Immersions

Visualizing and sonifying how an artificial ear hears music

University of Music Hochschule für Musik Karlsruhe

IMWI Institute for Music Informatics and Musicology

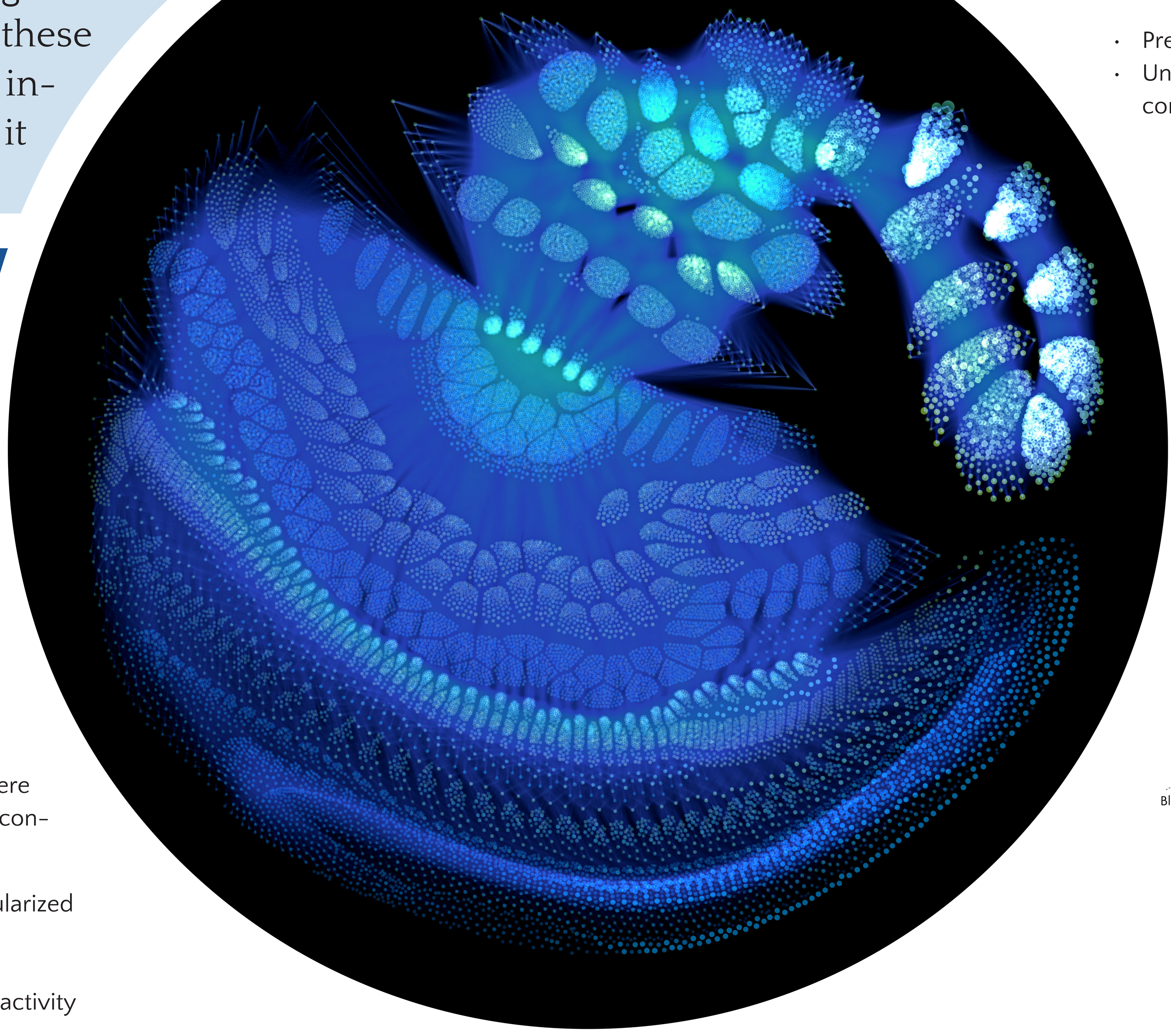
Vincent Herrmann

Insitute for Music Informatics and Musicology, University of Music Karlsruhe

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2019, Vancouver

The "artificial ear" is an audio processing neural network that was trained in an unsuperwised way using contrastive predictive coding. Its inner workings are revealed in two ways: An optimization procedure generates sounds that activate certain regions in the network - and simultaneously these activations are visualized. You can interact with the model and explore it in realtime.



Input Optimization

Features the model relies upon can be sonified by optimizing the input such that it activates certain neurons.

- Compute the gradient of some neuron or a group of neurons w.r.t. the input
- Perform gradient ascent / descent to maximize / minimize the activation of the selected neurons

- All preprocessing used on the raw sensory data has to be differentiable
- The constant-Q transform of the audio signal used here (see Model Architecture) can be implemented as a convolutional layer with fixed weights

For better results the optimization procedure can be regularized

- Temporal shifting of the input
- Randomly masking time or pitch regions
- Additional loss terms punishing noise or total neural activity

Contrastive Predictive Coding

An encoder and a predictive model for the encodings are jointly trained in an unsupervised way using contrastive loss.

- Predictive coding might be how perception works in the brain
- Unsupervised representation learning can be done using contrastive predictive coding

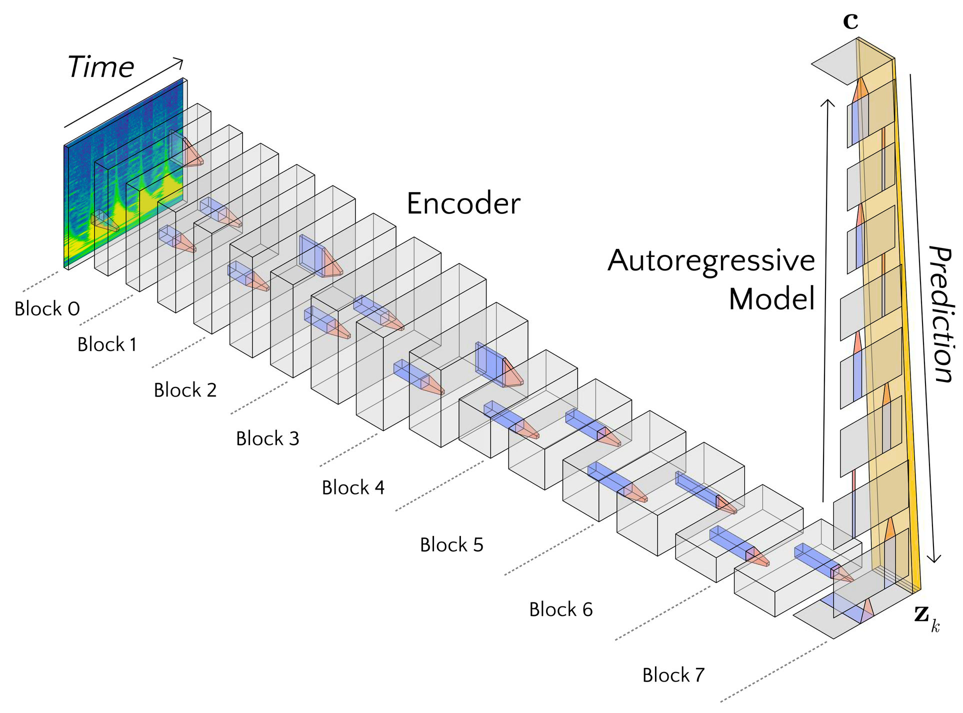
Two networks:

- The encoder network creates time step encodings $\mathbf{z}_t$
- The autoregressive model summarizes multiple $\mathbf{z}$ encodings into $\mathbf{c}$
- Future time steps $\mathbf{z}_k$ are predicted by linear transformations $\mathbf{M}_k$ of $\mathbf{c}$

Encoder and autoregressive model are jointly trained using a contrastive loss function:

$$\mathcal{L}_{CPC} = -\sum_{n,k} \log \frac{\exp(\mathbf{z}_{n,k}^T \mathbf{M}_k \mathbf{c}_n)}{\sum_m \exp(\mathbf{z}_{m,k}^T \mathbf{M}_k \mathbf{c}_n)}$$

Both m and n are indices of the minibatch dimension



Neural Layout

The artificial neurons are laid out in 2D using a multi-level force graph layout method.

Neurons and their connections of the network → vertices and edges of a graph

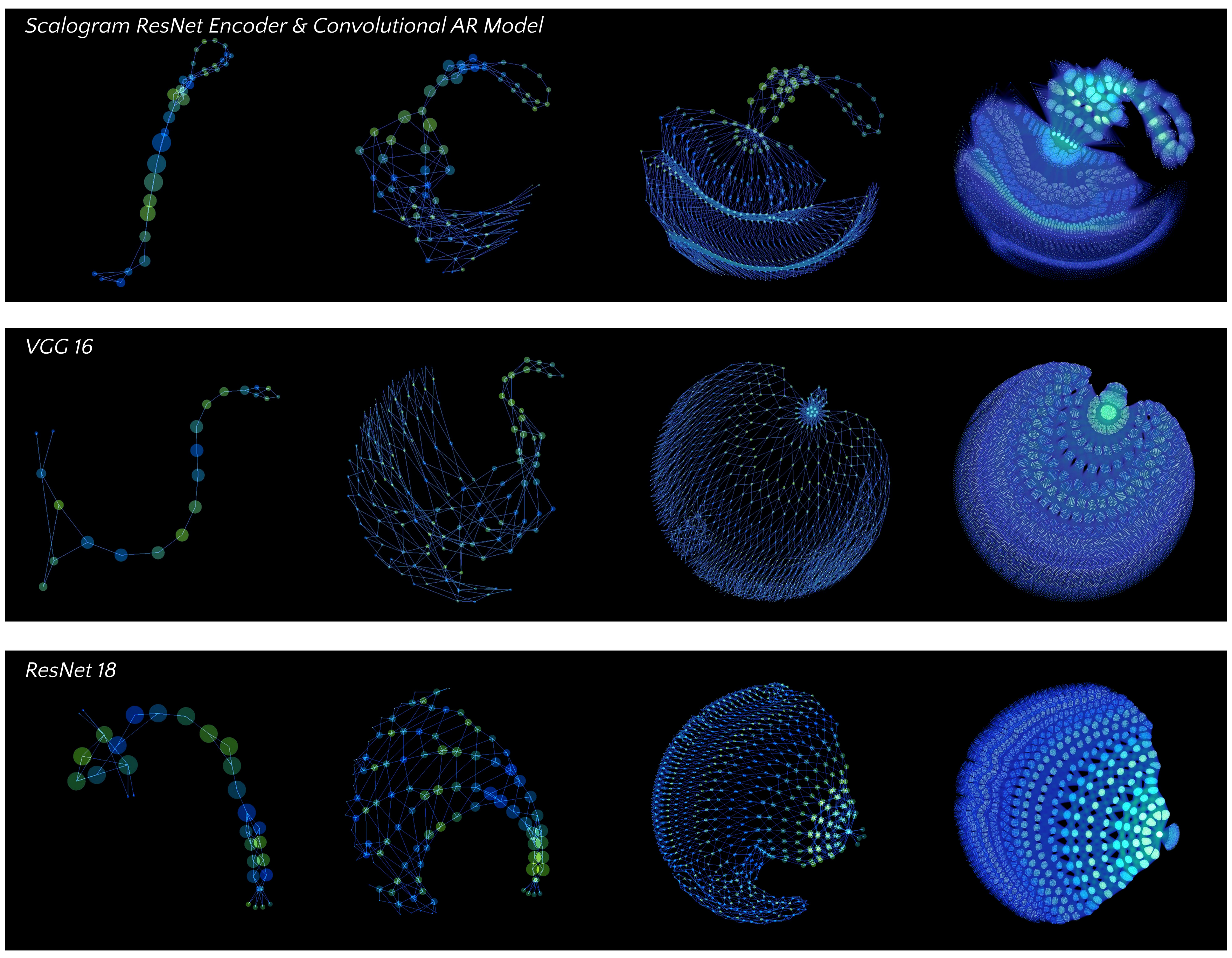
- The graph is simulated as a physical system
- Vertices repel each other, connections pull them together

The layout is calculated for progressive graph resolution levels to avoid local energy minima

- Low level graph approximations are created by iteratively merging vertices that are neighbouring in one of the dimensions (in our case channel and pitch)
- Higher resolution levels are reconstructed during the simulation in the reverse order
- Here the time dimension of the network is excluded - a new layout is calculated for each time step, leading to moving images

- A Barnes-Hut tree algorithm is used to reduce the computational complexity
- Implemented in PyTorch, can be accelerated using GPUs (code available!)
- Same method can be used for various kinds of neural network architectures

Four levels of the layout calculation for different network architectures



https://vincentherrmann.github.io/demos/immersions

Model Architecture

A modified ResNet with scalogram input serves as encoder. The autoregressive model is a 1D fully convolutional network.

Visualization with each layer in a different colour

Input:

- Scalogram (constant-Q or wavelet transform) of a four second long audio clip
- Close to the representation that the cochlea passes on to the brain

Encoder:

- ResNet-based architecture with 8 residual blocks
- Using 3x3 kernels (common) as well as 1x25 kernels supposed to detect overtone and harmonic structures

Autoregressive Model:

- Residual network with 1D (time) causal convolutions

Visualization & Interaction

The activations of the neural network are visualized in realtime. All aspects of the input optimization procedure can be interactively controlled.

- Neurons with high activation light up and are enlarged
- The visualization of the connections has to be precomputed, due to their high quantity (many millions per time frame)
- The selected target neurons for the input optimization are additionally highlighted

For the interactive mode, the input to the model is an audio clip that is continuously played in a loop

- The optimization procedure constantly generates new audio clips
- As soon as one clip has finished playing, the latest newly calculated clip is started
- Result is a slow acoustic morphing

Among the parameters that can be controlled via the GUI and a MIDI-controller are:

- The currently focussed neurons (in the dimensions pitch, time and channel)
- Audio clips serving as origin for the optimization
- Various regularization factors

Acknowledgements

I would like to thank Christoph Seibert and Leonard Herrmann for their important suggestions and help along the way, as well as the Cultural Office Karlsruhe and the state of Baden-Württemberg for sponsoring the project.