In: *Proceedings of the ICLR2021 Workshop on How Can Findings About The Brain Improve AI Systems? (Brain2AI@ICLR2021),* Number 18, virtual event, May 2021.

Cross-modal emotion recognition: How similar are patterns between DNNs and human fMRI data?

Christoph Korn, Sasa Redzepovic, Jan Gläscher
Institute for Systems Neuroscience
University Medical Center Hamburg-Eppendorf
Martinistr. 52, 20246 Hamburg, Germany
{c.korn,glaescher}@uke.de

Matthias Kerzel, Pablos Barros, Stefan Heinrich, Stefan Wermte
Department of Informatics
University of Hamburg
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
{kerzel,wermter}@informatik.uni-hamburg.de

Abstract

Deep neural networks (DNNs) have reached human-like performance in many perceptual classification tasks including emotion recognition from human faces and voices. Can the patterns in layers of DNNs inform us about the processes in the human brain and vice versa? Here, we set out to address this question for cross-modal emotion recognition. We obtained functional magnetic resonance imaging (fMRI) data from 43 human participants presented with 72 audio-visual stimuli of actors/actresses depicting six different emotions. The same stimuli were classified with high accuracy in our pre-trained DNNs built according to a cross-channel convolutional architecture. We used supervised learning to classify four properties of the audio-visual stimuli: The depicted emotion, the identity of the actors/actresses, their gender, and the spoken sentence. Inspired by recent studies using representational similarity analyses (RSA) for uni-modal stimuli, we assessed the similarities between the layers of the DNN and the fMRI data. As hypothesized, we identified gradients in pattern similarities along the different layers of the auditory, visual, and cross-modal channels of the DNNs. These gradients in similarities varied in expected ways between the four classification regimes: Overall, the DNNs relied more on the visual arm. For classifying spoken sentences, the DNN relied more on the auditory arm. Crucially, we found similarities between the different layers of the DNNs and the fMRI data in searchlight analyses. These pattern similarities varied along the brain regions involved in processing auditory, visual, and cross-modal stimuli. In sum, our findings highlight that emotion recognition from cross-modal stimuli elicits similar patterns in DNNs and neural signals. In a next step, we aim to assess how these patterns differ, which may open avenues for improving DNNs by incorporating patterns derived from the processing of cross-modal stimuli in the human brain.

1 Introduction

DNNs have reached (or even surpassed) human performance in many perceptual classification tasks. Similar to humans, cross-modal DNNs can combine information from multiple modalities such as visual and auditory features of the to-be-classified stimuli. Such DNNs excel at cross-modal emotion recognition from videos of humans who dynamically transmit their emotions via varying their facial expressions and via uttering sentences with affective prosody. (Barros & Wermter, 2016).

Recent studies have compared how DNNs and human brains represent uni-modal features during classification tasks for visual stimuli (Cichy et al., 2016) and for auditory stimuli (Kell et al., 2018). Methodologically, such comparisons are made possible by representational similarity analyses (RSA), which transform information representations from various sources of data into a common format, i.e., into stimulus-by-stimulus similarity matrices (Kriegeskorte et al., 2008).

Here, we harnessed this approach (Cichy & Teng, 2017) to test whether patterns in fMRI signals (obtained from participants viewing cross-modal videos of actors and actresses enacting basic emotions) correspond to patterns in DNNs trained for emotion classification on the same stimulus corpus. We expected gradients along the audio-visual processing streams for the different layers and channels of the DNNs.

2 Methods

2.1 fMRI

Forty-three participants were presented with 72 different audio-visual videos of three actors and three actresses enacting five basic emotions (happy, angry, fearful, disgusted, and surprised) plus neutral.

These audio-visual stimuli were taken from the highly controlled RAVDESS dataset (Livingstone & Russo, 2018). The RAVDESS dataset contains stimuli in which actors/actresses express emotions with high or low intensity. We chose 72 videos in which emotions were depicted with high intensity. Recognition accuracy was thus high in human participants, i.e., in our participants and in the participants tested by the team who developed the RAVDESS dataset (Livingstone & Russo, 2018). Specifically, for our experiment, we selected a subset of videos from the three ”best” actors and actresses with the highest emotion recognition by participants in the ”RAVDESS development study” (Livingstone & Russo, 2018). Each video was 3 s long. The videos showed dynamic facial expressions of the six emotions. Actors and actresses spoke one of two neutral sentences with the emotional prosody that corresponded to the facial expression.

The 72 stimuli were repeated in six scanning sessions (in counterbalanced order). Each scanning session included null events and took about 8 min. fMRI data was collected on a Siemens Prisma 3T scanner using a multiband EPI sequence.

fMRI data was analyzed using SPM and custom code based on a toolbox for RSA (Nili et al., 2014). Roughly speaking, RSA is based on correlations between the similarity matrices for fMRI data and the similarity matrices for layers of the DNNs. The magnitudes of these correlations are calculated for each searchlight position. Averages of these magnitudes across participants are and then tested against zero and depicted on brain maps.

2.2 DNNs

The DNNs were built based on a cross-channel convolutional architecture adapted from earlier work (Barros & Wermter, 2016). The DNNs process short video sequences and audio recordings in two separate channels of the network (the visual and the auditory arms). Audio recordings are transformed into spectrograms.

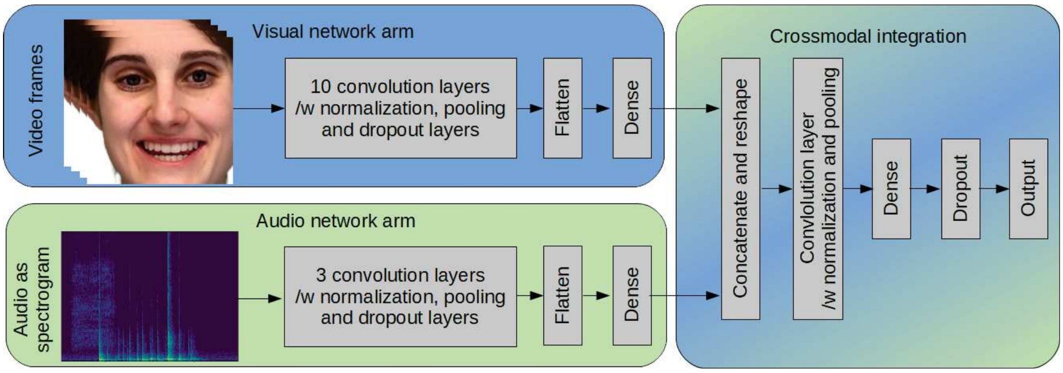
Video frames
Visual network arm
10 convolution layers /w normalization, pooling and dropout layers
Flatten
Dense
Audio as spectrogram
Audio network arm
3 convolution layers /w normalization, pooling and dropout layers
Flatten
Dense
Crossmodal integration
Concatenate and reshape
Convlolution layer /w normalization and pooling
Dense
Dropout
Output

Figure 1. Schematic outline of the DNN architecture that takes cross-modal information from video snippets (lengths: 3 s) with actors and actresses performing one of six emotional facial expressions along with voicing a sentence in the corresponding emotional prosody. The DNNs consist of layers in ”visual arm” and layers in an ”audito arm,” which are integrated in ”cross-modal” layers. Overall, the DNNs contain 29 layers.

The ”visual and the audio” arms of the network were pre-trained on a larger part of the RAVDESS corpus (excluding the 72 stimuli used in the fMRI experiment). All three parts of the DNNs (visual and the auditory arms plus the cross-modal part) were then trained and tuned on the 72 stimuli of the fMRI experiment (i.e., the ”cross-modal part” was not pre-trained). Four different training setups were realized during which supervised learning is used to classify the depicted emotion, the identity of the actors/actresses, their gender, and the spoken sentence. Training was repeated ten times and the best performing models were chosen for activation extraction. The DNNs showed a classification accuracy of avg. 0.82 (best model: 0.87) on emotion classification. ”Activations” from all layers of the DNNs were extracted for all 72 stimuli. These ”activations” for the 72 stimuli were correlated with each other, resulting in 72 x 72 square matrices. These ”similarity matrices” were used for RSA of the fMRI data. The lower triangular entries of the correlation matrices for all 29 layers were also correlated with each other, resulting in 29 x 29 square matrices.

3 Results

3.1 Pattern similarities within the DNNs

In a first step, we assessed the similarities (i.e., the correlation matrices) for the different layers from the DNNs for the four different classification regimes. Thereby, we characterized the unique contributions of the different layers for each classification regime, which was crucial for obtaining a benchmark of how similar patterns in DNNs and patterns in fMRI data were by design (of the DNNs).

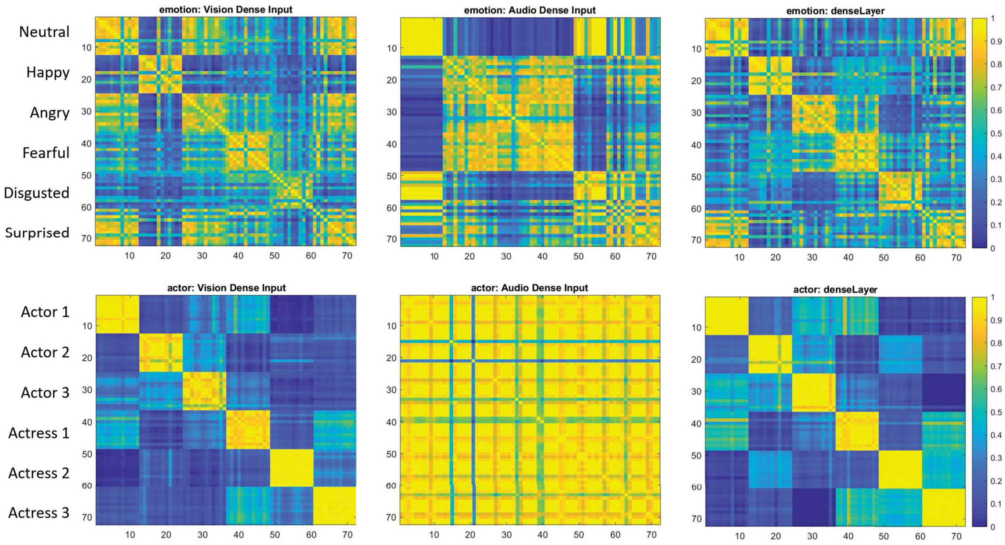

Figure 2. Similarity matrices of the “activations” in the units of three “upper” layers for the 72 audio-visual stimuli used in this study. The existence of the “yellow blocks” with high similarities (i.e., high correlations) along the diagonal indicates that “activations” were similar within—and rather distinct between—the six emotion categories. The upper row shows matrices for emotion classification. The lower row shows matrices for identity classification. That is, the correlation coefficients of the 72 stimuli are sorted according to emotion in the upper row and according to identity in the lower row. The first column shows an “upper” visual layer, the second column an “upper” auditory layer, and the third column an “upper” cross-modal integration layer.

The similarity patterns of the “upper” layers of the visual and the auditory arms showed that the classification distinguished between visual versus auditory features. For emotion classification, visual and auditory features by themselves provided rather decent separations of the six emotions (although the audito arm conflated happy, angry, and fearful). For identity classification, the auditory arm of our DNN failed and the similarity between stimuli in the cross-modal layers simply reflected the similarity in the visual arm. Nevertheless, identity classification resulted in more clear-cut similarity matrices than emotion classification, which might mirror the overall higher difficulty of recognizing emotions for DNNs (and for humans).

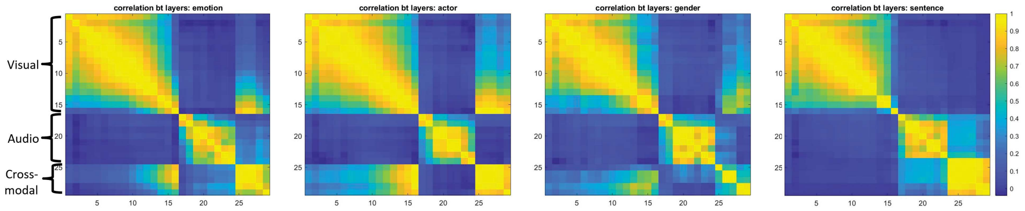

Figure 3. The similarity matrices between the 29 layers of the DNNs for the four classification regimes: emotion, identity, gender, and sentences. Neighboring layers within each arm are highly correlated by design. These matrices show how the layers in the three parts of the DNNs are related to each other. For the first three classification regimes, the visual arm seems more relevant than the auditory arm. As expected, for classifying sentences, the auditory arm is more relevant. Crucially, for emotion classification, layers from both visual and auditory arms correlate with the layers in the cross-modal part of the DNN.

3.2 Pattern similarities between the DNNs and fMRI signals

We used searchlight RSA to relate the patterns in various layers of the DNNs to fMRI signals. As a proof of concept, we depict one layer for each arm of the DNNs trained for emotion and identity classification.

Similarity patterns in the visual arm corresponded to brain regions in the visual processing stream, notably the fusiform face area. Similarity patterns in the auditory arm corresponded to regions in the auditory processing stream, notably parts of the superior temporal gyrus.

Crucially, cross-modal regions such as the temporo-parietal junction and the superior temporal sulcus were related to similarity patterns of the cross-modal integration layers in the DNN trained for emotion classification.

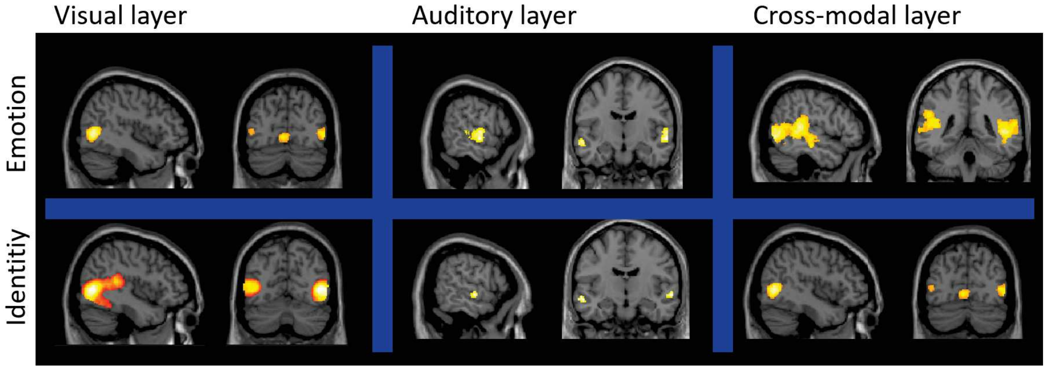

Figure 4. Searchlight RSA testing for the correspondence between different layers in the three arms of the DNNs trained for classification of the six emotions and for the identities of the six actors/actresses. Interestingly, the cross-modal layers were related to the temporo-parietal junction and the superior temporal sulcus. These findings corroborate notions that these regions are involved in combining visual and auditory processing.

4 Summary and outlook

Our current findings provide a proof of concept that “artificial and biological networks” can show similar patterns when presented with cross-modal emotional stimuli. Patterns of visual, auditory, and cross-modal layers of DNNs were related to fMRI signals in the fusiform face area, the superior temporal gyrus, and the temporo-parietal junction.

We are currently extracting similarity matrices for specific brain regions. Due to the variability and noise in fMRI signals, it is rather difficult to numerically compare the similarity matrices from fMRI data and DNNs in a straightforward way. Nevertheless, we can assess the relative contribution of the visual and auditory arms for regions in the fusiform face area, the superior temporal gyrus, the temporo-parietal junction, etc.

We will follow up on these promising results by refining (and hopefully improving) the employed DNNs. In particular, univariate fMRI analyses showed the well-described involvement of the amygdala and insula in processing emotional versus neutral stimuli. We therefore plan to use a "reverse translational approach" (form fMRI signals to DNNs) to test whether and how patterns derived from these brain regions could be captured by DNNs.

Acknowledgments

This work is funded by the German Research Foundation (DFG) within the collaborative research centre on cross-modal learning (TRR169).

References

Pablo Barros and Stefan Wermter. Developing crossmodal expression recognition based on a deep neural model. *Adaptive behavior*, 24(5):373–396, 2016.

Radoslaw Martin Cichy and Santani Teng. Resolving the neural dynamics of visual and auditory scene processing in the human brain: a methodological approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160108, 2017.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13, 2016.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.