Neurocomputing 253 (2017) 104–114

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Emotion-modulated attention improves expression recognition: A deep learning model

Pablo Barros*, German I. Parisi, Cornelius Weber, Stefan Wermter

Department of Informatics, University of Hamburg, Knowledge Technology, Vogt-Koelln-Strasse 30, Hamburg 22527, Germany

ARTICLE INFO

Article history:
Received 13 May 2016
Revised 30 December 2016
Accepted 23 January 2017
Available online 11 March 2017

Keywords:
Convolutional neural networks
Deep learning
Multimodal processing
Emotional attention
Emotion recognition

ABSTRACT

Spatial attention in humans and animals involves the visual pathway and the superior colliculus, which integrate multimodal information. Recent research has shown that affective stimuli play an important role in attentional mechanisms, and behavioral studies show that the focus of attention in a given region of the visual field is increased when affective stimuli are present. This work proposes a neurocomputational model that learns to attend to emotional expressions and to modulate emotion recognition. Our model consists of a deep architecture which implements convolutional neural networks to learn the location of emotional expressions in a cluttered scene. We performed a number of experiments for detecting regions of interest, based on emotion stimuli, and show that the attention model improves emotion expression recognition when used as emotional attention modulator. Finally, we analyze the internal representations of the learned neural filters and discuss their role in the performance of our model.

1. Introduction

Visual spatial attention allows animals and humans to process relevant environmental stimuli while suppressing irrelevant information. Several brain areas and neural mechanisms have been identified to be involved in the processing of spatial attention during perception [9]. For instance, it has been found that the superior colliculus (SC) – a midbrain structure responsible for the integration of audiovisual stimuli – plays a crucial role in spatial attention, more specifically in the process of target selection and estimating motor consequences such as saccades, i.e. quick eye movements to control the direction of fixation [24]. The integration of audiovisual stimuli in the SC has been extensively investigated from a neurophysiological perspective [30], with different computational approaches modeling the integration of multiple perceptual cues for triggering spatial attention in line with neurobehavioral evidence [4].

Converging findings suggest that selective attention is modulated by the affective significance of sensory inputs [32]. More specifically, it has been argued that emotional salience has a direct influence on attention and that neural processes responsible for emotional attention may supplement and even compete with other top-down mechanisms of perception. Behavioral studies have shown that people pay more attention to emotional rather than neutral stimuli and that these effects often are reflexive and involuntary, e.g. visual targets expressing an emotion such as happy or angry are found faster among distractors than targets without such emotional values [10,35]. Phelps et al. [27] showed that the visual detection threshold for low-contrast stimuli is improved if emotional cues are present, thus suggesting the existence of an emotion-driven mechanism to capture spatial attention. Additional studies suggest that in the case of limited attentional resources, emotional information is prioritized over non-affective cues [13,33]. These findings together indicate that emotional salience has a strong role in capturing attention, and emotional bias is also subject to a set of different non-affective regulatory effects.

Converging findings suggest that selective attention, processed in the SC, is modulated by the affective significance of sensory inputs [32]. In particular, it has been argued that emotional salience has a direct influence on attention and that neural processes responsible for emotional attention may supplement and even compete with other top-down mechanisms of recognition. Behavioral studies have shown that people pay more attention to emotional rather than neutral stimuli and that these effects often are reflexive and involuntary, e.g. visual targets expressing an emotion such as happy or angry are found faster among distractors than targets without such emotional values [33,35].

* Corresponding author.
E-mail addresses: barros@informatik.uni-hamburg.de (P. Barros), parisi@informatik.uni-hamburg.de (G.I. Parisi), weber@informatik.uni-hamburg.de (C. Weber), wermter@informatik.uni-hamburg.de (S. Wermter).

http://dx.doi.org/10.1016/j.neucom.2017.01.096

From a human–robot interaction (HRI) perspective, different computational models have been proposed for the detection and recognition of emotional expressions [1]. Different cues may carry emotional information such as face expressions, sound (voice pitch and intensity), and body movements [18]. It has been shown that the combination of these cues increases recognition accuracy [6], suggesting that models for the robust processing of emotional states should feature multimodal properties for the meaningful integration of a set of available perceptual cues. In this context, [12] established six universal emotions that exhibit invariance to cultural and racial factors: “Anger”, “Disgust”, “Fear”, “Happiness”, “Surprise”, and “Sadness”. However, for some HRI applications, emotional states were classified in terms of *positive* or *negative* emotions for triggering pro-active robot behaviors [3].

With the advance of deep learning networks, most of the recent work involves the use of neural architectures. The ones with the best performance and generalization apply different classifiers to different descriptors [7,22,26], and there is no consensus on a universal emotion recognition system. Most of these systems are applied to one modality only and reach a good performance for specific tasks. However, all these works rely on face detection models, which are not related to emotion recognition. Although such face detection models work well in controlled scenarios, they show poor performance when applied in complex scenarios [25,31].

As an extension of previous work aiming at emotion recognition [2], in this paper we investigated the modulation mechanisms of emotion-driven attention and implemented a deep neural architecture for the detection of emotional stimuli in a natural scene. We trained our model to distinguish between neutral and happy expressions conveyed by facial features and body movement and used this information as a modulator to our perception model, improving its recognition capabilities.

We show that although the input is composed of a single image sequence containing both, face and body movement cues, the model will autonomously learn separate cue-specific filters. In contrast to traditional deep learning models using discrete target labels for modulating the learning process, we use probability distributions that allow the model to estimate the location of interest, i.e. the region in the image that triggers selective attention. Interestingly, after using teaching signals with only one emotional expression in the image, experiments have shown that the model is able to produce congruent probability distributions for more than one expression present in the scene. We evaluated our system with a bi-modal face and body benchmark dataset, showing that the combination of facial properties and body movements significantly improves the detection of emotion-relevant areas in the image.

2. Deep emotional attention model

Our model combines the idea of hierarchical learning and selective emotional attention using convolutional neural networks (CNN). Our approach differs from traditional CNN-based approaches by two factors: first, the input stimuli are composed of the whole scene, which may or may not contain people expressing emotions. Second, the network is trained to (a) localize where the emotion expression is and (b) identify if the detected emotion expression is interesting enough to attract the attention of the model.

CNNs were used for several visual recognition tasks, starting with the Neocognitron proposed by Fukushima [15]. In most of the cases the CNNs were used to learn hierarchical descriptors from the input stimuli and describe the input data in a smaller, but highly abstract representation. In our work, we do not use the CNN as a classification technique. We employ the convolutional units as feature descriptors, but instead of learning hierarchical contours and shapes, which is commonly used in general image recognition tasks, they learn spatial information as discussed by Speck et al. [29].

We use our model to detect emotional events conveyed by face expressions and body movements. In this scenario, each convolutional unit learns how to process facial and movement features from the whole image. We differed our work from simple classification tasks by not tuning the convolutional units to describe forms, but rather to identify where an expression is located in the image. Therefore, we implement a hierarchical localization representation where each layer deals with a sub-region of the image. The first layers will learn how to detect Regions of Interest (ROI) which will then be fine-tuned in the deeper layers. Because the pooling units increase the spatial invariance, we only apply them in our last layers, which means that our first layers are only composed of convolutional units stacked together. In this Section 2 we describe our model, starting with common CNNs and our emotional attention model. To train our network as a localization model, we use a different learning strategy based on probability density functions, which will also be explained in this section.

2.1. Convolutional neural networks

Each layer of the CNN has a set of different convolutional units that increase the capability to learn different features from the same region in the image. This operation generates different filtered outputs, or filter maps, one for each unit. The pooling units in each of these filter maps are generating spatial invariance. Each set of filters acts in a receptive field in the input stimuli. The activation of each unit v_{nc}^{xy} at (x, y) of the nth filter in the cth layer is given by

$$v_{nc}^{xy} = max\left(b_{nc} + \sum_{m}\sum_{h=1}^{H}\sum_{w=1}^{W} w_{(c-1)m}^{hw} v_{(c-1)m}^{(x+h)(y+w)}, 0\right), \quad (1)$$

where max $(\cdot, 0)$ represents the rectified linear function, shown to be effective for training deep neural architectures [16], b_{nc} is the bias for the nth feature map of the cth layer, m indexes over the set of feature maps in the $(c-1)$ layer connected to the current layer c, $w_{(c-1)m}^{hw}$ is the weight of the connection between the unit (h, w) within a receptive field, connected to the previous layer $c-1$, and to the filter map m. H and W are the height and width of the receptive field.

In the pooling layers, a receptive field of the previous filter map is connected to a pooling unit in the current layer, reducing the dimensionality of the feature maps. The pooling units generate the maximum activation of the receptive field $u(x, y)$ which is defined as:

$$a_j = max_{n\times n}(v_{nc}u(x, y)), \quad (2)$$

where v_{nc} is the output of the convolutional unit. In this function, the pooling unit computes the maximum activation among the receptive field $u(x, y)$. The maximum operation down-samples the feature map maintaining the input structure.

2.2. Sequence processing

In a traditional CNN, convolutional units are connected to a single input stimulus generating a representation of that single input. As we are dealing with sequential data, our representation is a series of stimuli which may affect the representation of each other. Therefore, we use the cubic receptive fields, implemented in the convolutional units by stacking filters together to be applied in the same region of different input stimuli [21]. Then, it is possible to create a representation of the whole sequence, which will be tuned to adapt to the presence of patterns in the sequence as a whole