Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.00036

MỘT MÔ HÌNH HỌC SÂU CHO PHÁT HIỆN CẢM XÚC KHUÔN MẶT

Nguyễn Thị Duyên[1], Trương Xuân Nam[1], Nguyễn Thanh Tùng[1]

[1] Khoa Công nghệ thông tin, Trường ĐH Thủy lợi

duyennt02@wru.vn, namtx@tlu.edu.vn, tungnt@tlu.edu.vn

TÓM TẮT: *Phát hiện cảm xúc khuôn mặt sử dụng các phương pháp học máy là chủ đề quan trọng trong lĩnh vực thị giác máy tính. Trong những năm gần đây, học sâu (Deep learning) đã thể hiện được ưu thế trong bài toán xử lý dữ liệu ảnh, âm thanh cả trong nghiên cứu và công nghiệp. Trong bài báo này, một mô hình học sâu với kiến trúc mạng tích chập được giới thiệu với thiết kế gồm 8 khối chính, trong đó 7 khối mạng tích chập và khối cuối là đầu ra softmax. Kiến trúc này hướng đến việc nhận dạng các thành phần trên mặt và cảm xúc của khuôn mặt. Tập dữ liệu phổ biến về nhận dạng mặt người FER-2013 được dùng trong quá trình thực nghiệm, kết quả cho thấy việc phát hiện cảm xúc khuôn mặt của mô hình đề xuất đạt độ chính xác tương đương với những mô hình tốt nhất đã được công bố.*

Từ khóa: *Học sâu, nhận dạng cảm xúc, cảm xúc khuôn mặt, mạng tích chập.*

I. GIỚI THIỆU

Bài toán phát hiện cảm xúc khuôn mặt đã có lịch sử nghiên cứu lâu dài. Từ năm 1964, Bledsoe [1] là người đầu tiên xây dựng chương trình nhận dạng khuôn mặt tự động kết hợp với hệ thống máy tính, bằng cách phân loại khuôn mặt trên cơ sở mốc chuẩn được nhập vào bằng tay. Các thông số để phân loại là khoảng cách chuẩn, tỉ lệ giữa các điểm như góc, mắt, miệng, chóp mũi và chóp cằm. Sau này, tại Bell Labs đã phát triển một kĩ thuật dựa trên vector với 21 thuộc tính khuôn mặt được phát hiện bằng cách sử dụng kỹ thuật phân loại tiêu chuẩn mẫu. Các thuộc tính được lựa chọn đánh giá chủ yếu là: màu tóc, chiều dài của đôi tai, độ dày môi... Năm 1986, hệ thống WISARD dựa trên mạng nơron đã có thể nhận biết được tình trạng và biểu cảm khuôn mặt một cách hạn chế.

Phát hiện cảm xúc khuôn mặt là bước phát triển tiếp sau của việc phát hiện khuôn mặt, tuy nhiên có nhiều quan điểm trong việc định nghĩa khái niệm cảm xúc, vốn rất không rõ ràng. Matsumoto [2] phân chia cảm xúc khuôn mặt thành 7 nhóm thể hiện chính: Vui vẻ, Ngạc nhiên, Hài lòng, Buồn bực, Cáu giận, Phẫn nộ và Sợ hãi. Tuy nhiên, nhóm của Mase và Pentland [3] cho rằng chỉ 4 loại cảm xúc được thể hiện một cách rõ ràng là Hạnh phúc, Ngạc nhiên, Giận giữ và Căm phẫn; các loại cảm xúc khác thường không rõ ràng và tùy thuộc nhiều vào kinh nghiệm của người quan sát (tức là không thể định lượng một cách chính xác). Cơ sở dữ liệu Radboud Faces Database thì phân chia cảm xúc khuôn mặt thành 8 loại: Tức giận, Căm phẫn, Sợ hãi, Hạnh phúc, Buồn rầu, Bất ngờ, Khinh miệt và Trung lập. Dataset Kaggle FER-F2013 [4] thì lại chỉ có 7 loại cảm xúc: Giận dữ, Căm phẫn, Sợ hãi, Hạnh phúc, Buồn rầu, Bất ngờ và Trung lập.

Do việc định nghĩa khái niệm cảm xúc tương đối mờ, nên việc đánh giá chất lượng các phương pháp phát hiện cảm xúc rất tùy thuộc vào tập dữ liệu huấn luyện và kiểm tra. Ví dụ như báo cáo của Mase [3] đề xuất phương pháp nhận diện cảm xúc dựa trên các đặc trưng chuyển động cơ mặt và chỉ sử dụng phương pháp K-láng giềng gần nhất nhưng đạt mức độ chính xác lên đến 88%. Trong khi đó, với tập dữ liệu FER-2013, phương pháp tốt nhất hiện nay sử dụng RBM (máy boltzmann hạn chế) đạt độ chính xác 71%, mô hình này sử dụng khoảng 5 triệu tham số [4], các phương pháp còn lại đều cho kết quả dưới 70%.

Trong bài báo này, chúng tôi thử nghiệm một kiến trúc học sâu dựa trên nhiều lớp tích chập (ConvNet) để phát hiện cảm xúc khuôn mặt. Dữ liệu thu được từ webcam sẽ được định vị khuôn mặt bằng phương pháp haar cascade [5] từ thư viện OpenCV [6], sau đó dữ liệu được chuyển vào mạng học sâu với đầu ra xác suất (softmax), trả về xác suất của 7 loại cảm xúc do hệ thống tính toán được. Kết quả thử nghiệm trên bộ dữ liệu FER-2013 đạt 66.3%, nằm trong TOP5 mô hình học máy tốt nhất của dataset này.

II. HỌC SÂU VÀ BÀI TOÁN PHÁT HIỆN CẢM XÚC KHUÔN MẶT

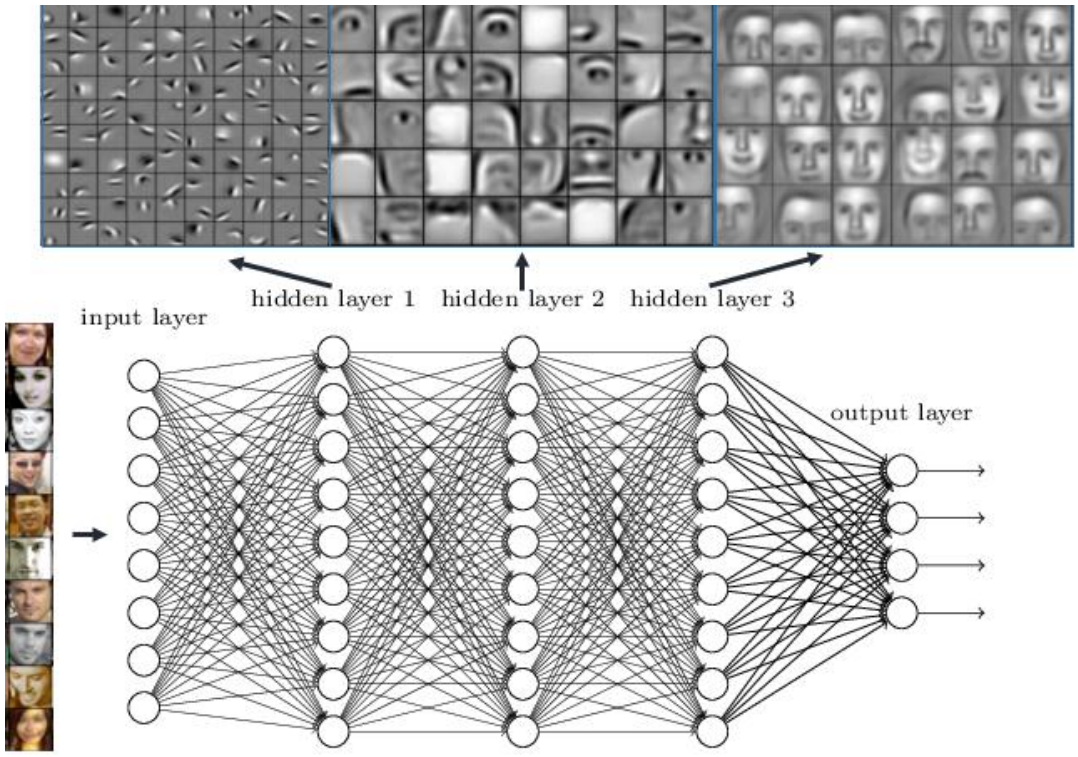
A. Học sâu (deep learning)

Học sâu (deep learning) là tập các thuật toán học máy với ý tưởng xây dựng mô hình dữ liệu có mức độ trừu tượng cao dựa trên các dữ liệu có mức độ trừu tượng hóa thấp hơn, bằng cách phân lớp dữ liệu và các biến đổi phi tuyến [10].

Nghiên cứu từ rất lâu cho thấy mạng nơron được chứng minh khả năng xấp xỉ vạn năng chỉ với không quá 4 lớp, nhưng chưa có phương pháp nào cụ thể ước lượng số nơron cần thiết trên mỗi lớp [10]. Việc nghiên cứu về các mạng có số lớp lớn chỉ trở nên phổ biến sau thành công của mạng AlexNet khi mô hình này thắng giải ImageNet 2012 với khoảng cách rất xa so với các mô hình cạnh tranh [11], mặc dù kiến trúc CNN đã được LeCun giới thiệu từ trước đó rất lâu [12].

Ngoài kiến trúc CNN, các mô hình mạng học sâu còn nhiều dạng kiến trúc khác như các lớp truyền thẳng kết nối đầy đủ (fully connected layer), RNN, LSTM, GRU, DBN,…[10]

Hình 1 biểu diễn một mô hình học sâu tiêu biểu [13] sử dụng trong nhận dạng mặt người, trong đó dữ liệu đầu vào của mạng có thể là dữ liệu ở dạng thô nhất là các điểm ảnh RGB (thậm chí không cần qua tiền xử lý). Các đặc trưng được tổ hợp và tạo thành các chi tiết nhỏ ở lớp ẩn đầu tiên, sau đó tiếp tục được tái tạo và tổ hợp mức chi tiết lớn ở lớp ẩn thứ hai, và cuối cùng các hình ảnh đặc trưng của toàn bộ khuôn mặt ở lớp ẩn thứ 3. Lớp output cho ra đánh giá xác suất khuôn mặt thuộc phân lớp nào (người nào).



Hình 1. Một mô hình học sâu trong nhận dạng mặt người

Một mô hình học sâu thường có 3 nhiệm vụ được kết hợp trong một kiến trúc mạng duy nhất:

- Các lớp đặc trưng (features): có nhiệm vụ chuyển đổi các đặc trưng thành dạng dữ liệu phù hợp để xử lý, chẳng hạn như các tầng tích chập (convolution), mẫu (subsampling), pooling,…
- Các lớp mô hình (modeling): sử dụng các thuật toán học để khái quát hóa dữ liệu, chẳng hạn nơron network, restricted BM, DBN, autoencoder,…
- Các lớp giải mã (decoding): dựa trên dữ liệu khái quát biến đổi thành đầu ra (markov random field hoặc những công cụ tương tự).

Các mạng học sâu đều có cấu trúc xác định trước, như vậy bài toán tập huấn vẫn là việc xác định giá trị các tham số trên mạng. Hiện chưa có phương pháp tập huấn nào cho phép điều chỉnh cấu trúc mạng hiệu quả.

B. Bài toán phát hiện cảm xúc khuôn mặt

Đây là một bài toán phân lớp tương đối tiêu chuẩn, đã được nghiên cứu trong một thời gian khá dài. Một hệ thống nhận diện cảm xúc khuôn mặt thường được triển khai gồm 3 bước.

1. Nhận ảnh và tiền xử lý: Ảnh khuôn mặt được lấy từ nguồn dữ liệu tĩnh (chẳng hạn như từ file, database), hoặc động (từ livestream, webcam, camera,…), nguồn dữ liệu này có thể trải qua một số bước tiền xử lý nhằm tăng chất lượng hình ảnh để giúp việc phát hiện cảm xúc trở nên hiệu quả hơn.

2. Trích xuất các đặc trưng: Bước rất quan trọng, đặc biệt với các phương pháp truyền thống, các đặc trưng khuôn mặt được tính toán dựa trên các thuật toán có sẵn, kết quả thường là một vector đặc trưng làm đầu vào cho bước sau.

3. Phân lớp và nhận diện cảm xúc: Đây là một bài toán phân lớp điển hình, rất nhiều các thuật toán có thể áp dụng trong bước này như KNN, SVM, LDA, HMM,…

Một vấn đề lớn đối với bài toán phát hiện cảm xúc khuôn mặt là thiếu sót các dataset tiêu chuẩn đủ lớn và sự chuẩn hóa các loại cảm xúc. Một trong những dataset đầu tiên cho bài toán này (năm 2009) là CK+ chỉ có 593 loạt ảnh,

bộ dataset MMI cũng chỉ có 740 ảnh và 2900 video. Một số dataset xuất hiện gần đây có số lượng mẫu lớn hơn như EmotionNet [14] có 1 triệu mẫu hoặc AffectNet [15] có 450 nghìn mẫu. Các dataset cũng có nhiều khác biệt nhau về số lượng và cách phân loại cảm xúc, cũng như cách tính hiệu suất của các phương pháp phân loại.

C. Ứng dụng học sâu vào bài toán phát hiện cảm xúc khuôn mặt

Các mạng học sâu được ứng dụng rộng rãi vào bài toán phát hiện cảm xúc khuôn mặt, đặc biệt các loại mạng phù hợp với việc xử lý dữ liệu hình ảnh như CNN, DBN (deep belief network), DAE (deep autoencoder). Ngoài ra, một số tác giả sử dụng kết quả của các pre-trained model như AlexNet, VGG-face, GoogleNet,… và sử dụng các đặc trưng được trích xuất từ các mô hình này làm đầu vào cho hệ thống phân loại của họ [16].

Tuy được ứng dụng rộng rãi, nhưng bài toán phát hiện cảm xúc khuôn mặt vẫn là một thách thức lớn vì độ chính xác của những hệ thống hiện nay vẫn còn khá thấp; chẳng hạn như mô hình CNN của Liu et al. [17] cho dataset MMI mới đạt khoảng 78,5 % (tốt nhất cho dataset này); mô hình kết hợp VGG16-LSTM của Vielzeuf et al. [18] cho dataset AffectNet mới đạt được 48,6 % (tốt nhất cho dataset này).

III. MÔ HÌNH ĐỀ XUẤT VÀ KẾT QUẢ THỰC NGHIỆM

A. Dataset FER-2013

Dữ liệu FER-2013 được công bố bởi trang Kaggle trong khuôn khổ workshop của hội thảo ICML 2013. Dữ liệu gồm các ảnh đa cấp xám cỡ 48x48 chỉ gồm khuôn mặt hầu như được căn giữa ảnh và tỉ lệ khuôn mặt được điều chỉnh chiếm phần lớn diện tích của ảnh. Một ảnh sẽ được gán nhãn nằm một trong bảy loại cảm xúc giá trị từ 0 đến 6 (0: giận dữ, 1: căm phẫn, 2: sợ hãi, 3: hạnh phúc, 4: buồn rầu, 5: bất ngờ, 6: trung lập).

Bộ dữ liệu này gồm 28.709 mẫu huấn luyện, mẫu kiểm tra công khai có 3.589 ảnh. Khi thực hiện đánh giá mô hình, Kaggle sẽ sử dụng một bộ kiểm tra khác cũng có 3.589 ảnh, vì vậy kết quả đánh giá của ban giám khảo có thể có sai lệch so với sử dụng bộ test công khai, một số trường hợp đặc biệt sai lệch có thể lên đến 5% [4].

Chúng tôi sử dụng bộ dữ liệu này cho mô hình thử nghiệm vì bộ dữ liệu có số mẫu khá lớn, phù hợp với việc huấn luyện với mạng học sâu, vốn đòi hỏi nhiều mẫu hơn các phương pháp học máy thông thường. Ngoài ra, bộ dữ liệu được cấu trúc dễ dàng xử lý bởi thư viện Keras/TensorFlow và có nhiều kết quả đối chứng khi thực hiện so sánh mô hình của chúng tôi với các kết quả của những nhóm nghiên cứu khác.

B. Mô hình đề xuất

Kiến trúc đề xuất của chúng tôi gồm 8 khối chính được thể hiện tại Hình 4, trong đó có 7 khối CNN và khối cuối là đầu ra softmax, xem tại Hình 2. Đầu tiên, ảnh 48x48 đa cấp xám được chuyển vào khối A, khối có 32 filter, sử dụng kernel filter cỡ 3x3, hàm kích hoạt ReLU, kết quả tính toán được chuyển qua một lớp batch normalization. Khối A được thiết kế với ý đồ tạo ra 32 đặc trưng cơ bản cho việc phát hiện cảm xúc khuôn mặt. Khối B được thiết kế tương tự khối A, ngoại trừ việc sử dụng 64 filter, mục tiêu của khối này giúp tổ hợp các đặc trưng cơ bản thành các đặc trưng phức tạp hơn.

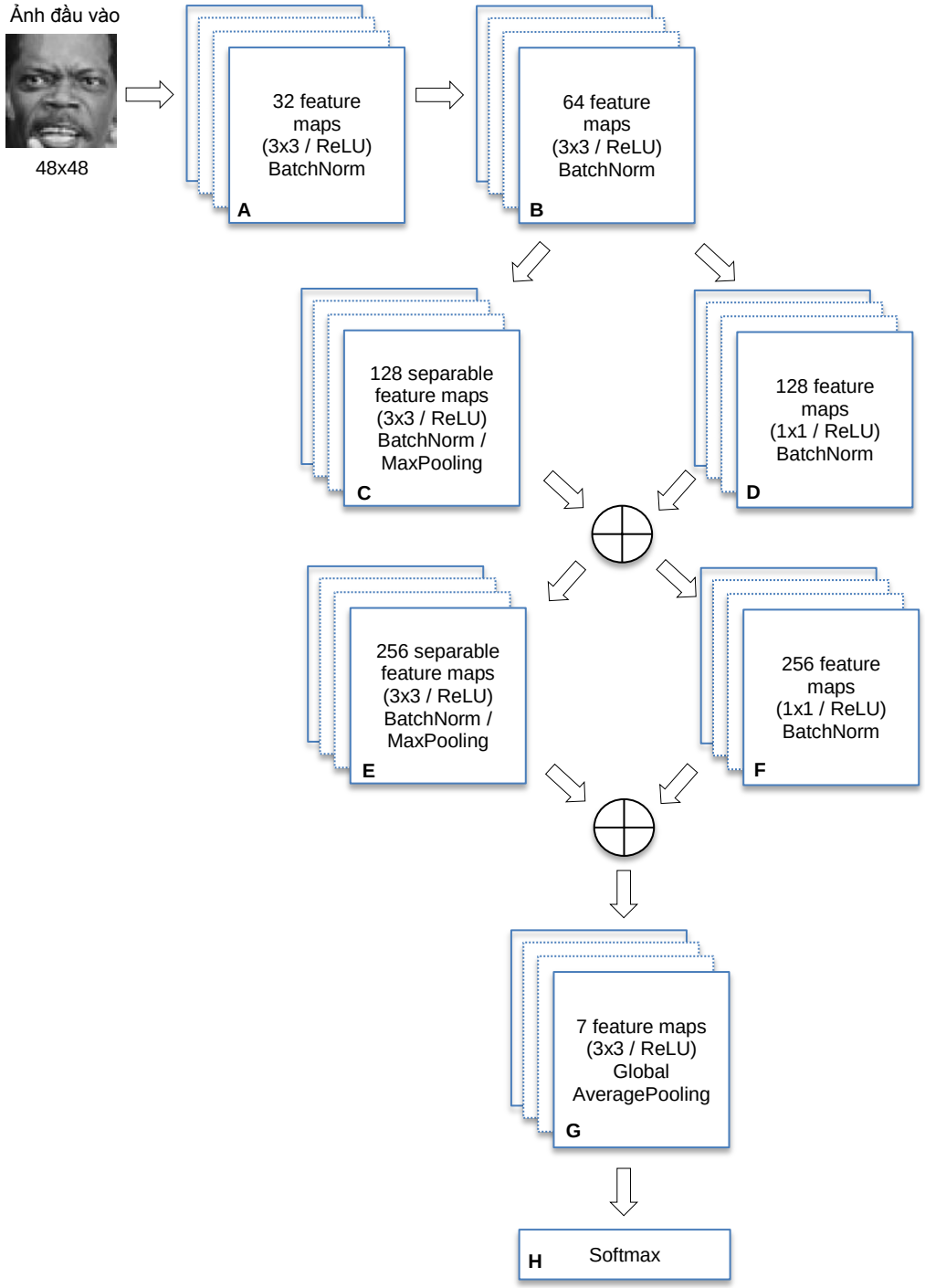
Kết quả đầu ra khối B được xử lý độc lập trong 2 khối C và D, khối C là một depthwise separable CNN 128 filter [7], sau đó được chuẩn hóa bởi một lớp batch normalization và max pooling. Khối D chỉ là một filter nhằm điều chỉnh trọng số của đặc trưng khi tính gộp kết quả với khối C. Khối E và F cũng được thiết kế tương tự như vậy.

Cuối cùng, chúng tôi sử dụng khối F có 7 filter (tương ứng với 7 loại cảm xúc), kết quả tính toán của CNN được chuyển vào một global average pooling (chuyển kết quả 2D thành vector), kết quả này được xử lý qua một lớp softmax để trả về xác suất của từng loại cảm xúc.

Mạng được huấn luyện end-to-end với batch_size = 128, epochs = 100. Sau 70 lượt huấn luyện hầu như kết quả trên tập test không thay đổi.

Kết quả thử nghiệm trên dữ liệu kiểm tra đạt mức độ chính xác khoảng 66.3% (trung bình 5 lần huấn luyện). Trong quá trình huấn luyện độ chính xác thường xuyên cao hơn kết quả kiểm nghiệm trên bộ kiểm tra, nhưng không quá sai khác.

Mô hình sau khi huấn luyện cũng được kiểm tra với dữ liệu ngẫu nhiên từ dataset CK+ [8] và RaFD [9] với kết quả khoảng 61% và 52% (kết quả với RaFD thấp hơn một chút có lẽ vì bộ dữ liệu này có ảnh không chụp thẳng mặt).



Hình 2. Kiến trúc CNN đề xuất dùng cho việc phát hiện cảm xúc khuôn mặt

C. Kết quả thực nghiệm

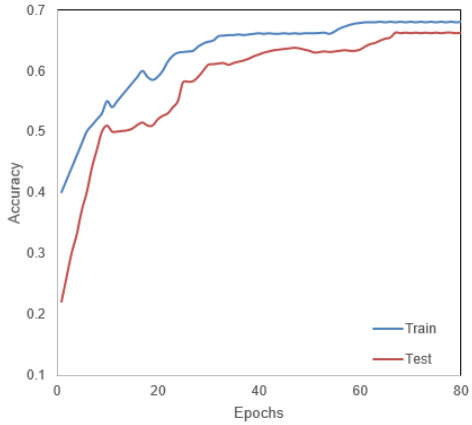
Để triển khai huấn luyện và thử nghiệm mô hình đề xuất, ngôn ngữ Python và thư viện Keras/TensorFlow được sử dụng cho việc xây dựng mô hình mạng CNN. Dữ liệu FER-2013 được tiền xử lý không đáng kể, ngoại trừ việc chuyển đổi đa cấp xám từ dạng số nguyên 0 đến 255 về miền số thực [0, 1] nhằm hỗ trợ tốt hơn cho dữ liệu đầu vào của mạng tích chập.

Ngôn ngữ Python kết hợp thêm OpenCV cũng được sử dụng để viết chương trình minh họa hỗ trợ cho việc xử lý dữ liệu đầu vào từ webcam/camera. Quá trình xử lý qua 5 bước như sau:

1. Ảnh đầu vào được chuyển thành đa cấp xám;
2. Dùng haar cascade (OpenCV) tìm kiếm vùng mặt người trên ảnh đầu vào;
3. Vùng ảnh mặt người được chuyển đổi về kích thước 48x48;
4. Ảnh 48x48 đa cấp xám chuyển đổi về miền [0, 1] sau đó đưa vào mô hình CNN;
5. Đầu ra của CNN là xác suất của các cảm xúc, chọn cảm xúc có xác xuất cao nhất làm kết quả cuối cùng.

Tất cả các thí nghiệm được chạy trên máy trạm sử dụng bộ xử lý Intel i9-7920X, RAM 64 GB và GPU GTX 1080 Ti, hệ điều hành Ubuntu 18.04; các thư viện hỗ trợ Keras 2.2.4, TensorFlow 1.12, CUDA 10.0.130, cuDNN 7.4.1.

Mô hình được huấn luyện với epochs = 100, tuy nhiên kết quả về độ chính xác trên tập huấn luyện và tập kiểm tra gần như ổn định sau bước 70 khi kiểm nghiệm thực tế. Độ chính xác trên tập dữ liệu kiểm tra không bị giảm sau khi mạng đã ổn định, như vậy có thể thấy mô hình không bị hiện tượng quá khớp. Muốn tăng độ chính xác của mô hình, chúng tôi điều chỉnh phù hợp về số filter trên mỗi lớp và có thể tăng thêm một số lớp ẩn trong mạng CNN nhằm tăng khả năng nhận biết các cấu trúc phức tạp trên khuôn mặt.



Hình 3. Biến động về độ chính xác của mô hình trên tập huấn luyện và tập kiểm tra theo số lượt huấn luyện

Kết quả thử nghiệm thực tế cho thấy mô hình khá nhạy khi nhận biết cảm xúc hạnh phúc (happy), khá kém với cảm xúc căm phẫn (disgust). Việc hầu hết các mô hình được công bố với tập dữ liệu FER-2013 đều chỉ đạt độ chính xác thấp (dưới 70%), điều này có thể cho thấy bộ dữ liệu này có những yếu tố mất cân bằng hoặc nhiễu khi gán nhãn dữ liệu.

Hình 3. Một kết quả phát hiện cảm xúc khuôn mặt

IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, một kiến trúc mô hình CNN ứng dụng vào việc phát hiện cảm xúc khuôn mặt người được đề xuất, tập ảnh 48x48 điểm ảnh đa cấp xám được sử dụng trong thực nghiệm để đánh giá mô hình. Đây là một bài toán có tính ứng dụng cao có thể áp dụng trong nhiều vấn đề thực tế, đặc biệt liên quan đến việc cảm nhận phản hồi của khách hàng. Tuy chất lượng phân loại cảm xúc khuôn mặt người chưa cao, tuy nhiên lợi thế là mô hình không quá lớn (dưới 4 MB) nên có thể tiếp tục tối ưu để triển khai trên các thiết bị cầm tay, mô hình cũng có thể cài đặt trên các thiết bị nhúng vốn yêu cầu chặt chẽ về bộ nhớ.

Trong thời gian tới, chúng tôi sẽ tập trung vào việc nâng cao chất lượng của mô hình, có thể chuyển đổi sang ứng dụng GAN thay vì chỉ sử dụng thuần CNN. Ngoài ra, có thể bổ sung các lớp mạng hỗ trợ khả năng phân biệt giới tính cùng lúc với cảm xúc (nghiên cứu cho thấy cảm xúc khuôn mặt là khác nhau với giới tính nam hoặc nữ), điều này cũng sẽ tăng độ chính xác của mô hình.

Việc mô hình được huấn luyện trên một dataset nhưng làm việc được với các dataset khác cho thấy mô hình đã học được các đặc trưng phù hợp với khuôn mặt người; tuy nhiên hầu hết các dataset hiện nay đều sử dụng các khuôn mặt người phương Tây, chúng tôi sẽ tiến hành xây dựng dataset bổ sung với khuôn mặt người châu Á, để phong phú thêm dữ liệu huấn luyện và nâng cao chất lượng nhận dạng.

TÀI LIỆU THAM KHẢO

[1] Bledsoe, W. W (1964). "The Model Method in Facial Recognition", *Technical Report PRI 15*. Panoramic Research, Inc., Palo Alto, California.

[2] Matsumoto, David, and Hyi Sung Hwang (2011). "Reading facial expressions of emotion", *Psychological Science Agenda*, Vol 25, No5, pp. 10-18.

[3] K. Mase, A. Pentland (1991), "Recognition of facial expression from optical flow", *IEEE TRANSACTIONS on Information and Systems*, Vol E74-D, No10, pp. 3474-3483.

[4] I Goodfellow, D Erhan, PL Carrier, A Courville, M Mirza, B Hamner, W Cukierski, Y Tang, DH Lee, Y Zhou, C Ramaiah, F Feng, R Li, X Wang, D Athanasakis, J Shawe-Taylor, M Milakov, J Park, R Ionescu, M Popescu, C Grozea, J Bergstra, J Xie, L Romaszko, B Xu, Z Chuang, and Y. Bengio (2013). "Challenges in Representation Learning: A report on three machine learning contests." arXiv 2013.

[5] Paul Viola and Michael Jones (2001). "Rapid Object Detection using a Boosted Cascade of Simple Features", Conference on Computer vision and Pattern recognition 2001.

[6] Docs, OpenCV. "Face Detection Using Haar Cascades.", OpenCV: Face Detection Using Haar Cascades, 4 Aug. 2017.

[7] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". arXiv 2017.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression", *Proceedings of IEEE on CVPR for Human Communicative Behavior Analysis*, San Francisco, USA, 2010.

[9] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., & van Knippenberg, A. (2010). *Presentation and validation of the Radboud Faces Database*. Cognition & Emotion, 24(8), 1377-1388. DOI: 10.1080/02699930903485076.

[10] Bengio, Yoshua. "Learning Deep Architectures for AI". *Foundations and Trends in Machine Learning: Vol. 2: No. 1*, pp 1-127, (2009).

[11] Krizhevsky, Alex. "ImageNet Classification with Deep Convolutional Neural Networks". Retrieved 17 November 2013.

[12] Y LeCun, L Bottou, Y Bengio, P Haffner (1998). "Gradient-based learning applied to document recognition", Proceedings of the IEEE 86 (11), p2278-2324.

[13] Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Y. Ng, "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations", ICML 2009.

[14] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016.

[15] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. PP, no. 99, pp. 1-1, 2017.

[16] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C¸ . Gulc¸ehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari et al., "Combining modality specific deep neural networks for emotion recognition in video," in Proceedings of the 15th ACM on International conference on multimodal interaction. ACM, 2013, pp. 543-550.

[17] X. Liu, B. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 522-531.

[18] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017, pp. 569-576.

FACIAL EMOTION RECOGNITION USING DEEP LEARNING

Nguyen Thi Duyen, Truong Xuan Nam, Nguyen Thanh Tung

ABSTRACT: *Facial emotion recognition plays an important role for the fields of computer vision and artificial intelligence. Deep learning models have shown the best results for dealing with supervised and unsupervised problems in both research and industry recent years. In this paper, a convolutional network architecture with 8 blocks is presented, the final block is the outcome softmax. This architecture is designed for the facial emotion recognition. The FER-2013 dataset has been used for conducting our experiments. The results show that the our Deep learning architecture provide a potential results when compared with the best Deep learning models on this kind of the dataset for human face recognition.*