập trình Python và sử dụng thư viện Pytorch. Thời gian huấn luyện và thời gian kiểm thử tương ứng là 41.3 giây và 0.0041 giây.

Bảng 1. Chi tiết cấu hình hệ thống máy tính thực nghiệm

Thiết bị	Tham số
CPU	Intel core i7 CPU 920 2,67GHz.
RAM	RAM: 16GB
GPU	GeForce GTX TITAN X

Cấu trúc và các thiết lập chi tiết của các thông số cho mô hình CNN được minh họa trong Bảng 2.

Chúng tôi áp dụng tác vụ ‘dropout’ trong lớp tích chập cuối để tránh hiện tượng quá khớp (overfitting). Dropout là quá trình vô hiệu hóa một tỉ lệ các nơ ron trong quá trình huấn luyện giúp cho mô hình có khả năng tổng quát hóa (generalizaion) tốt hơn.

Bảng 2. Thiết lập các tham số cho mô hình mạng nơ ron tích chập

Lớp	Tham số
Đầu vào	Huấn luyện: $8580 \times (7 \times 32)$ Kiểm thử: $2860 \times (7 \times 32)$
Tích chập 1	$64 \times (1 \times 7)$, ReLU
Tích chập 2	$96 \times (1 \times 5)$, ReLU
Tích chập 3	$128 \times (1 \times 3)$, ReLU
FC1	256×64 , ReLU, Dropout(0,5)
FC2	64×2

Để so sánh, chúng tôi chọn một mô hình học máy thông dụng hiện nay là Support Vector Machine (SVM). Các thông số tối ưu của SVM được lựa chọn bằng tìm kiếm lưới, đó là: kernel = 'poly'; degree = 3; C = 0,35; coef0 = 0,125, gamma = 0,0625.

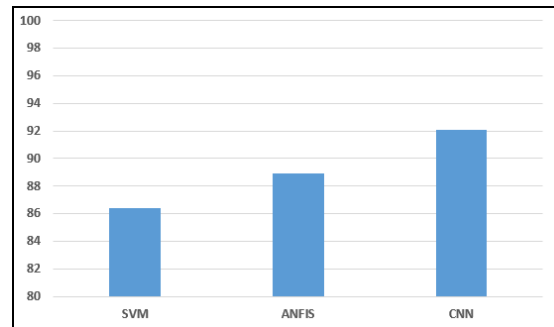
Kết quả và bàn luận

So sánh kết quả thực nghiệm của các mô hình được biểu thị ở Bảng 3 và Hình 2.

Bảng 3. So sánh độ chính xác của các mô hình SVM, ANFIS và CNN

Mô hình	Độ chính xác (%)
SVM	86,38
ANFIS	99,93
CNN	92,09

Độ chính xác của mô hình đề xuất cũng được so sánh với mô hình mạng nơ ron thích nghi mờ (ANFIS) sử dụng trong nghiên cứu [12] có sử dụng cùng tập dữ liệu nhận dạng. Trong nghiên cứu này, các tác giả đã công bố kết quả nhận dạng tốt nhất là 88,93%.



Hình 2. So sánh độ chính xác của các mô hình nhận dạng SVM, ANFIS và CNN cho nhận dạng cảm xúc trong video

Kết quả thực nghiệm cho thấy, mô hình mạng nơ ron tích chập CNN cho độ chính xác nhận dạng vượt trội so với các mô hình nhận dạng được so sánh. Kết quả trên có được là nhờ khả năng trích xuất những đặc trưng hữu ích bậc cao của các lớp tích chập giúp cho việc phân lớp trở nên hiệu quả hơn.

KẾT LUẬN

Bài toán phân loại và nhận dạng cảm xúc trong video là bài toán khó do yêu cầu về trích chọn đặc trưng phù hợp cũng như yêu cầu độ phức tạp tính toán cao. Trong nghiên cứu này chúng tôi đã đề xuất một phương pháp trích chọn đặc trưng mới cho dữ liệu video và cùng với đó là một mô hình bộ nhận dạng mới dựa trên mô hình mạng nơ ron tích chập. Kết quả thực nghiệm cho thấy những đề xuất mới đã cho độ chính xác nhận dạng vượt

trội so với những mô hình học máy truyền thống. Mô hình mới dựa trên CNN cũng cho phép tận dụng khả năng xử lý song song trên bộ xử lý đồ họa GPU làm tăng tốc độ tính toán, nhờ đó có thể áp dụng hệ thống nhận dạng cảm xúc trong video theo thời gian thực. Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu và cải thiện khả năng nhận dạng cảm xúc trong video bằng cách áp dụng các mô hình học sâu tiên tiến hơn, đồng thời có thể kết hợp các loại dữ liệu khác như văn bản, âm thanh cùng với dữ liệu hình ảnh để cải thiện hiệu suất của bộ nhận dạng.

TÀI LIỆU THAM KHẢO

1. X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han (2010), "The wisdom of social multimedia," *Proc. Int. Conf. Multimed. - MM '10*, pp. 1235.
2. J. Yuan, S. McDonough, Q. You, and J. Luo (2013), "Sentribute: image sentiment analysis from a mid-level perspective," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pp. 10.
3. S. Siersdorfer, E. Minack, F. Deng, and J. Hare (2010), "Analyzing and Predicting Sentiment of Images on the Social Web," *Proc. 18th ACM Int. Conf. Multimed.*, pp. 715–718.
4. D. Borth, T. Chen, R.-R. Ji, and S.-F. Chang (2013), "Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," *ACM Int. Conf. Multimed. (ACM MM)*, pp. 459–460.
5. L.-P. Morency, R. Mihalcea, and P. Doshi (2011), "Towards multimodal sentiment analysis," *Proc. 13th Int. Conf. multimodal interfaces - ICMI '11*, pp. 169.
6. D. Joshi *et al.* (2011), "Aesthetics and emotions in images," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115.
7. J. Schmidhuber (2015), "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117.
8. J. Mandhyani, L. Khatri, V. Ludhrani, R. Nagdev, and P. S. Sahu (2017), "Image Sentiment Analysis," *Int. J. Eng. Sci. Comput.*, vol. 7, no. 2, pp. 4566–4569.
9. V. Campos, B. Jou, and X. Giró-i-Nieto (2017), "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vis. Comput.*, vol. 65, pp. 15–22.
10. Z. Cai, D. Cao, and R. Ji, (2015), *Video (GIF) Sentiment Analysis using Large-Scale Mid-Level Ontology*.
11. M. Wollmer *et al.* (2013), "You tube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53.
12. G. Lee, M. Kwon, S. Kavuri Sri, M. Lee, S. Kavuri, and M. Lee (2014), "Emotion recognition based on 3D fuzzy visual and EEG features in movie clips," *Neurocomputing*, vol. 144, pp. 560–568.
13. W. Wang, Y. Yu, and J. Zhang (2005), "A new SVM based emotional classification of image," *J. Electron.*, vol. 22, no. 1, pp. 98–104.
14. a Kumar and G. K. H. Pang (2002), "Defect detection in textured materials using gabor filters," *Ieee Trans. Ind. Appl.*, vol. 38, no. 2, pp. 425–440.
15. W. N. Wang, Y. L. Yu, and S. M. Jiang (2007), "Image retrieval by emotional semantics: A study of emotional space and feature extraction," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 4, pp. 3534–3539.

SUMMARY

**EMOTION IDENTIFICATION IN VIDEO
USING CONVOLUTIONAL NEURAL NETWORK****Nguyen Thi Thu Hien^{1*}, Nguyen Thi Phuong Nhung²**¹*University of Education – TNU,*²*University of Technology - TNU*

Emotion analysis and identification in a video are based on visual features of the video clip to identify and classify the emotion in this clip. The analysis requires sufficient feature extraction from visual data. Video data usually in high dimension and require high-performance computing with an efficient algorithm, especially needed in real-time analysis. In this study, we propose a new feature extraction method suitable for emotion classification in video problem. We also propose a new classification model base on the improvement of modern deep learning is Convolutional Neural Network. The new proposals are verified on an experiment of video emotion classification and show the better performance compared with traditional machine learning.

Keywords: *emotion classification; emotion identification; feature extraction; convolutional neural network; deep learning*

Ngày nhận bài: 14/5/2018; **Ngày phản biện:** 27/5/2018; **Ngày duyệt đăng:** 31/5/2018

* Tel: 0982 203129, Email: thuhiensptn@gmail.com