Современные подходы распознавания человеческих эмоций с помощью глубоких нейронных сетей

Г.А. Альгашев
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
algashev@live.com

А.О. Корепанов
Самарский национальный исследовательский университет им. академика С.П. Королева,
Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН
Самара, Россия
andrew.korepanov@gmail.com

А.В. Никоноров
Самарский национальный исследовательский университет им. академика С.П. Королева,
Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН
Самара, Россия
artniko@gmail.com

***Аннотация* — В данной работе рассмотрены актуальные дискретные и многомерные модели классификации эмоций, а также исследованы основные подходы для построения систем распознавания человеческих эмоций. Рассмотрены современные способы анализа видео и аудио информации для решения задачи распознавания эмоций с помощью глубоких нейронных сетей.**

Ключевые слова — распознавание эмоций, классификация эмоций, нейронные сети, свёрточные сети, рекуррентные сети.

1. Введение

Распознавание эмоций человека является одной из задач машинного обучения. Интерес к этой теме каждый год увеличивается, т.к. её можно применить ко многим сферам деятельности: отслеживание состояние водителей во время управления транспортным средством, мониторинг обучающихся, проходящих обучение в онлайн форме, маркетинговые исследования и т. д.

Прежде чем решать задачу распознавания эмоций, необходимо определить, какую модель классификации эмоций необходимо использовать для распознавания. В настоящее время популярны два типа моделей эмоций: дискретные и многомерные.

Дискретные модели состоят из фиксированного набора базовых эмоций. В машинном обучении популярны модели, состоящие из 6 базовых эмоций человека (гнев, радость, удивления, отвращение, горе и страх) «нейтральной» категории и категории «другое», куда попадают все остальные эмоции, отличные от базовых. Считается, что все остальные эмоции можно выразить через комбинацию базовых эмоций. Разумеется такие модели очень легки для понимания и восприятия, но они не могут дать численную оценку силе эмоции.

Многомерные модели были созданы, чтобы решить данную проблему. Одной из популярных является модель Дж. Рассела, в ней вводится две оси: первая отвечает за знак эмоции (от негативной к положительной), а вторая ось отвечает за интенсивность эмоции (от низкой к высокой).

Следующим важным аспектом является то, на основе каких данных производится распознавание эмоций. Самым распространённым является распознавание по изображению, видео или аудио.

Построение классификатора эмоций по изображениям является не сложной задачей, однако фотографии не всегда могут точно отображать истинную эмоцию человека. Поэтому для получения более точных результатов всё чаще и чаще используют видео, где эмоцию можно распознать уже по последовательности кадров.

Но одного видео может быть тоже недостаточно для распознавания эмоции, поэтому часто вместе с видео информацией используют и аудио информацию.

Так как задача распознавания эмоций является достаточно популярной, то в свободном доступе можно найти данные, на основе которых можно обучать модели. Вот некоторые из них: OMG-Emotion challenge [1], AffectNet [2], AFEW-VA [3], Aff-Wild2 [4]. Преимущества этих моделей в том, что они содержат в себе видео и аудио составляющие, а также размечены для распознавания по многомерной модели Рассела.

2. Основные подходы построения систем для распознавания эмоций

Классическим подходом для решения задачи распознавания эмоций является классификация по ключевым точкам на лице человека. Соответственно для получения этих точек необходимо использовать один из алгоритмов: PDM [5], CML [6], AAM [7], CNN [8]. Необходимо разместить точки на лице таким образом, чтобы они захватывали мимику. Далее эти точки необходимо подать на классификатор для получения результатах [9]. В качестве классификатора могут выступать SVM или Random Forest.

Минусом данного подхода является то, что использование только ключевых точек без визуальной составляющей ведёт к потере информации. Чтобы решить данную проблему, в найденных точках вычисляют дескрипторы: LBP [10], HOG [11], SIFT [12], LATCH [13]. После ряда преобразований, полученные данные подаются для классификации.

Использование данного подхода для распознавания эмоций является устаревшим и в настоящее время многие современные модели используют глубокие свёрточные сети.

Преимущества использования глубоких свёрточных сетей заключается в том, что зачастую нет необходимости создавать и обучать модель с нуля. Для решения задачи можно взять уже готовую архитектура, изменить её под свою задачу и до обучить на своих данных. Поэтому для распознавания эмоций в качестве основы можно взять уже готовы сети для распознавания лиц [14].

Как мы ранее отметили, построить классификатор для изображений не является сложной задачей, однако лучше работать с видео информацией. Поэтому популярность получили следующие два подхода для классификации эмоций на основе глубоких свёрточных сетей.

В первом подходе на вход свёрточной сети подаются изображения (кадры видео), а на выходе сеть выдаёт высокоуровневые признаки изображения. Далее полученные высокоуровневые признаки подаются на вход рекуррентной сети (чаще всего используют LSTM сеть) для учёта временной составляющей. И уже рекуррентная сеть выдаёт информацию о том, какая эмоция была распознана [15].

Второй подход подразумевает использование 3D-CNN [16], на которую подают кадры из видео. Особенностью использования 3D-CNN является использование свёрток для преобразования четырёхмерных данных в трёхмерные карты признаков.

Также можно и объединить оба подхода для решения задачи [17].

Для распознавания эмоций по аудио чаще всего применяют два подхода. В первом случае анализу поддаётся не исходная звуковая волна, а разнообразные наборы признаков: F0, MFCC, LPC, i-вектора и др. После извлечения данных признаков их можно подать на вход SVM или Random Forest для классификации.

Второй подход подразумевает использование свёрточных сетей [18]. Для этого звуковой сигнал преобразовывают в изображение (чаще всего в спектрограмму). Дальше используется тот же принцип, что и для распознавания по видео.

Имея две модели для распознавания эмоций по видео и аудио информации теперь необходимо объединить их результаты для получения итогового ответа. Самым очевидным и простым вариантом является объединения результатов, например получить их среднее значение. На практике используют более сложные алгоритмы для объединения результатов, такие как Multiple Kernel Learning [19] и ModDrop [20].

3. Заключение

Так как пока задача распознавания эмоций не решена полностью, то исследования в этой направлении являются актуальными. Видно, что современные модели преимущественно используются комбинацию глубоких свёрточных сетей и рекуррентных сетей, а так же объединение результатов моделей, работающих с видео и аудио информацией. И что для более корректного определения эмоции человека необходимо использовать многомерную классификацию, например модель Рассела.

Литература

[1] One-Minute Gradual-Emotion Behavior Challenge [Electronoc resource]. – Access mode: https://www2.informatik.uni-hamburg.de/wtm/OMG-EmotionChallenge (04.09.2022).

[2] AffectNet [Electronic resource]. – Access mode: http://mohammadmahoor.com/affectnet (10.09. 2022).

[3] AFEW-VA database for valence and arousal estimation In-The-Wild [Electronic resource]. – Access mode: https://ibug.doc.ic.ac.uk /resources/afew-va-database (21.09. 2022).

[4] Aff-Wild2 database [Electronic resource]. – Access mode: https://ibug.doc.ic.ac.uk/resources/aff-wild2 (27.09. 2022).

[5] Point Distribution Model [Electronic resource]. – Access mode: https://www.menpo.org/menpofit/pdm.html (03.10. 2022).

[6] OpenFace 2.2.0: a facial behavior analysis toolkit [Electronic resource]. – Access mode: https://github.com/TadasBaltrusaitis /OpenFace (05.10. 2022).

[7] Active Appearance Model [Electronic resource]. – Access mode: https://www.menpo.org/menpofit/aam.html (05.10. 2022).

[8] How far we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks) [Electronic resource]. – Access mode: https://github.com/1adrianb/2D-and-3D-face-alignment (14.10. 2022).

[9] Ko, B.C. A Brief Review of Facial Emotion Recognition Based on Visual Information / B.C. Ko // Sensors. – 2018. – Vol. 2. – P. 401.

[10] Local Binary Pattern for texture classification [Electronic resource]. – Access mode: https://scikit-image.org/docs/dev/auto_examples/ features_detection/plot_local_binary_pattern.html (14.10. 2022).

[11] Histogram of Oriented Gradients [Electronic resource]. – Access mode: https://scikit-image.org/docs/dev/auto_examples/ features_detection/plot_hog.html (14.10. 2022).

[12] Image Classification in Python with Visual Bag of Words (VBoW) [Electronic resource]. – Access mode: https://ianlondon.github.io/ blog/how-to-sift-opencv/ (22.10. 2022).

[13] Performance Evaluation of Binary Descriptors – Introducing the LATCH descriptor [Electronic resource]. – Access mode: https://gilscvblog.com/2015/11/07/performance-evaluation-of-binary-descriptor-introducing-the-latch-descriptor/ (22.10. 2022).

[14] Knyazev, B. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video / B. Knyazev., R. Shvetsov, N. Efremova, A. Kuharenko // Computer Vision and Pattern Recognition. – 2017.

[15] Ouyang, X. Audio-visual emotion recognition using deep transfer learning and multiple temporal models / X. Ouyang, S. Kawaai, E. Goh, S. Shen, W. Ding, H. Ming, D.Y. Huang // In Proceedings of the 19th ACM International Conference on Multimodal Interaction. – 2017. – P. 577–582.

[16] Li. Y. Deep Learning of Human Emotion Recognition in Videos / Y. Li // Technical Report, Uppsala University. – 2018. – P. 42.

[17] Hasani, B. Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks / B. Hasani, M.H. Mahoor // IEEE Conference on Computer Vision and Pattern Recognition Workshops. – 2017. –P. 2278-2288

[18] Niu, Y. A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks // Y. Niu, D. Zou, Y. Niu, Z. He, H. Tan // Computer Vision and Pattern Recognition. – 2017.

[19] Poria, S. A. Convolutional MKL based multimodal emotion recognition and sentiment analysis / S. Poria, I. Chaturvedi, E. Cambria, A. Hussain // 2016 IEEE 16th International Conference on Data Mining (ICDM). – 2016. – P. 439-448

[20] Neverova, N., Wolf, C., Taylor, G., Nebout, F. ModDrop: adaptive multi-modal gesture recognition // Computer Vision and Pattern Recognition // URL: https://arxiv.org/pdf/1501.00102.pdf.