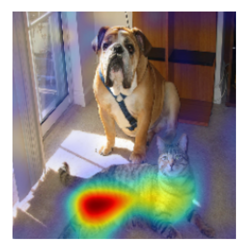
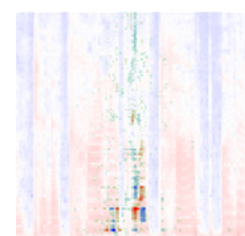
Universität Potsdam

Introduction to Deep Learning

Introspection


Andreas Krug, M.Sc.

January 21, 2019

Orga

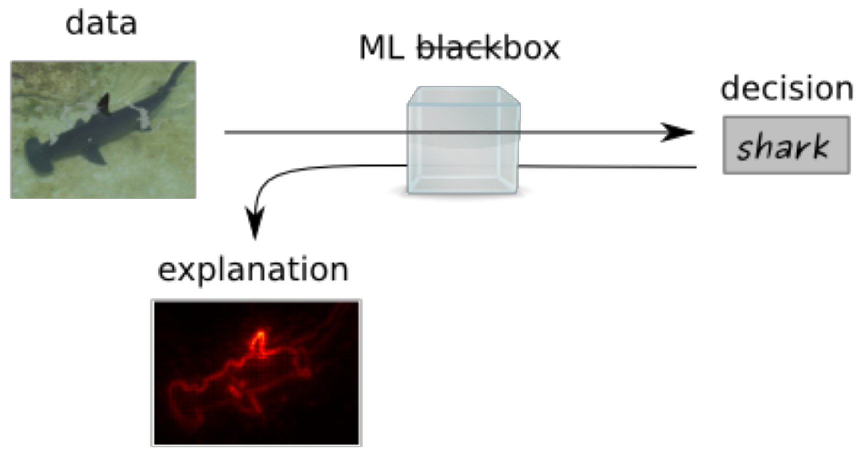
- Exam date based on the **8 votes** on Doodle
  - Mon, February 25
  - Mon, March 18
  - Wed, March 27
- https://moodle2.uni-potsdam.de/
  “Introduction to Deep Learning“ in WS18/19
  password: see Mattermost
- Scheduler will be available soon

Introspection

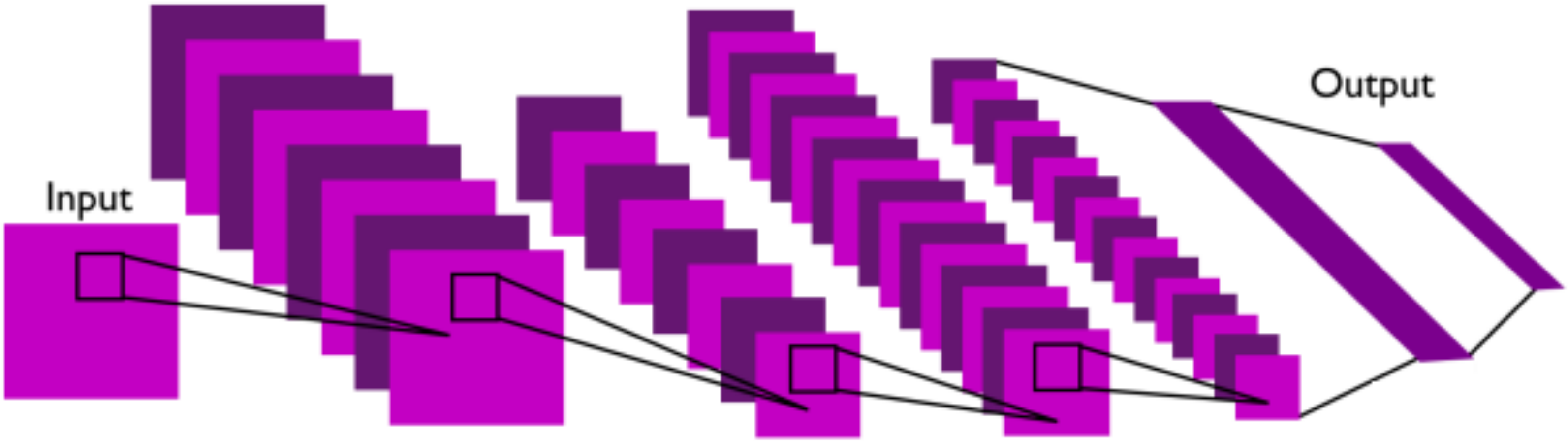
Types of introspection

feature visualization

Saliency maps
e.g. layerwise relevance propagation (LRP)

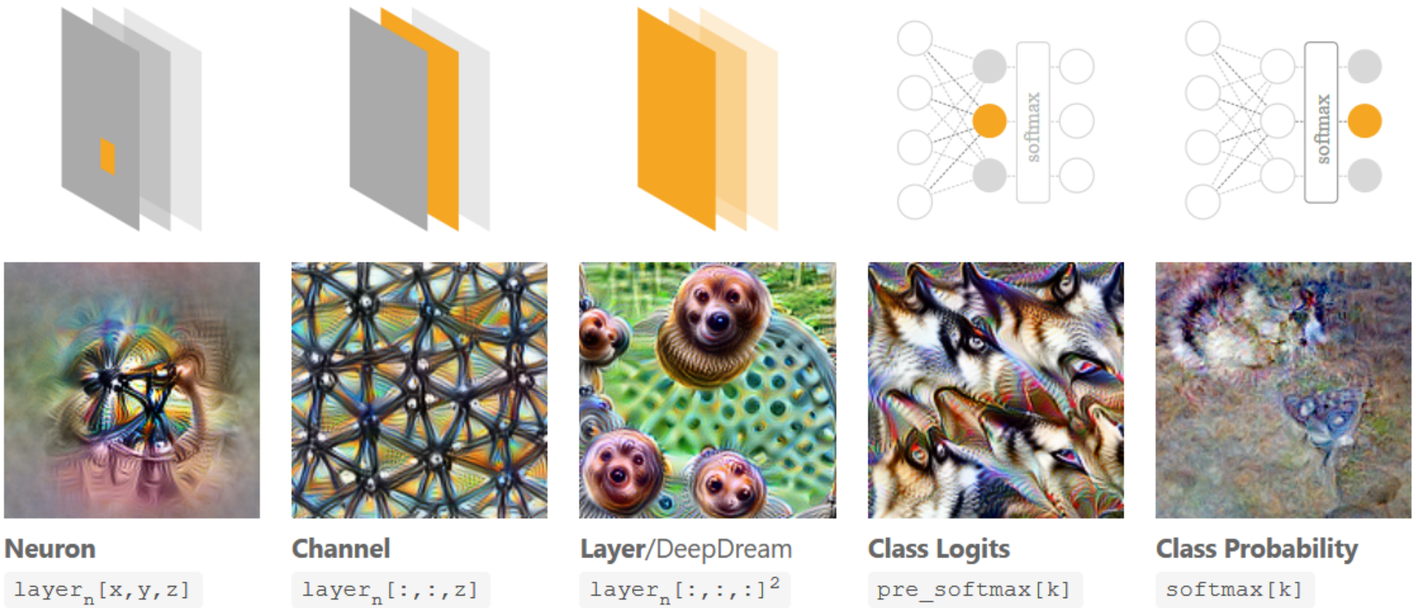


Feature visualization



feature visualization by optimization
(find the input that optimizes a particular part of the network)

Feature visualization



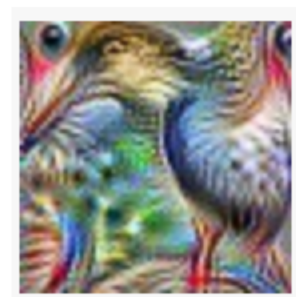
[https://distill.pub/2017/feature-visualization/]

Feature visualization

What's the main problem with the (vanilla) optimization approach?
How do we solve this?

unregularized optimization is unnatural

vs



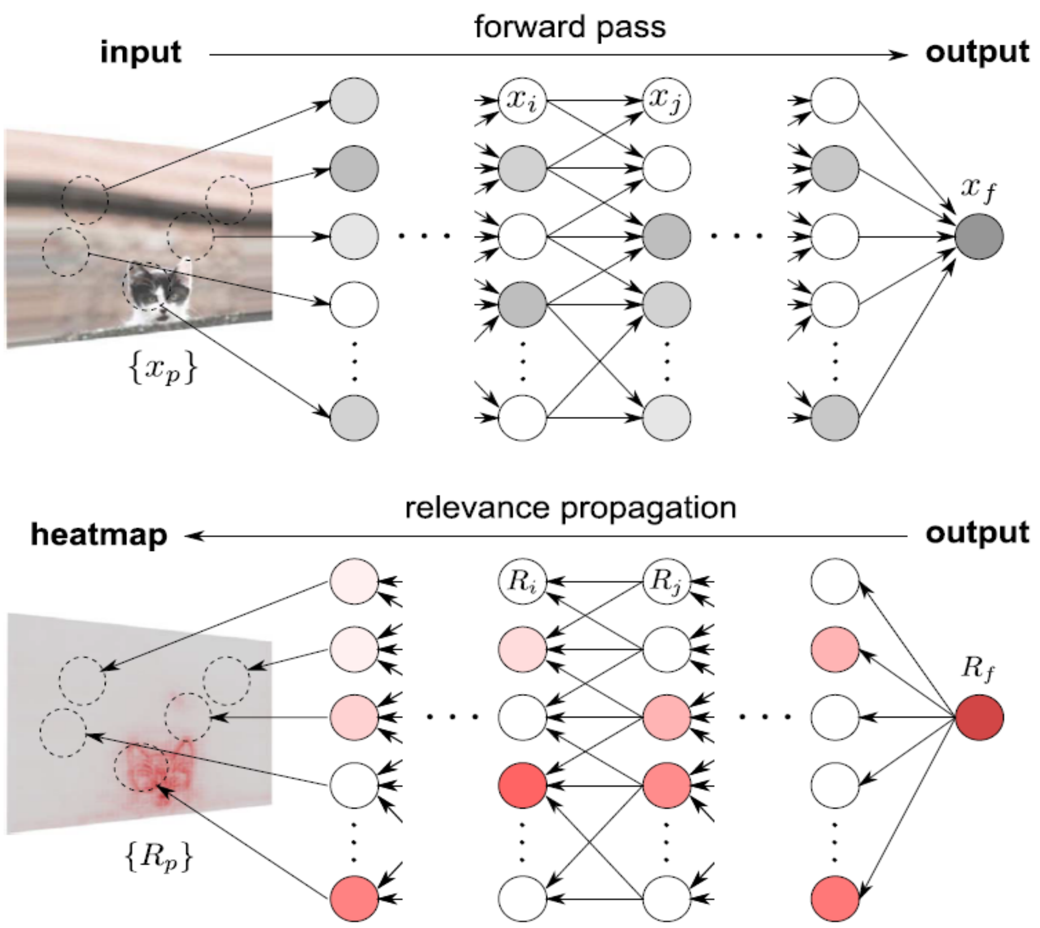
regularization methods

frequency penalization

transformation robustness

learned prior

layerwise relevance propagation (LRP)



[Montavon et al. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition.]

deep Taylor decomposition and LRP

What's the difference?

deep Taylor decomposition

$$R_d^{(1)} = (x - x_0)_{(d)} \cdot \frac{\partial f}{\partial x_{(d)}}(x_0)$$

- root point x_0 must be determined
- computationally efficient (backprop)

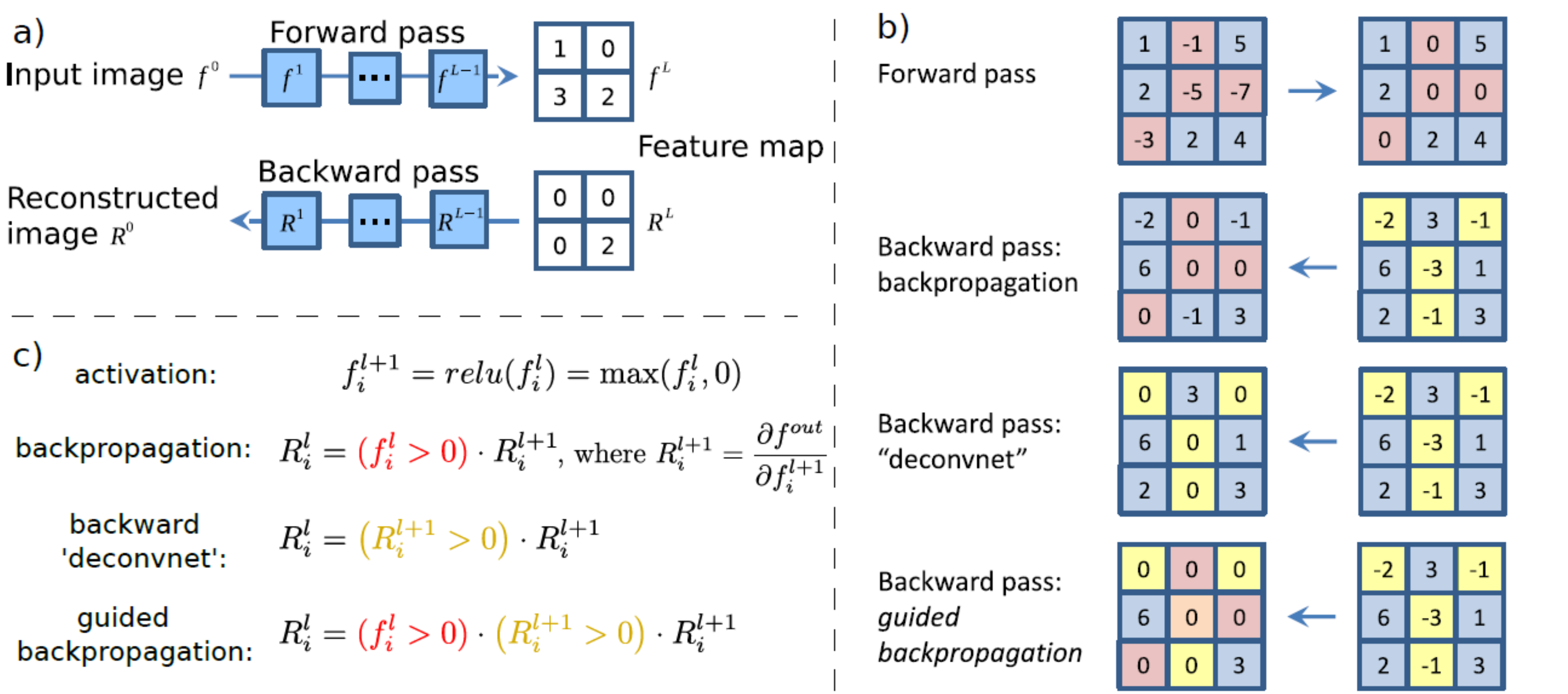
layerwise relevance propagation

$$R_{i \leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j} \cdot R_j^{(l+1)}$$

- no root point needed
- computationally expensive

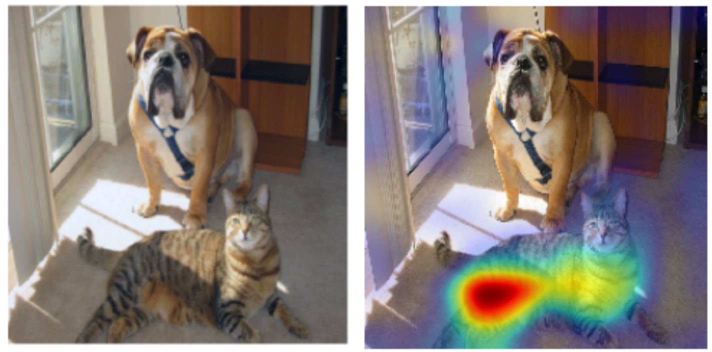
Universität Potsdam

other types of propagating output signal back to the input



[Springenberg et al. (2014). Striving for Simplicity: The All Convolutional Net]

GradCAM: Gradient-weighted Class Activation Mapping



(a) Original Image (c) Grad-CAM ‘Cat’

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

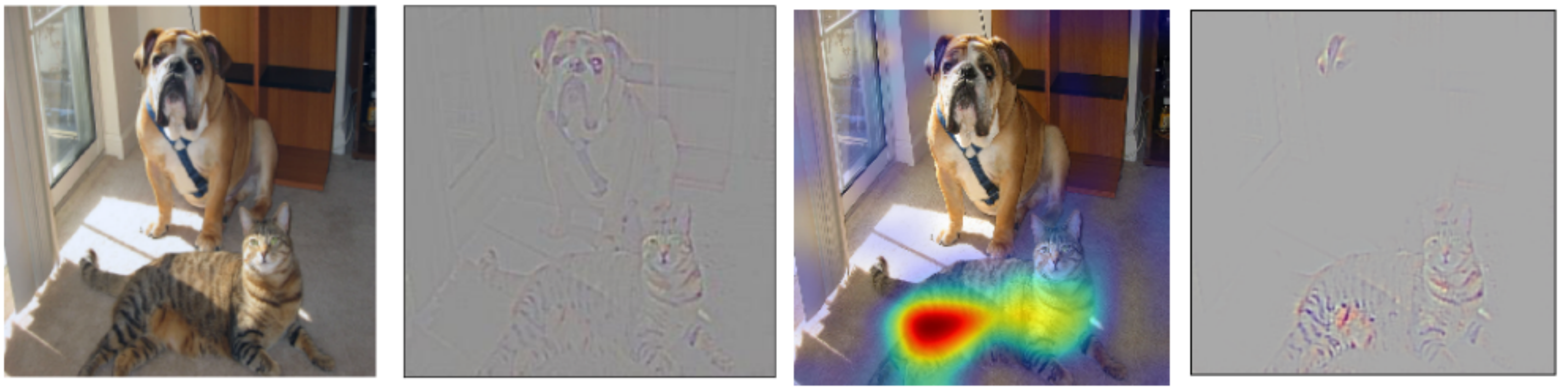
importance of feature map A^k for class c

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k \right)}_{\text{linear combination}}$$

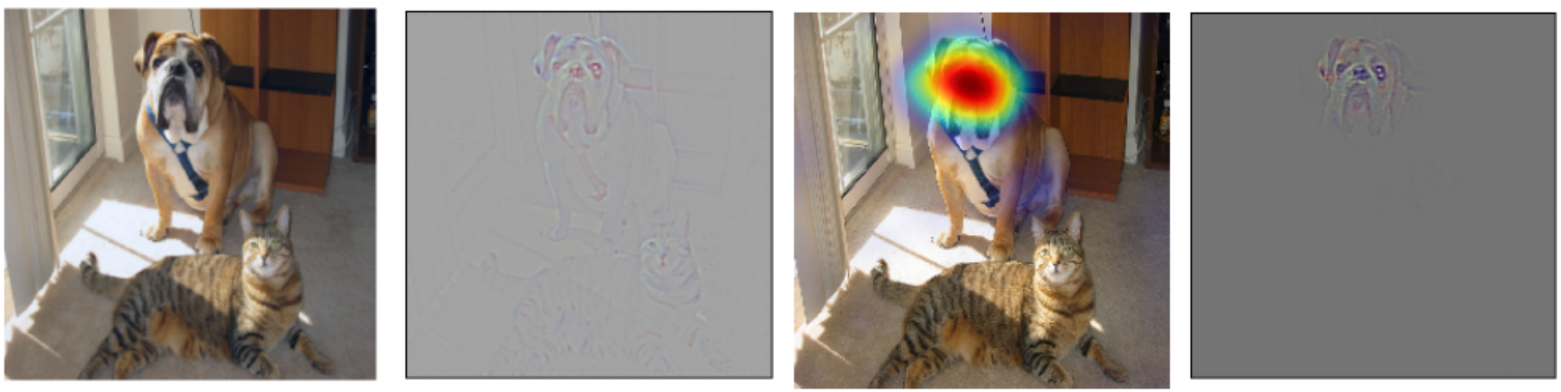
combine all feature maps A^k in one layer as weighted sum

GradCAM: Gradient-weighted Class Activation Mapping

Universität Potsdam



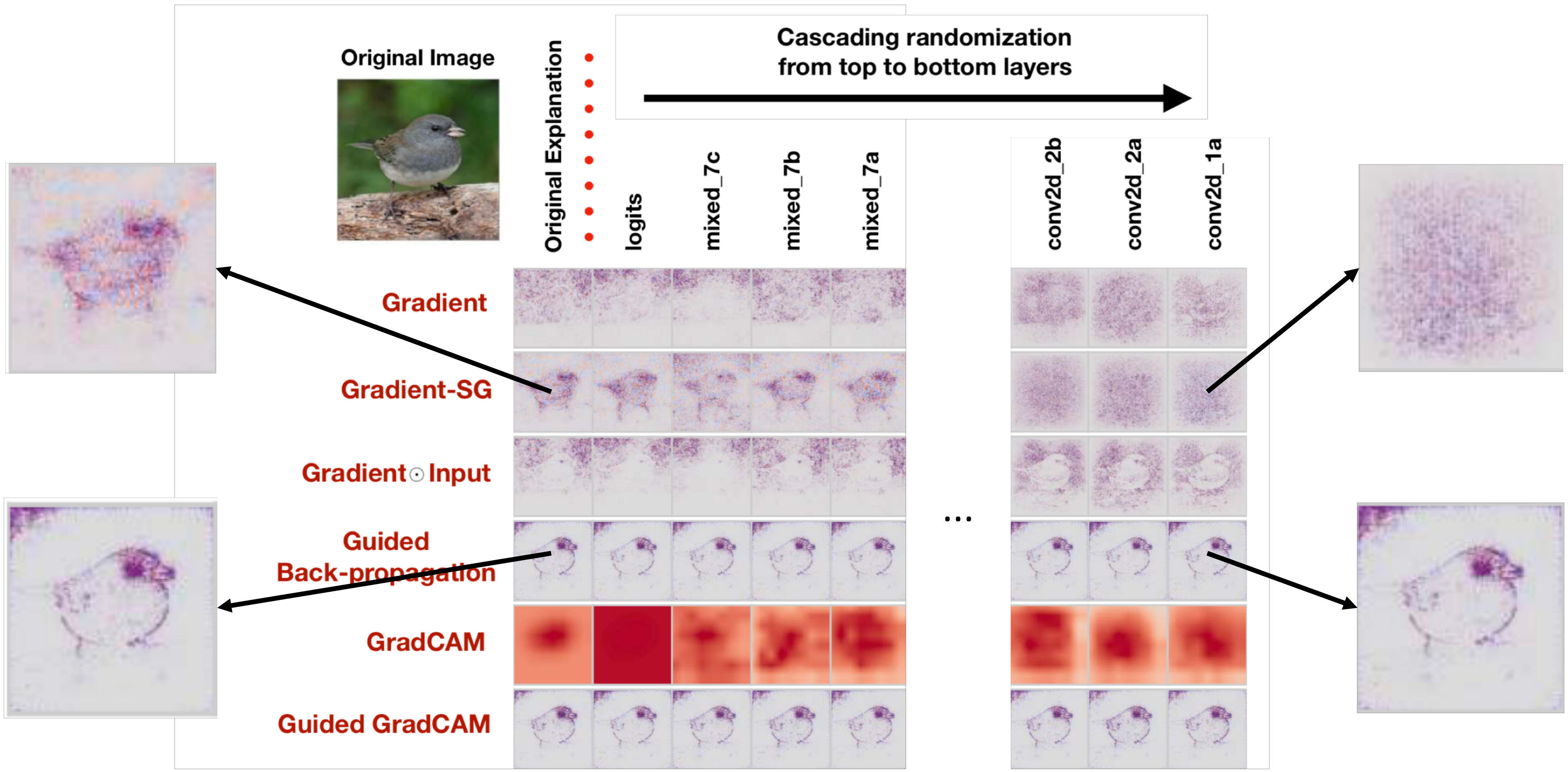
(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat'



(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog'

[Selvaraju et al. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.]

Sanity checks for introspection



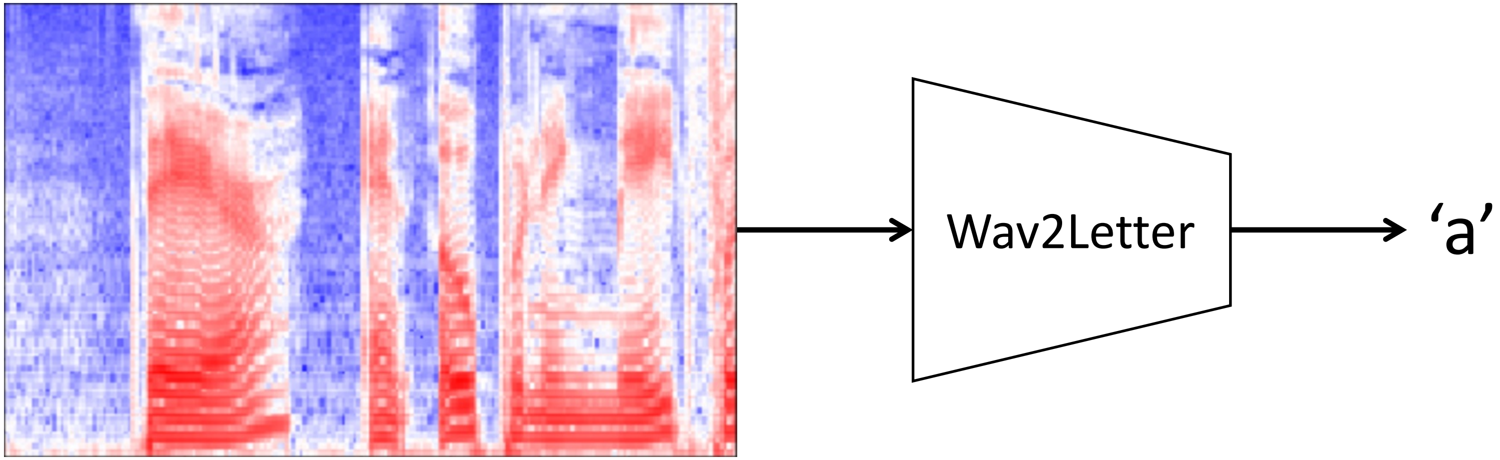
[Adebayo et al. (2018). Sanity Checks for Saliency Maps.]

Problems

- those models
  - sometimes need particular architectures
    (e.g. only 2D-convolution with max-pooling)
  - mostly use ReLUs and a positive input space
    (which pixels positively influence an output class)
  - are mostly evaluated only for images
    (visually interpretable)
- not well applicable for
  - other activation functions (allowing negative activation)
  - real-valued input space (negative values)
  - visually hardly interpretable data (e.g. waveforms)

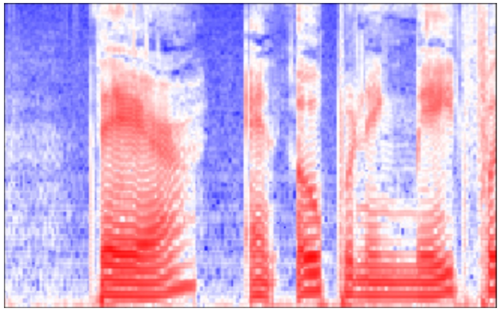
Introspection for ASR

- apply introspection to a image-like ASR task
- Speech recognizer based on Wav2Letter
- Audio (z-normalized Mel-Spectrogram) → 1D-ConvNet → letter prediction

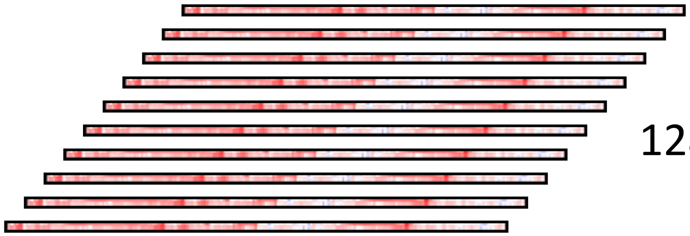


Introspection for ASR

- not 2D



=



- large layers for non-Taylor LRP

- ReLUs, but negative input values

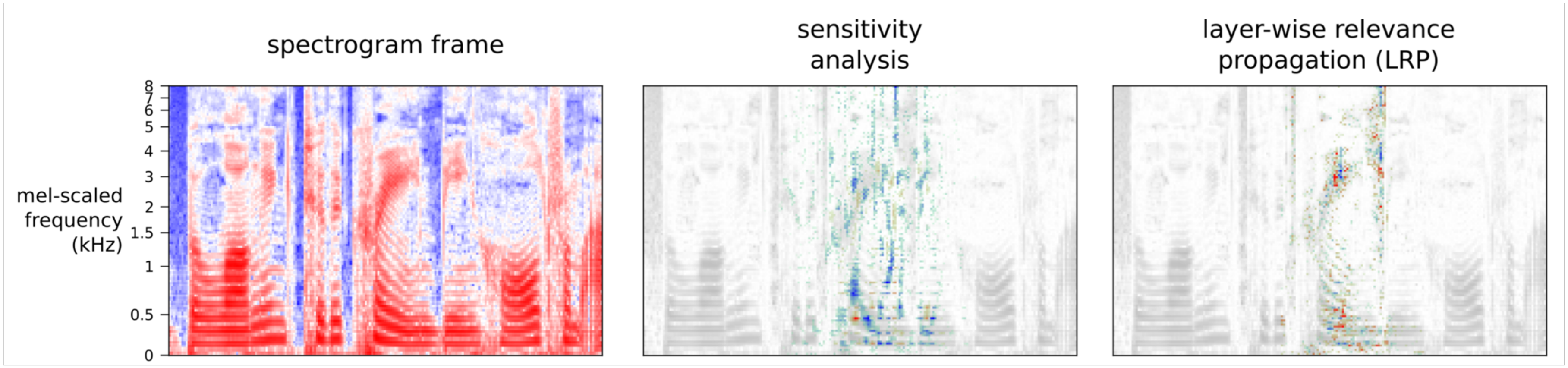
- batch norm between convolution and ReLU activation

- predicting several 100 letters per sentence (from overlapping windows)

Introspection for ASR

predicted as letter 'a'

w.r.t. output logit for letter 'a'



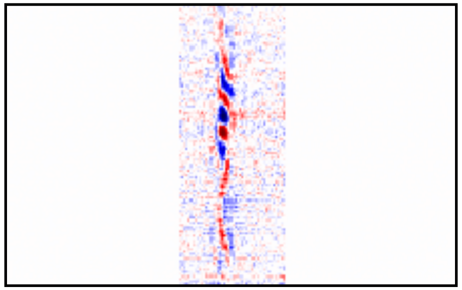
introspection methods that work for images are not easy to interpret

and only tell about one specific input (at one time step)

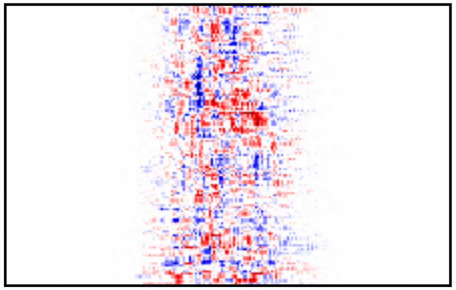
Introspection for ASR

- ‘global’ introspection
- feature visualization → optimal input for a particular letter

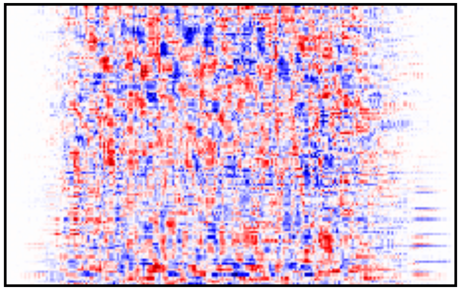
optimal inputs (weakly regularized)



first layer



a middle layer

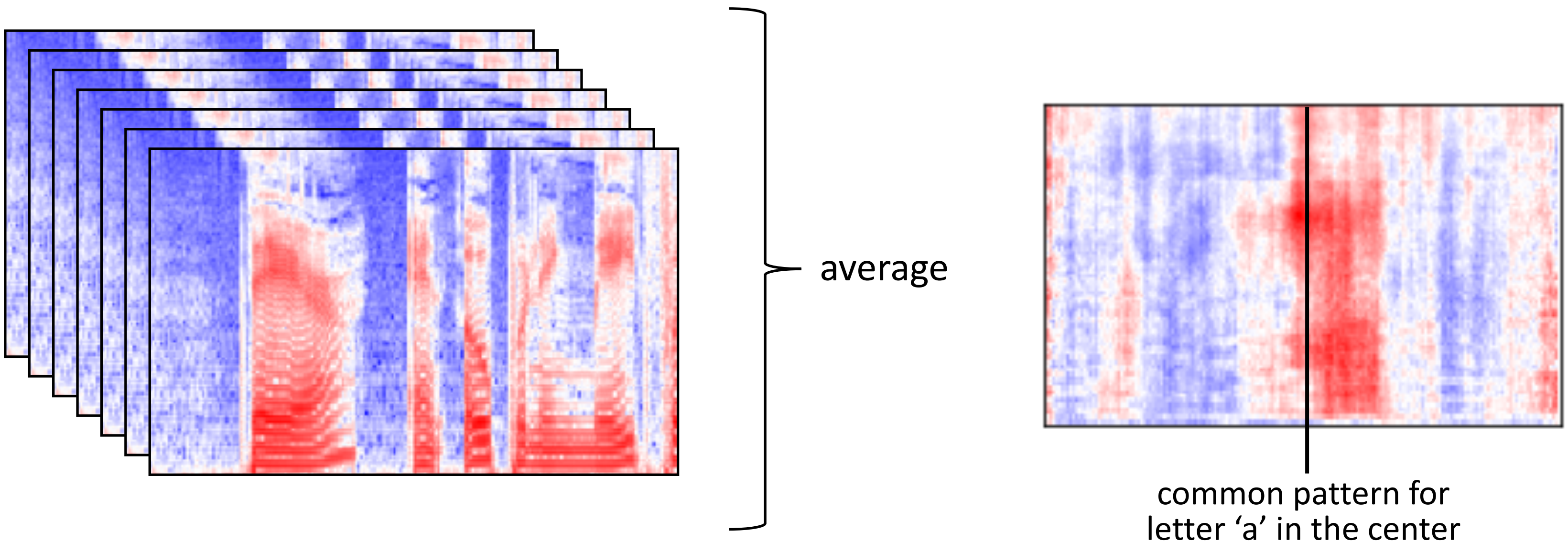


output layer

- nothing useful to learn from this (in contrast to images)

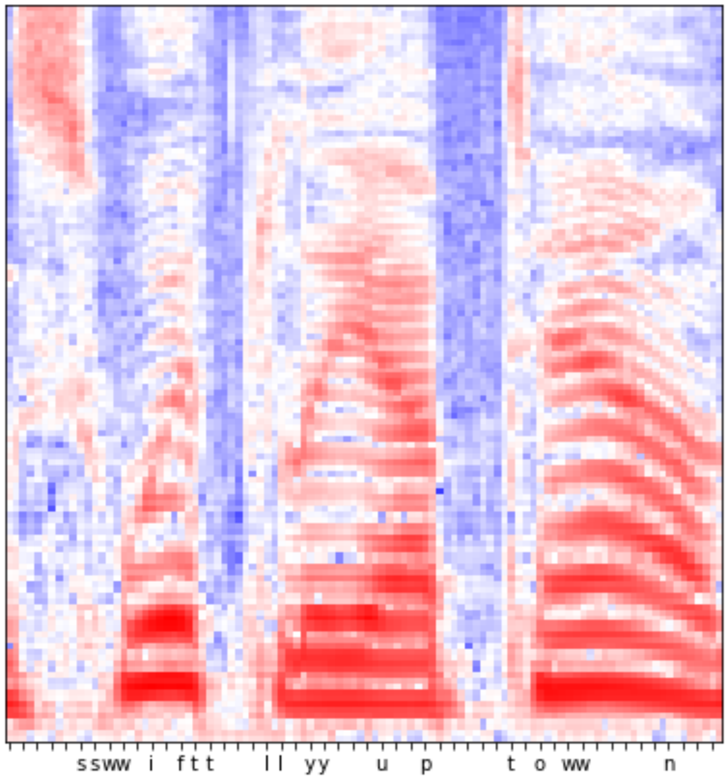
Introspection for ASR

- ‘global’ introspection
- using (all) test/training data and their predictions



- Requirement: prediction and spectrogram need to be aligned

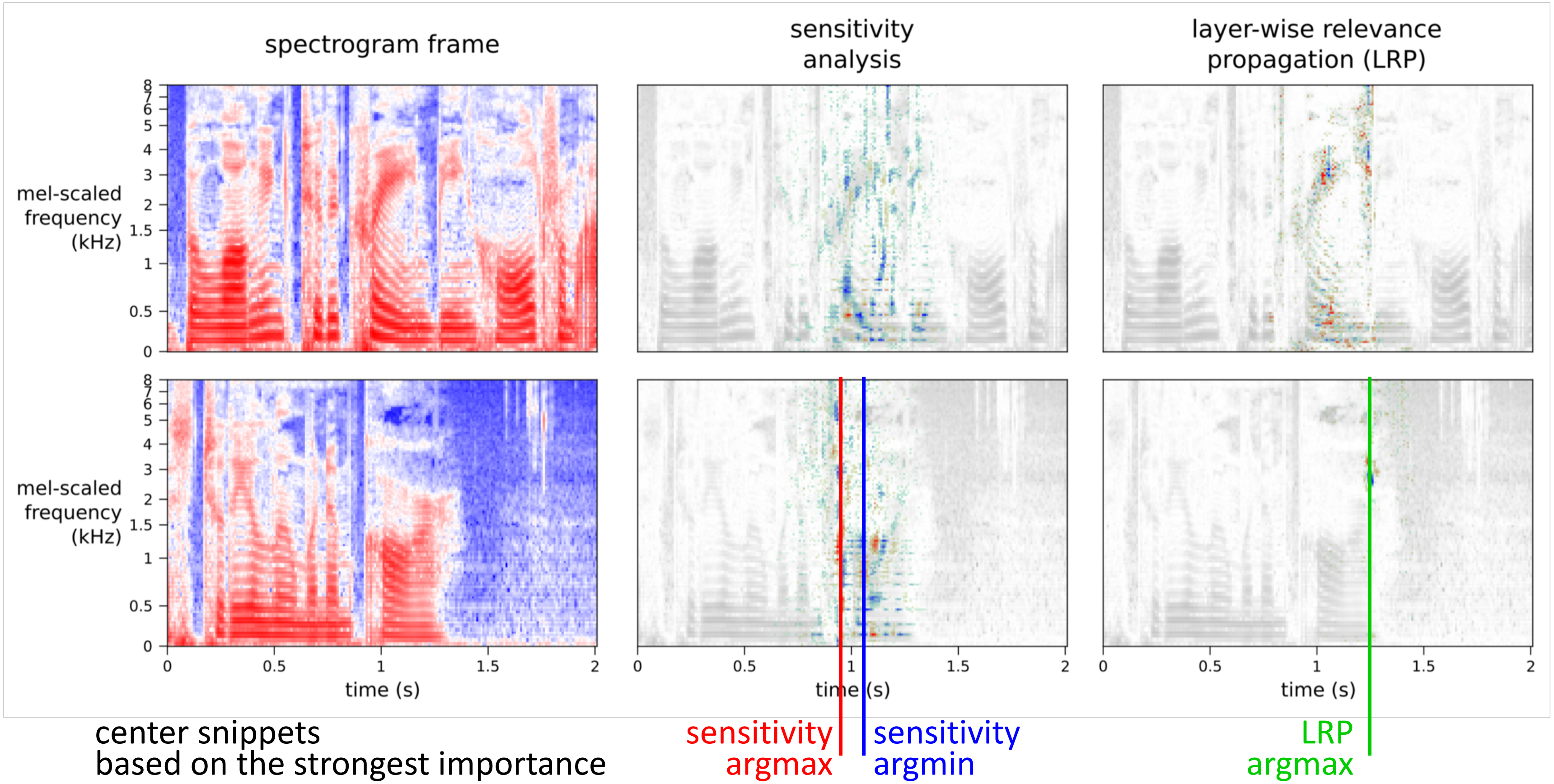
Introspection for ASR



! predictions are not aligned with the position in the spectrogram

→ sensitivity/LRP could help here

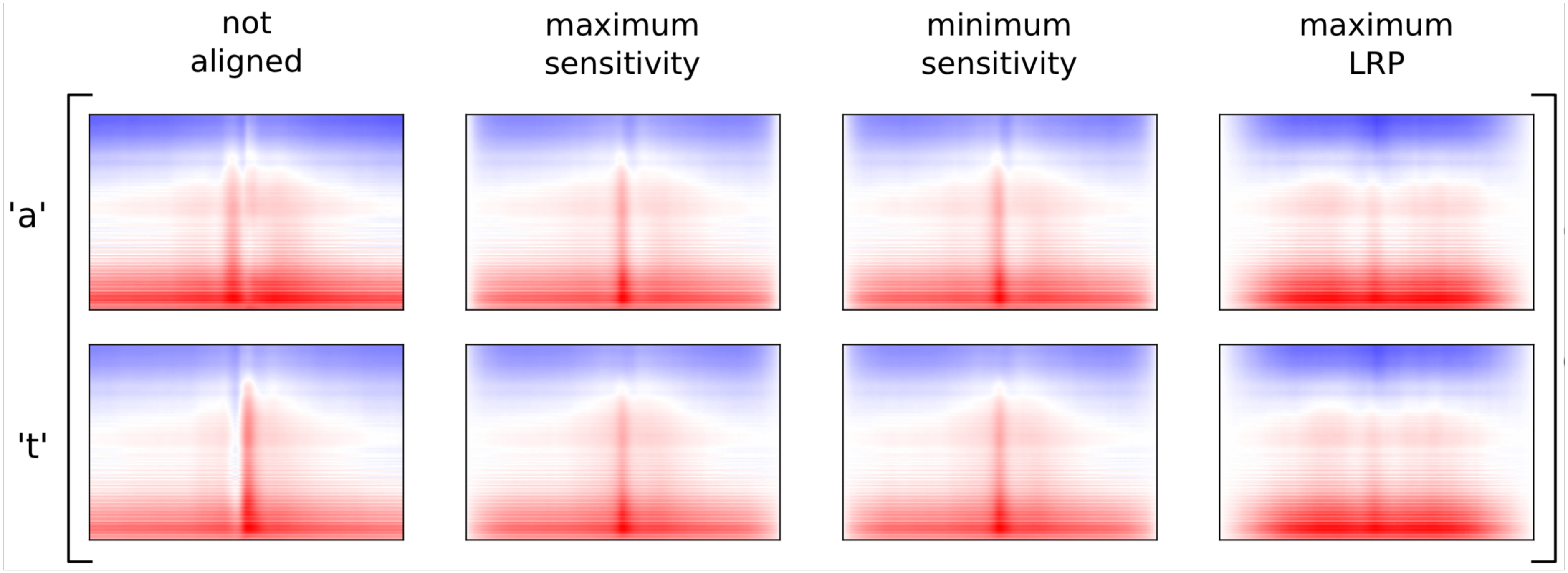
Introspection for ASR



center snippets
based on the strongest importance

Introspection for ASR

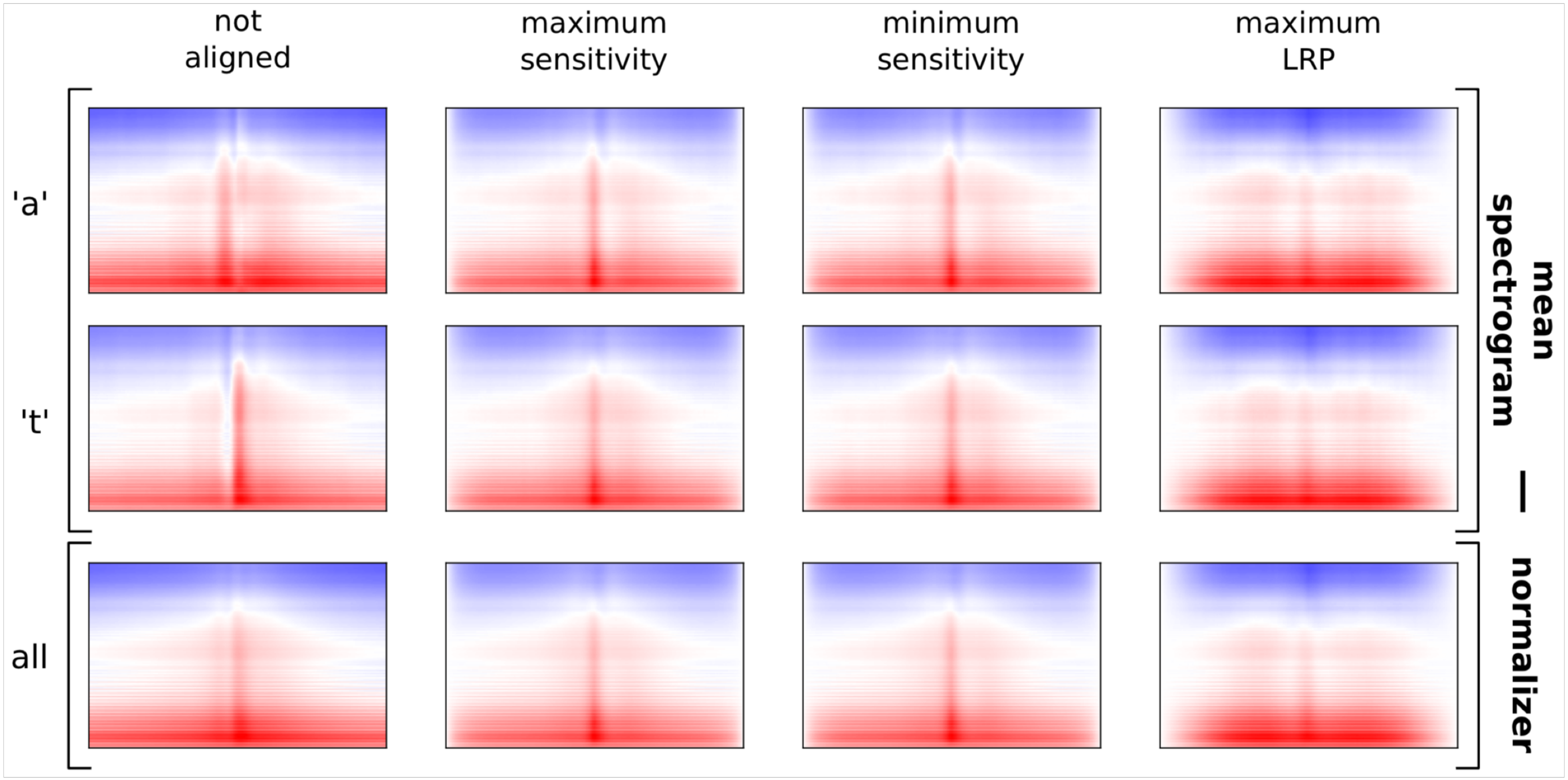
with the aligned spectrogram frames, we can try to find common patterns



letter-specific information are overshadowed by being a letter at all

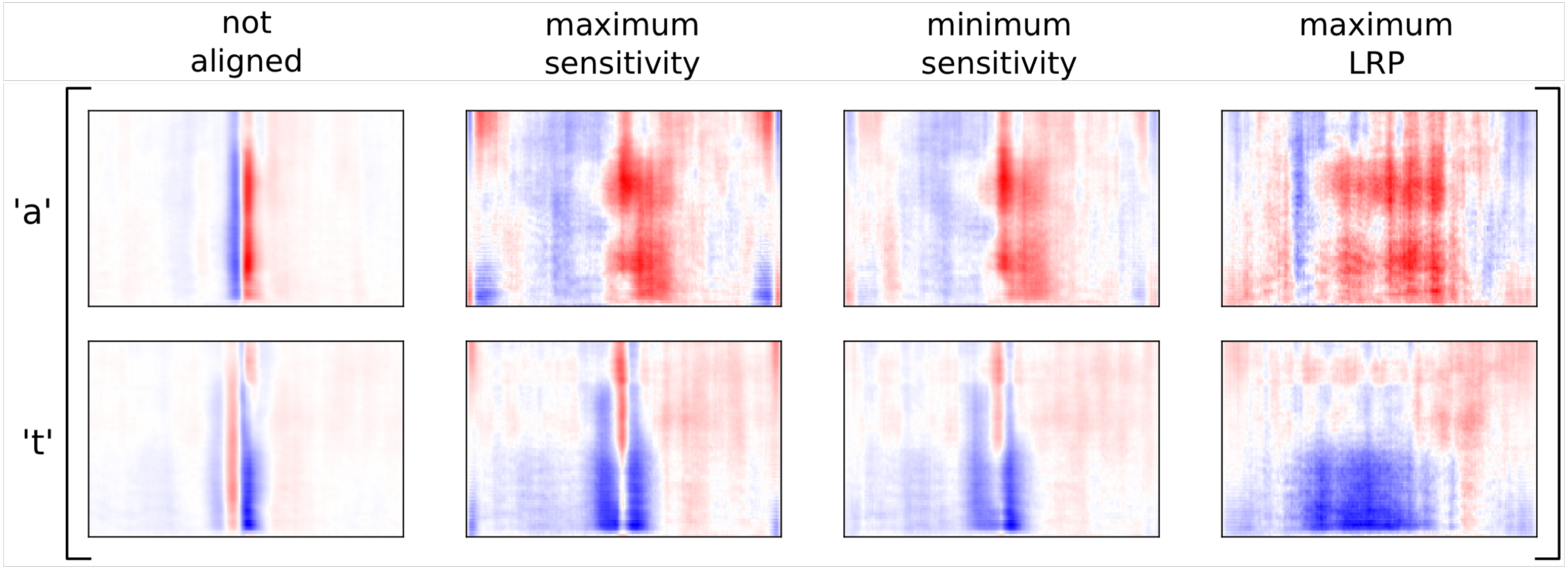
Introspection for ASR

subtract the average over all spectrogram frames for all letters



Introspection for ASR

after subtracting the average over all spectrogram frames of all letters

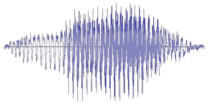


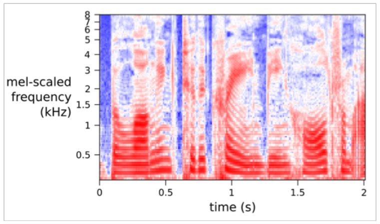
we get letter-specific patterns (not spectrograms!)
from which we can learn something about the model

[Krug et al. (2018). Introspection for convolutional automatic speech recognition.]

Neuron Activation Profiles

- We have:
  - little intuition about input signal
  - more intuition about the output
    ‘SPEECH’ → /S P IY CH/

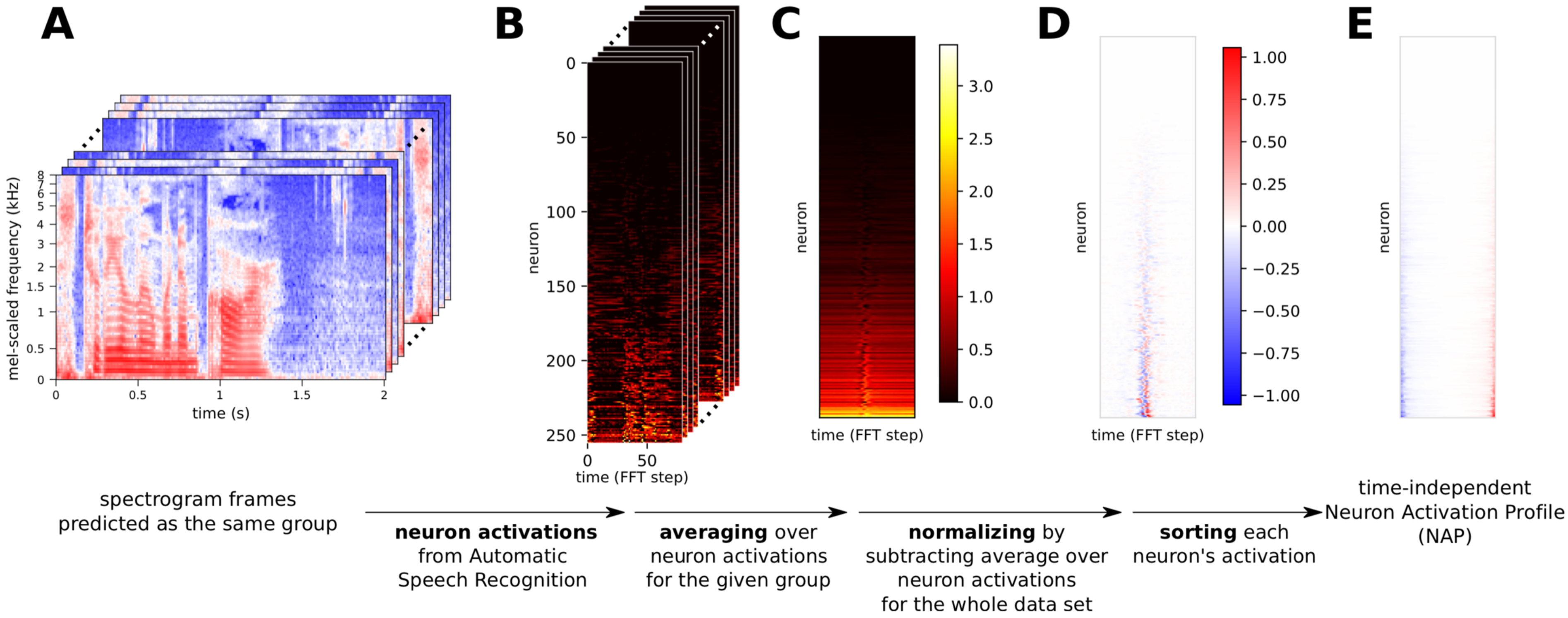




- Perform introspection on the output space?

- Comparison of characteristic neuron responses to letters and phonemes by clustering

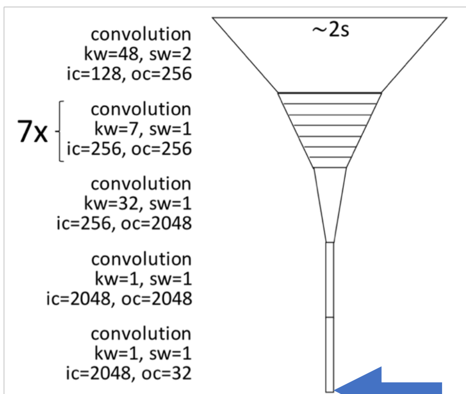
Neuron Activation Profiles



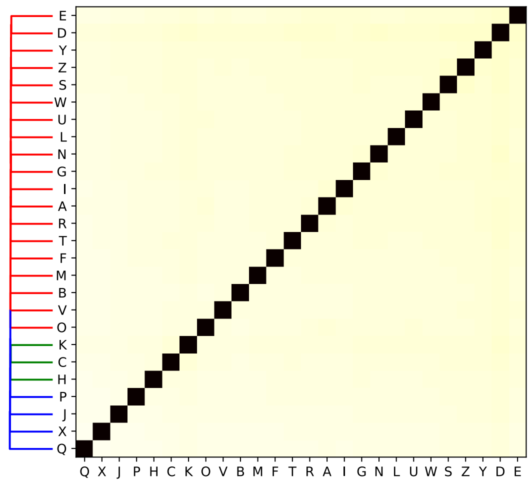
[Krug et al. (2018). Neuron Activation Profiles for Interpreting Convolutional Speech Recognition Models.]

Neuron Activation Profiles

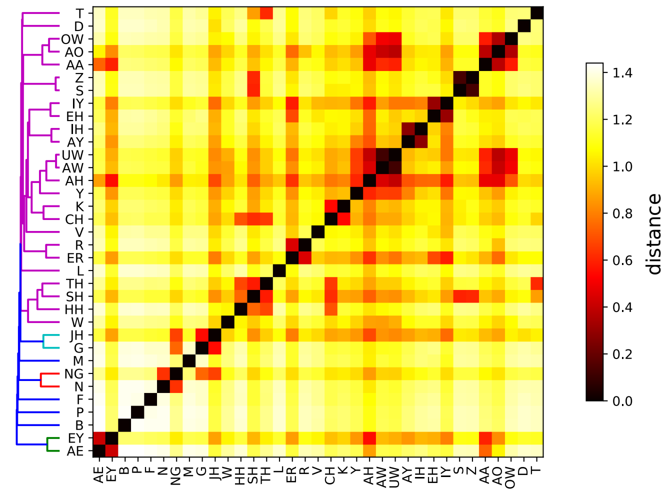
Clustering of NAPs in the output layer



letters



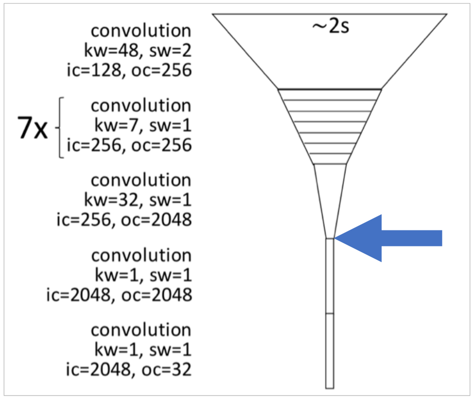
phonemes



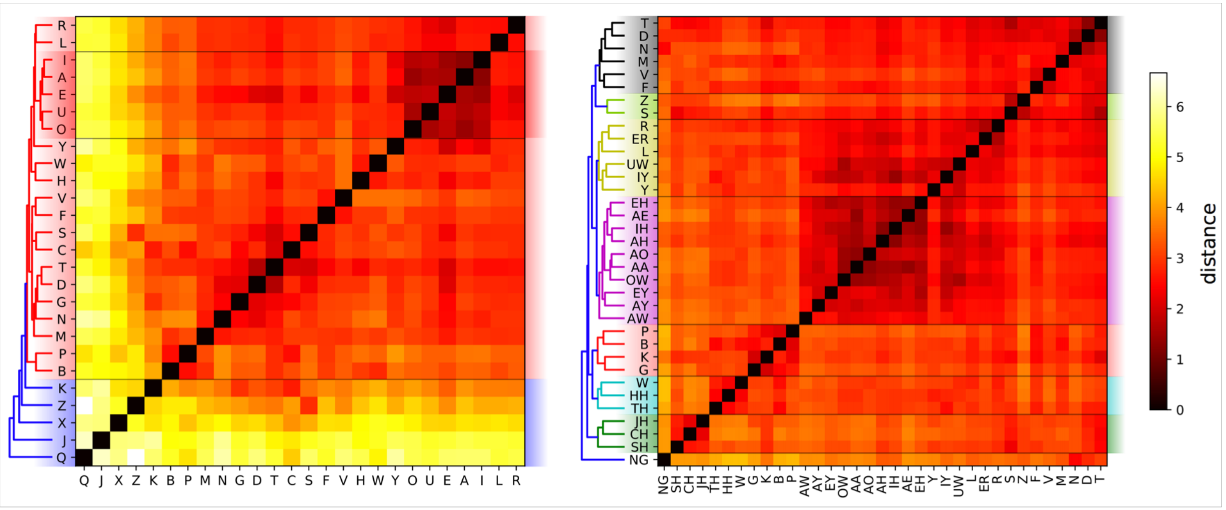
- (obviously) letters are best encoded in the output layer
- phonemes NAPs are only similar for very small sets of phonemes

Neuron Activation Profiles

Clustering of NAPs in the 9th layer



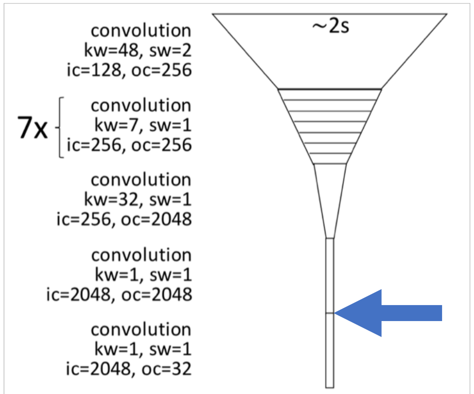
letters | phonemes

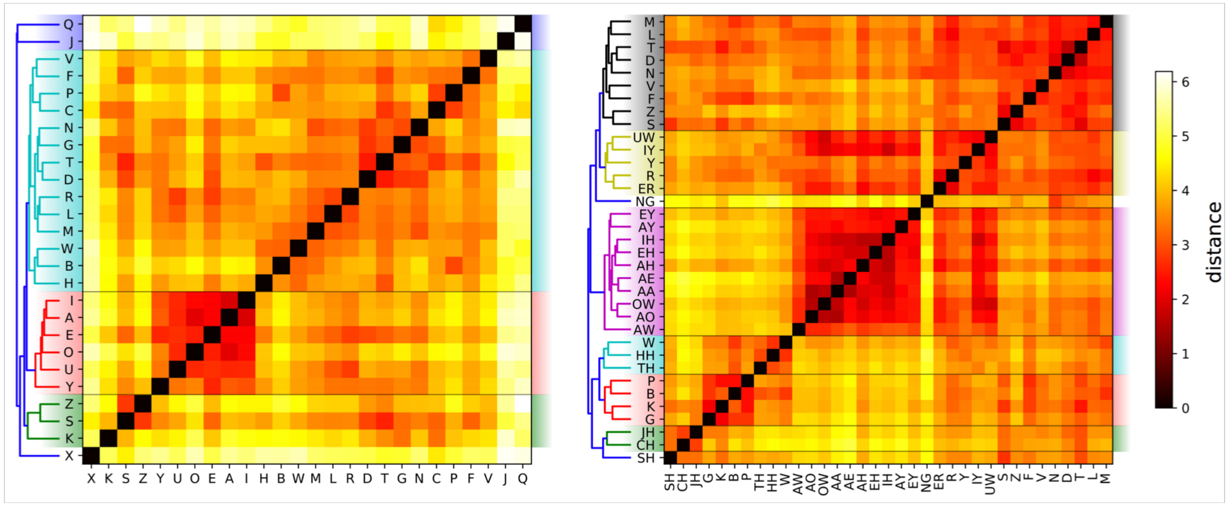


- clusters of similar phonemes emerge
- no distinct clustering of NAPs for letters

Neuron Activation Profiles

Clustering of NAPs in the 10th layer





- phoneme clusters become more distinct
  - cluster of vowel letters emerges

Universität Potsdam

Assignments

- Reading on the topics you liked to discuss again
  look for additional material (videos, blogs)
  post questions on MM!
  - RNNs, LSTMs
  - Attention (& Transformer)
  - backpropagation
- Work on your projects
- PEP-evaluation until last session (Feb 04)
  Link on Mattermost

Slides & assignments on: https://mlcogup.github.io/idl_ws18/schedule