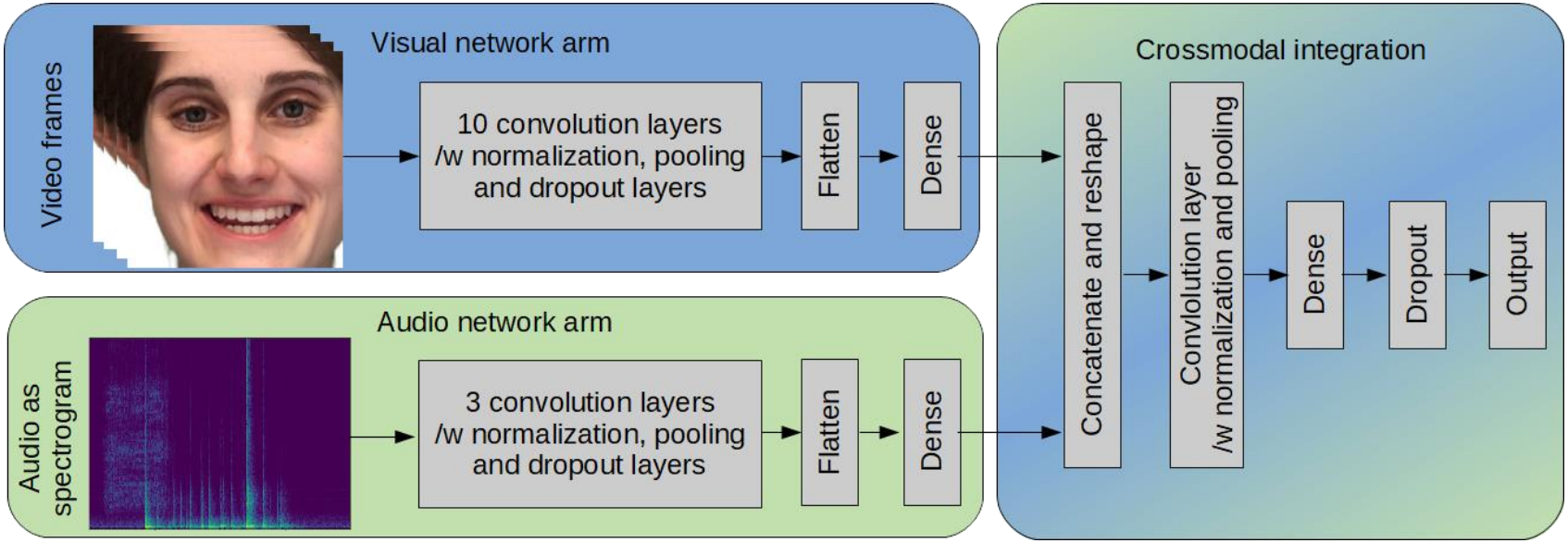
# Cross-modal emotion recognition: How similar are patterns between DNNs and human fMRI data?

Christoph W. Korn, Saša Redžepović, Jan Gläscher
Institute for Systems Neuroscience
University Medical Center Hamburg-Eppendorf
Martinistr. 52, 20246 Hamburg, Germany

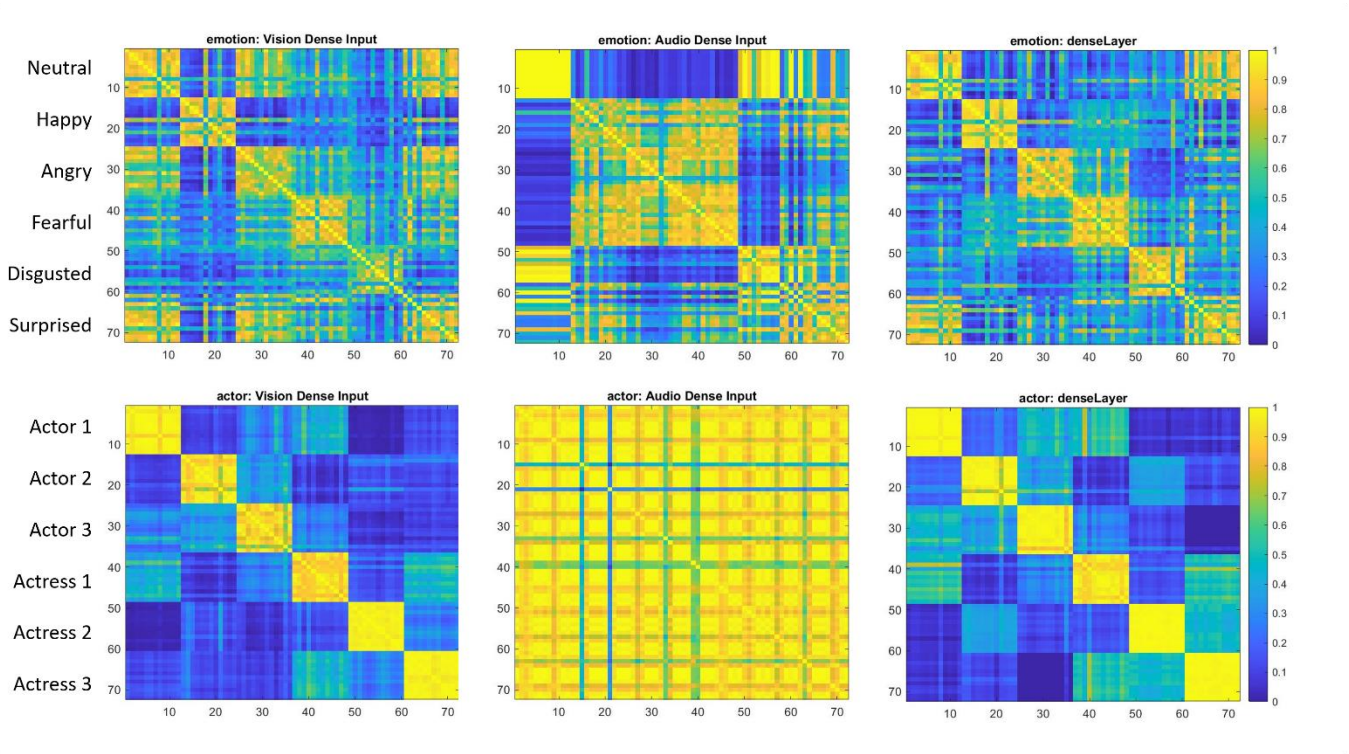
Matthias Kerzel, Pablo Barros, Stefan Heinrich, Stefan Wermter
Department of Informatics
University of Hamburg
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany

# Outline of the 2-armed DNN architecture



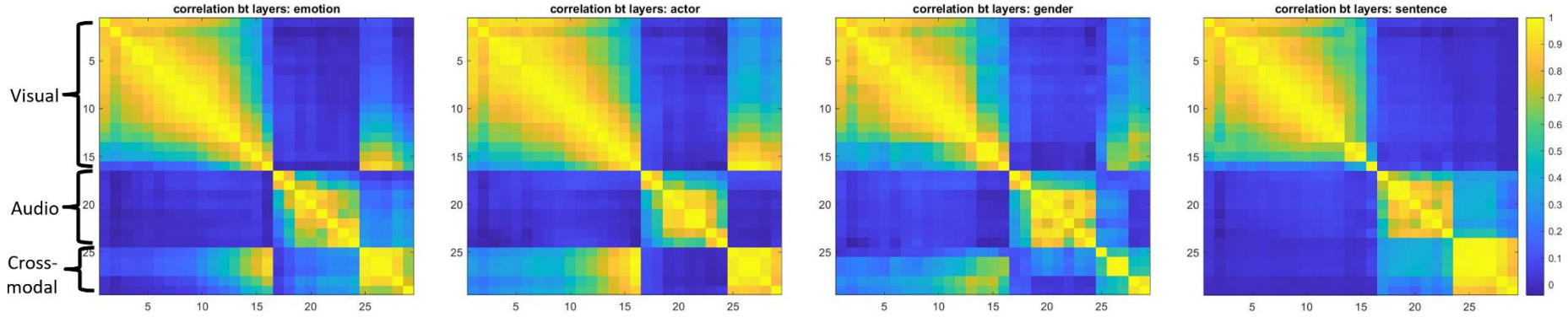

- DNN takes cross-modal information from video snippets (lengths: 3 s)
- Actors and actresses perform 1 of 6 emotional facial expressions along with voicing a sentence in the corresponding emotional prosody
- Overall, the DNNs contain 29 layers

# Similarity of the “activations” in 3 “upper” DNN layers



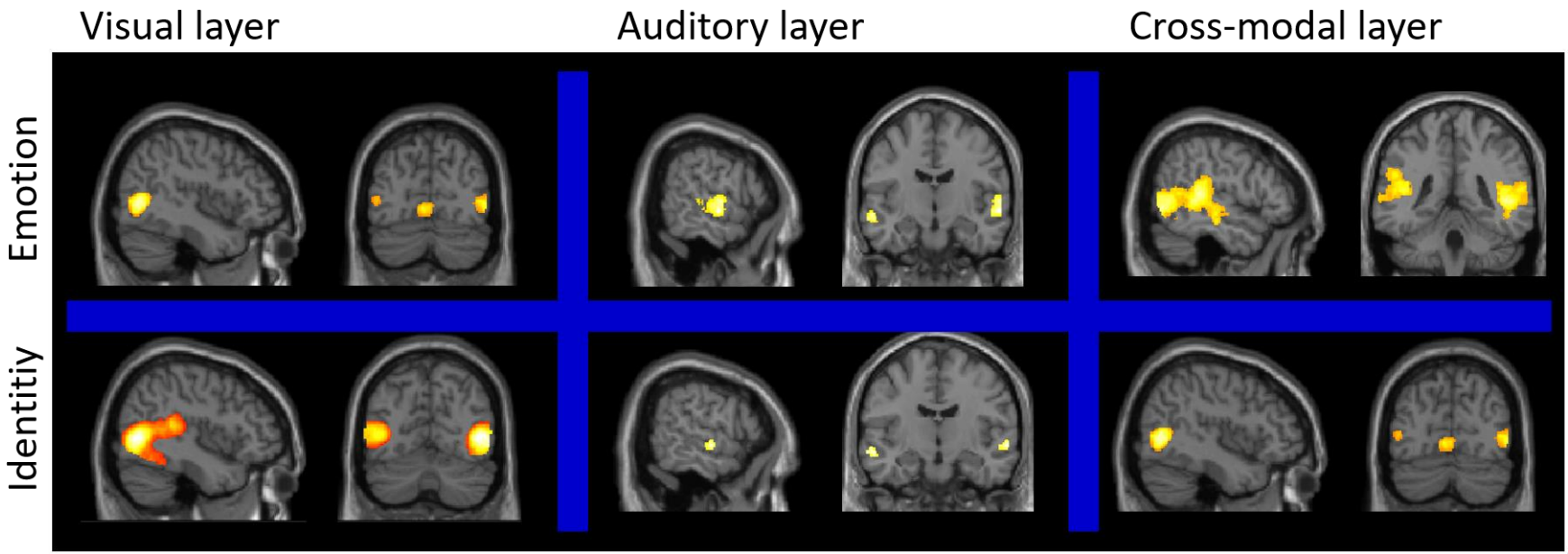

- 72 audio-visual stimuli
- 2 classification regimes: emotion & identity
- “Activations” were similar within (and rather distinct between) the 6 emotions
- Overall, visual arm was more decisive than auditory arm

# Similarities between 29 layers of the DNNs




- 4 classification regimes: emotion, identity, gender, and sentences
- For classifying emotion, identity, and gender, the visual arm seemed more relevant than the auditory arm
- For classifying sentences, the auditory arm was more relevant
- Crucially, for emotion classification, layers from both visual and auditory arms correlated with the layers in the cross-modal part of the DNN

# Searchlight RSA: Correspondence between DNNs & fMRI data




- Different layers in the 3 parts of the DNNs trained for classification of 6 emotions and for the identities of 6 six actors/actresses
- Cross-modal layers were related to the temporo-parietal junction (TPJ) and the superior temporal sulcus (STS)
- Proof-of-principle findings corroborating the notion that these regions are involved in combining visual and auditory processing