Interpreting intermediate convolutional layers in unsupervised acoustic word classification

Gašper Beguš & Alan Zhou

{begus,azhou314}@berkeley.edu

SPE-76, May 12, 2022

Introduction

1. Understanding how deep neural networks learn representations internally
2. Most existing studies focus on the visual domain

Introduction

1. **A technique to visualize and interpret intermediate layers of unsupervised deep convolutional classifiers**
2. Introduce Generalized Additive Mixed Models (GAMMs) to the paradigm
3. Using non-linear regression, we infer underlying distributions for each word
4. Analyze both absolute values and shapes of individual words at different convolutional layers
5. Hypothesis testing

Introduction

1. A technique to visualize and interpret intermediate layers of unsupervised deep convolutional classifiers
2. **Introduce Generalized Additive Mixed Models (GAMMs) to the paradigm**
3. Using non-linear regression, we infer underlying distributions for each word
4. Analyze both absolute values and shapes of individual words at different convolutional layers
5. Hypothesis testing

Introduction

1. A technique to visualize and interpret intermediate layers of unsupervised deep convolutional classifiers
2. Introduce Generalized Additive Mixed Models (GAMMs) to the paradigm
3. **Using non-linear regression, we infer underlying distributions for each word**
4. Analyze both absolute values and shapes of individual words at different convolutional layers
5. Hypothesis testing

Introduction

1. A technique to visualize and interpret intermediate layers of unsupervised deep convolutional classifiers
2. Introduce Generalized Additive Mixed Models (GAMMs) to the paradigm
3. Using non-linear regression, we infer underlying distributions for each word
4. **Analyze both absolute values and shapes of individual words at different convolutional layers**
5. Hypothesis testing

Introduction

1. A technique to visualize and interpret intermediate layers of unsupervised deep convolutional classifiers
2. Introduce Generalized Additive Mixed Models (GAMMs) to the paradigm
3. Using non-linear regression, we infer underlying distributions for each word
4. Analyze both absolute values and shapes of individual words at different convolutional layers
5. **Hypothesis testing**

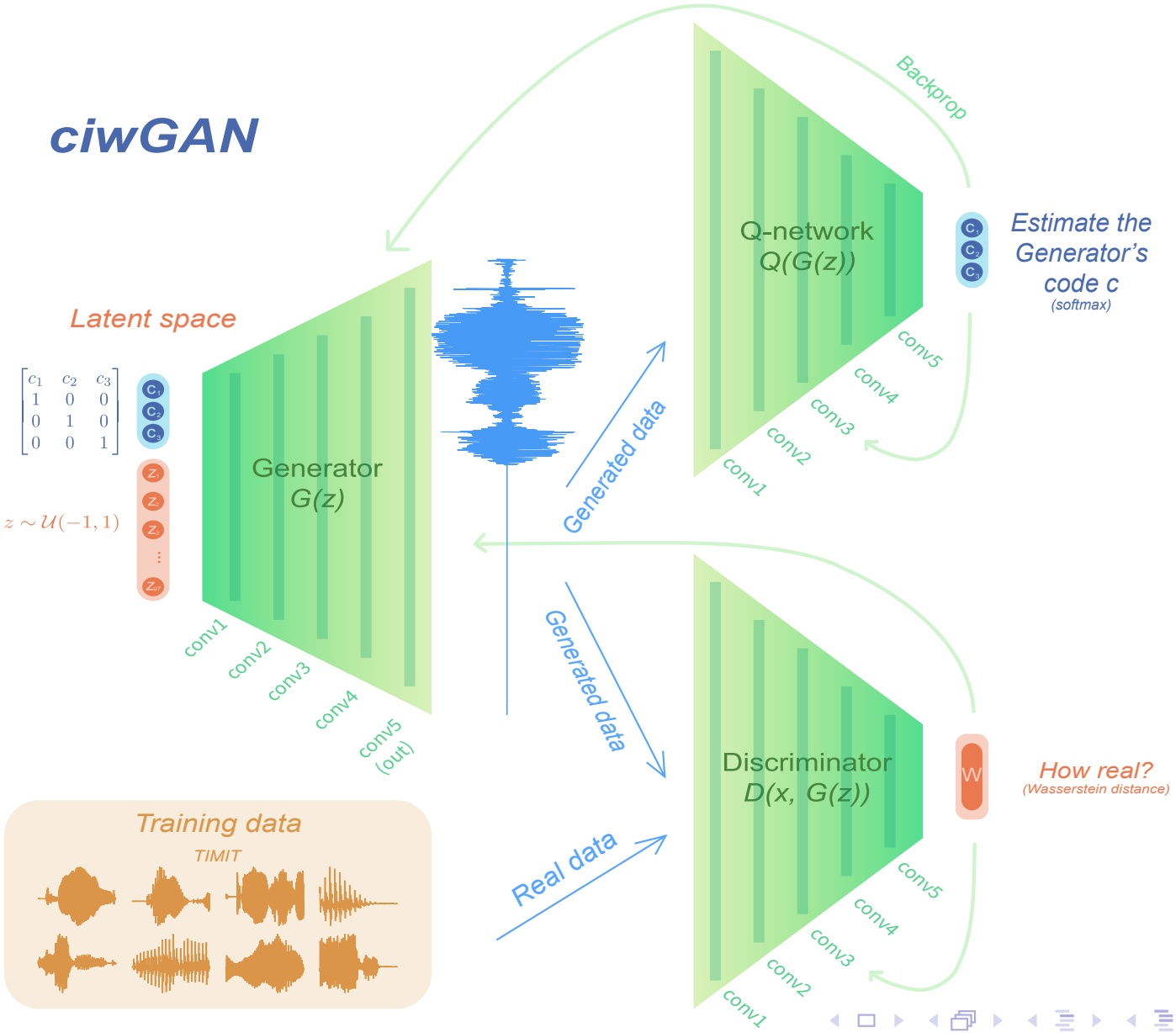
Proposal

Proposal

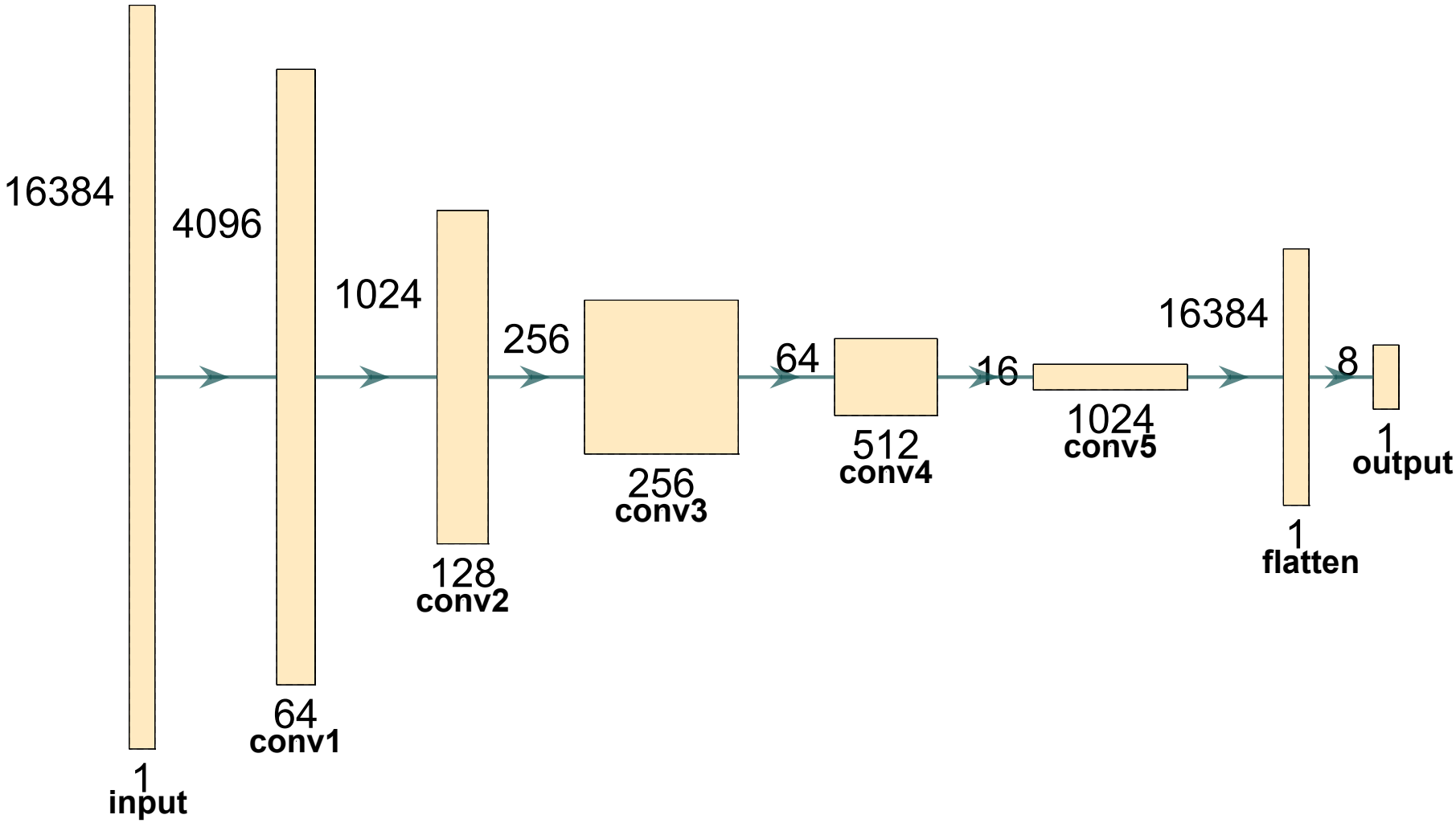
Averaging over individual feature maps in each convolutional layer after the Leaky ReLU activation in the Q-network (classifier) yields highly interpretable time-series data that summarizes which linguistic properties are encoded at which layer.

- The technique applied to the Generator: Beguš and Zhou (2021)

The model (Beguš, 2021)



The Q-network

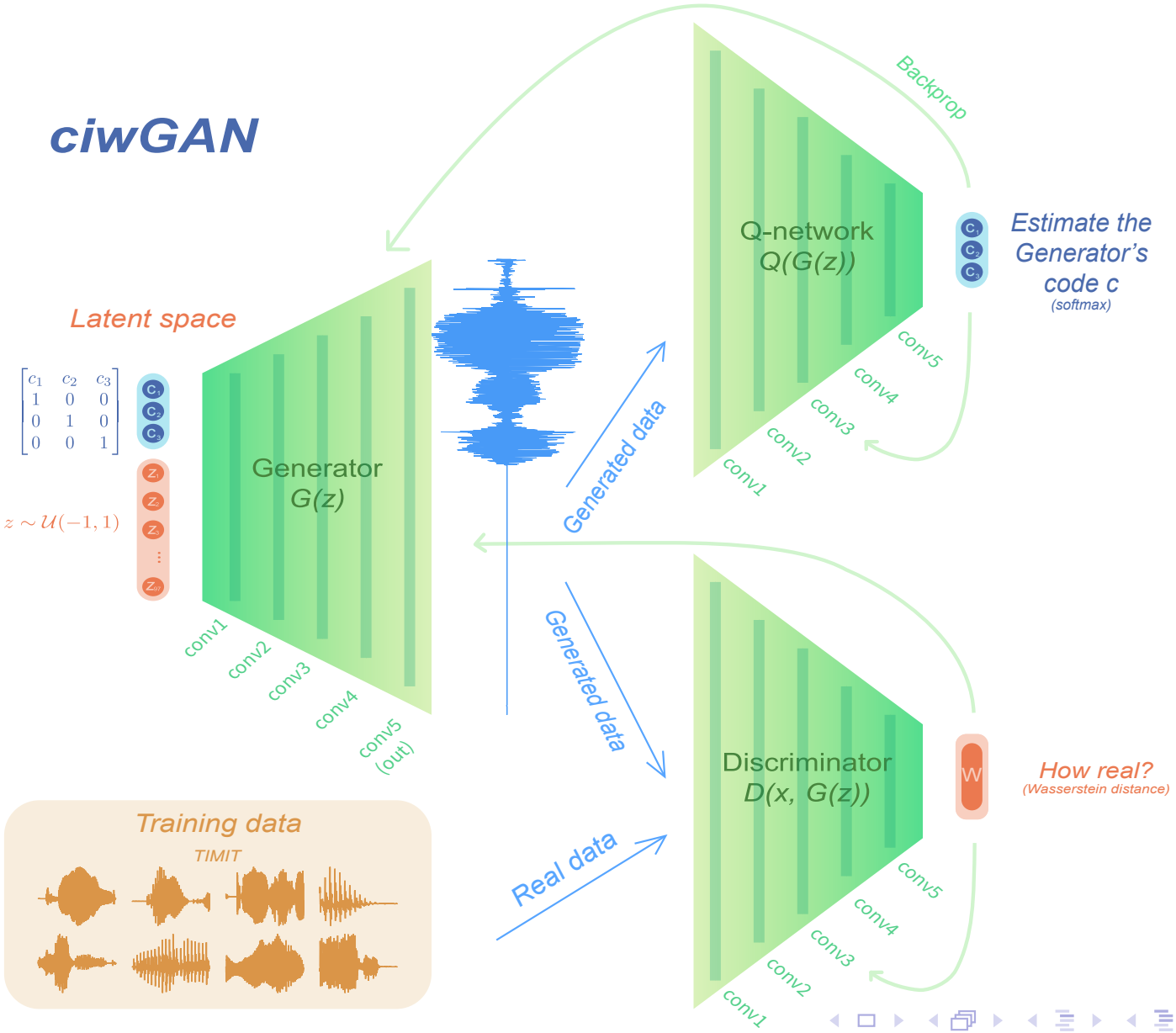


(Goodfellow et al., 2014; Radford et al., 2015; Donahue et al., 2019)

Training data

- ciwGAN trained on eight unlabeled single words from TIMIT Garofolo et al. (1993):
  ask, dark, greasy, oily, rag, wash, water, and year
- 80% (altogether 4,052 tokens) **training** data
- 20% (altogether 1,067 tokens) **test** data

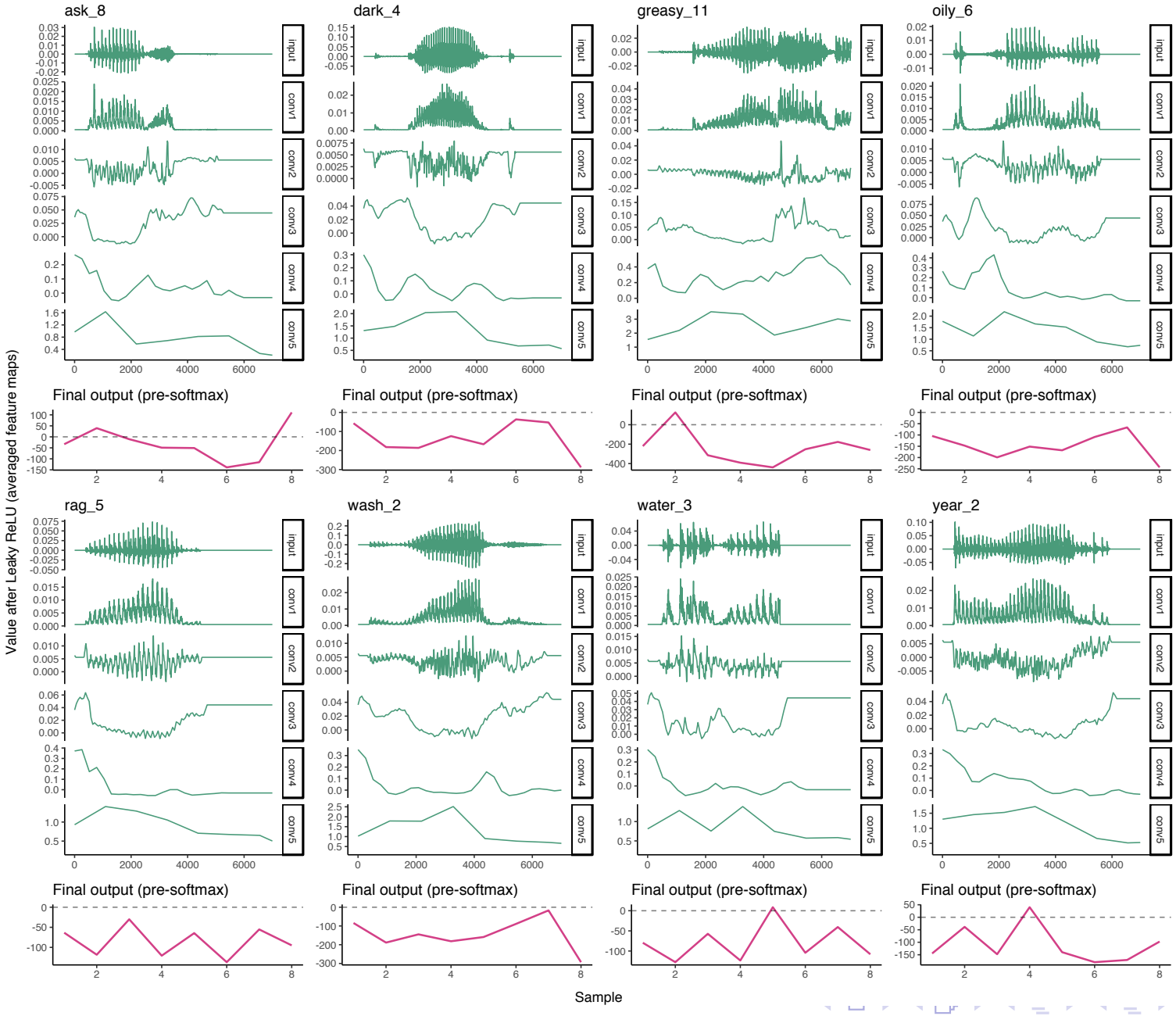
The model (Beguš, 2021)



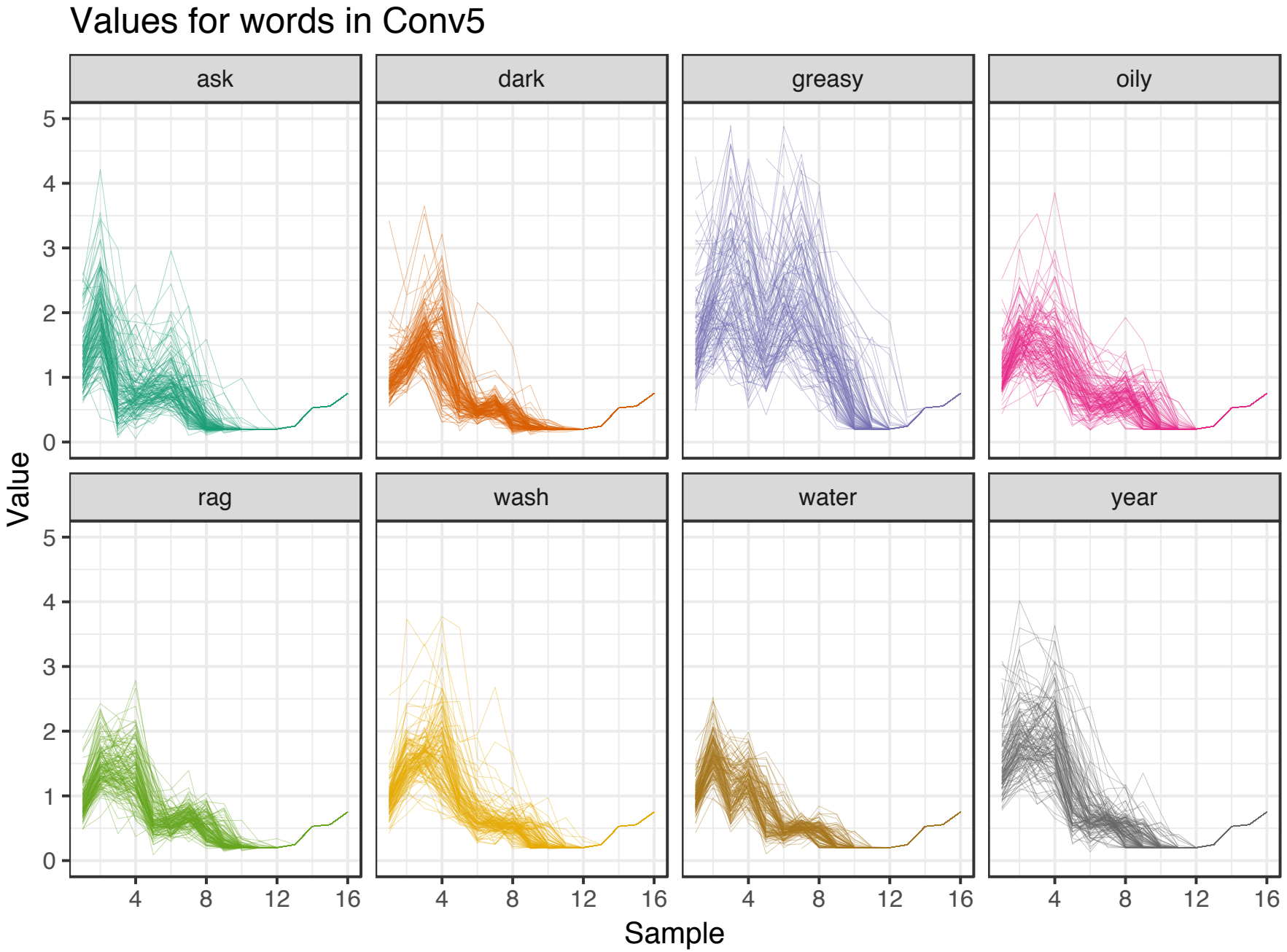
Averaging over feature maps

$$t = \frac{1}{||C||} \sum_{i=1}^{||C||} C_i \tag{1}$$

Individual words



Several trajectories

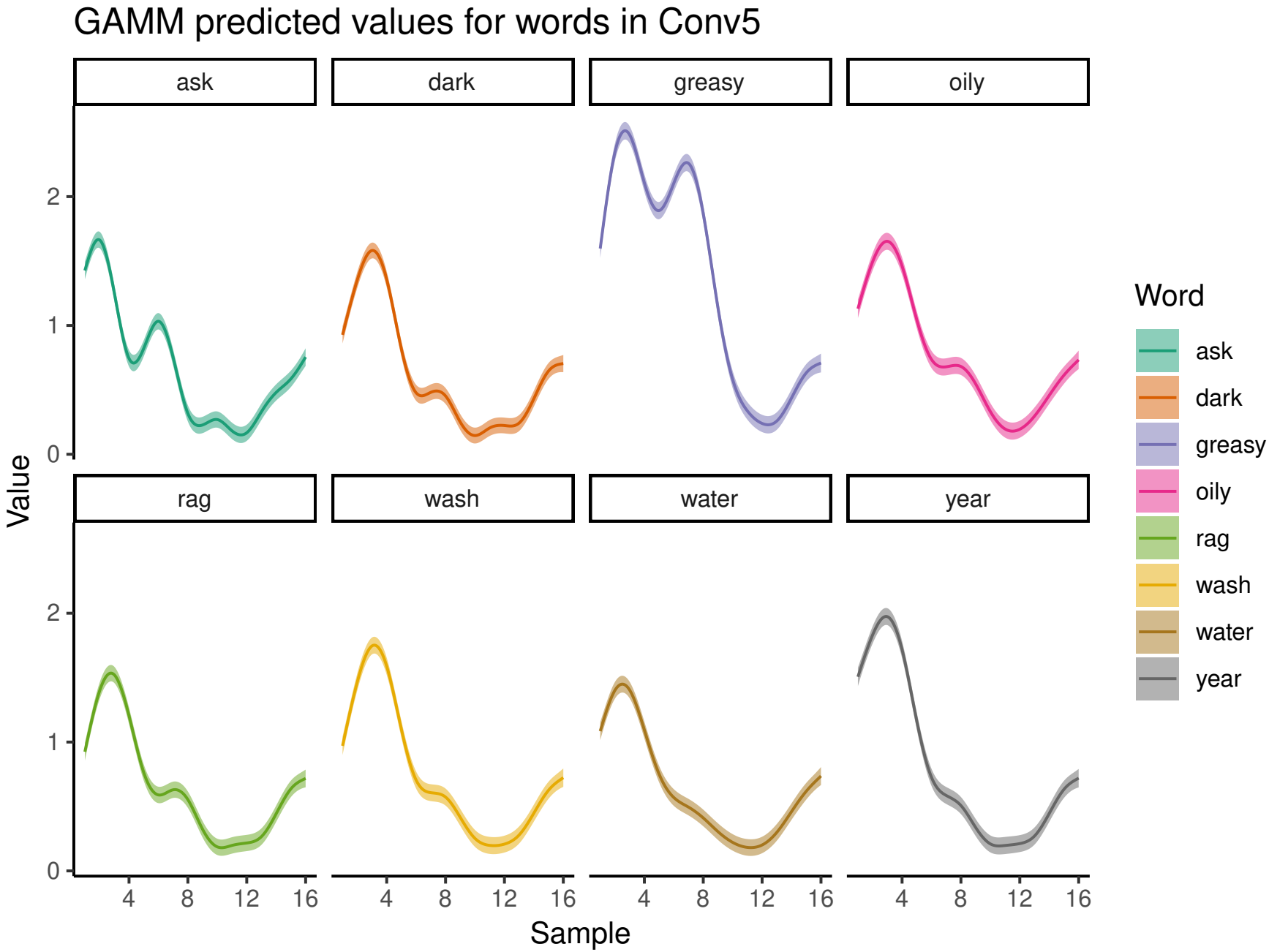


Generalized Additive Mixed Models

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	0.6832	0.0215	31.7138	< 0.0001
Word=dark	-0.0409	0.0300	-1.3626	0.1730
Word=greasy	0.6323	0.0314	20.1116	< 0.0001
Word=oily	0.0811	0.0312	2.6021	0.0093
Word=rag	-0.0240	0.0305	-0.7870	0.4313
Word=wash	0.0603	0.0311	1.9425	0.0521
Word=water	-0.0464	0.0307	-1.5126	0.1304
Word=year	0.1298	0.0311	4.1712	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Sample)	8.9700	8.9934	235.5282	< 0.0001
s(Sample):Word=dark	8.9684	8.9936	159.7072	< 0.0001
s(Sample):Word=greasy	8.9446	8.9886	231.4914	< 0.0001
s(Sample):Word=oily	8.9562	8.9911	122.4444	< 0.0001
s(Sample):Word=rag	8.9534	8.9904	110.0716	< 0.0001
s(Sample):Word=wash	8.9625	8.9924	165.8256	< 0.0001
s(Sample):Word=water	8.9309	8.9855	71.5510	< 0.0001
s(Sample):Word=year	8.9613	8.9922	140.2448	< 0.0001
s(Sample,UniqueWord)	5327.3250	9595.0000	4.8275	< 0.0001

Table: Regression estimates of the GAMM model.

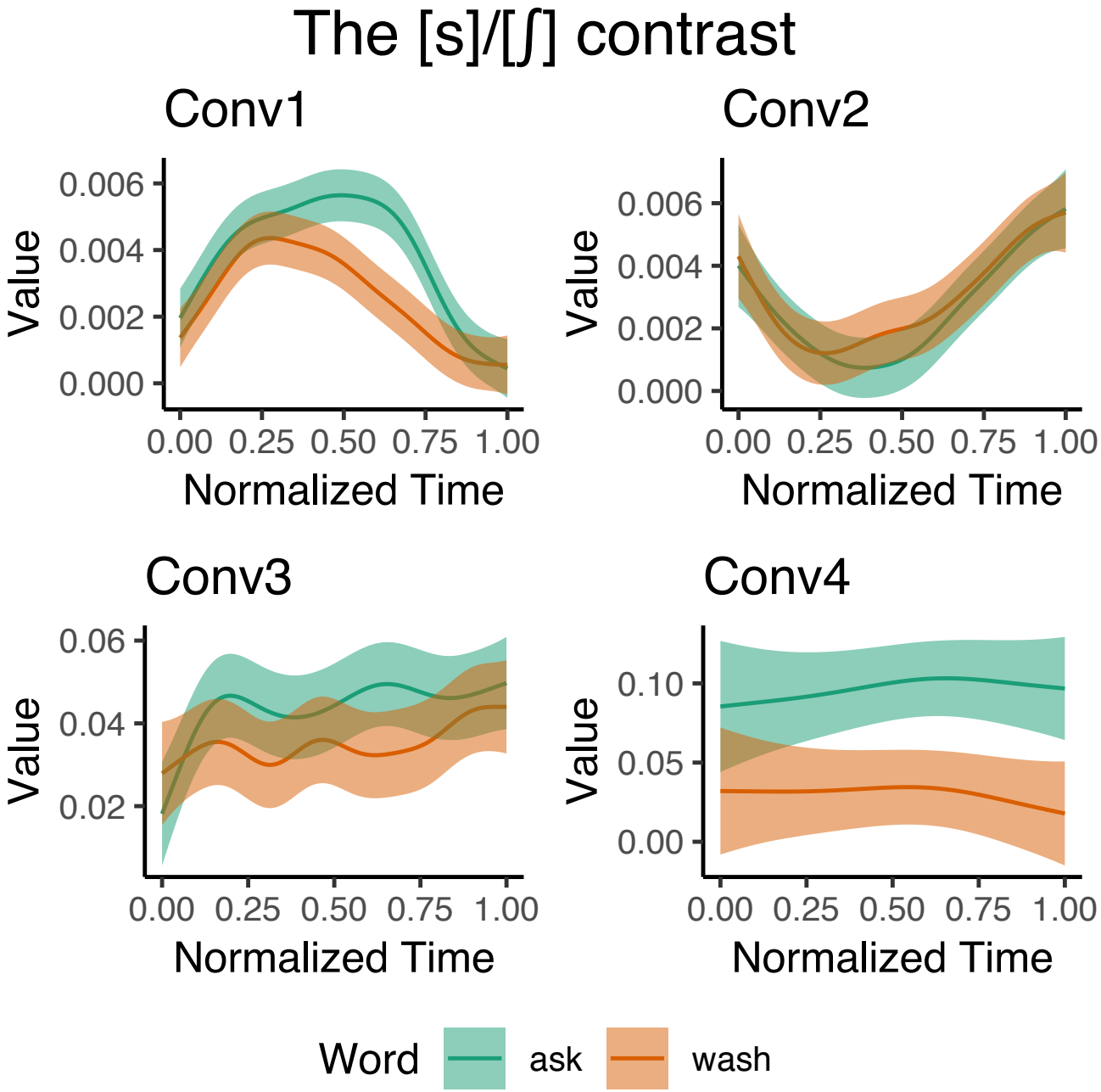
Inferring shapes of words



Shape encoding vs. value encoding

- With the technique, we can test both word-level and phone-level encoding
- **Shape encoding** vs. **value encoding**

Shape encoding vs. value encoding



Generalized Additive Mixed Models

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = ask	0.6832	0.0215	31.7138	< 0.0001
Word=dark	-0.0409	0.0300	-1.3626	0.1730
Word=greasy	0.6323	0.0314	20.1116	< 0.0001
Word=oily	0.0811	0.0312	2.6021	0.0093
Word=rag	-0.0240	0.0305	-0.7870	0.4313
Word=wash	0.0603	0.0311	1.9425	0.0521
Word=water	-0.0464	0.0307	-1.5126	0.1304
Word=year	0.1298	0.0311	4.1712	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Sample)	8.9700	8.9934	235.5282	< 0.0001
s(Sample):Word=dark	8.9684	8.9936	159.7072	< 0.0001
s(Sample):Word=greasy	8.9446	8.9886	231.4914	< 0.0001
s(Sample):Word=oily	8.9562	8.9911	122.4444	< 0.0001
s(Sample):Word=rag	8.9534	8.9904	110.0716	< 0.0001
s(Sample):Word=wash	8.9625	8.9924	165.8256	< 0.0001
s(Sample):Word=water	8.9309	8.9855	71.5510	< 0.0001
s(Sample):Word=year	8.9613	8.9922	140.2448	< 0.0001
s(Sample,UniqueWord)	5327.3250	9595.0000	4.8275	< 0.0001

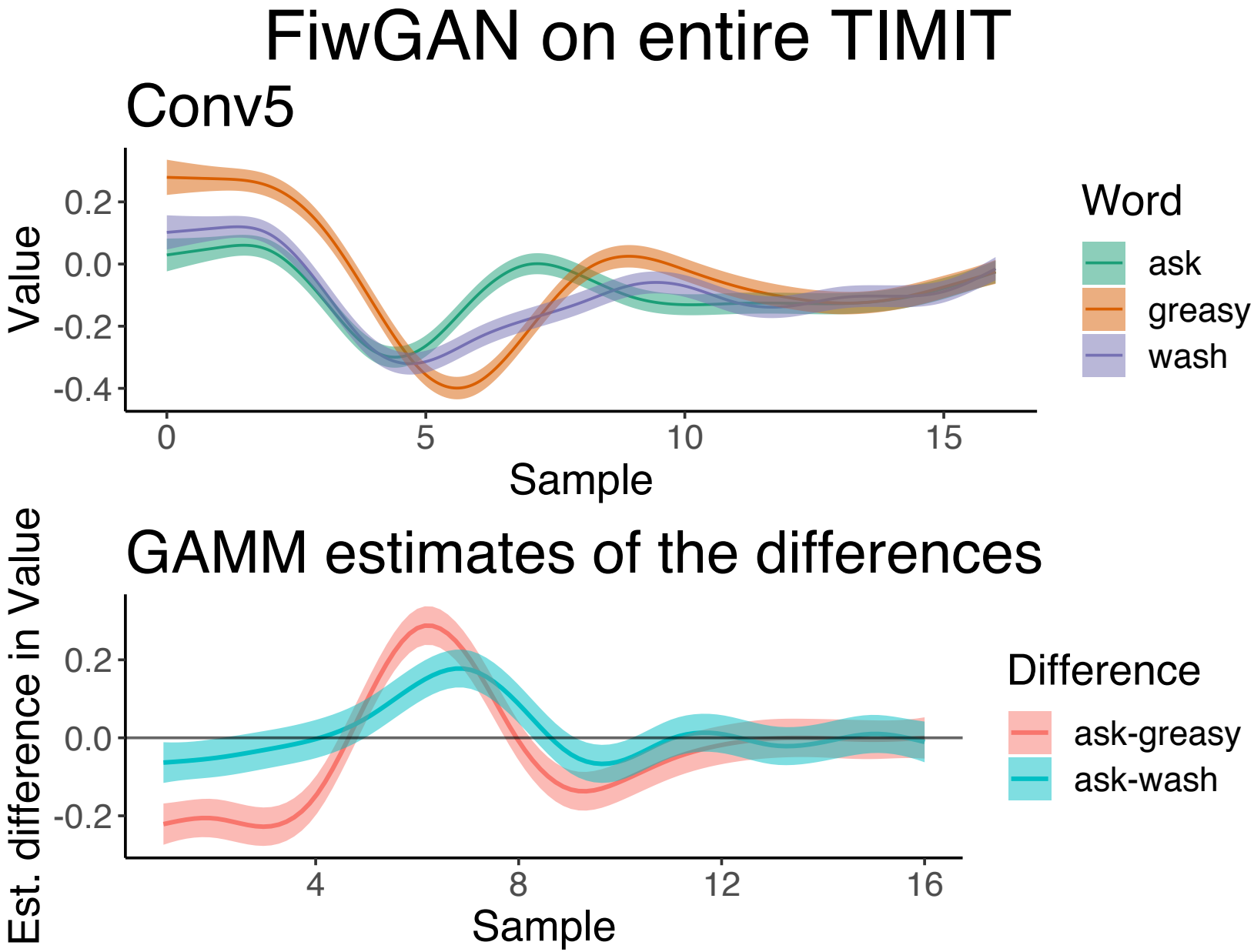
Table: Regression estimates of the GAMM model.

Generalized Additive Mixed Models

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept) = **ask**	0.6832	0.0215	31.7138	< 0.0001
Word=dark	-0.0409	0.0300	-1.3626	0.1730
Word=greasy	0.6323	0.0314	20.1116	$<$ **0.0001**
Word=oily	0.0811	0.0312	2.6021	**0.0093**
Word=rag	-0.0240	0.0305	-0.7870	0.4313
Word=wash	0.0603	0.0311	1.9425	0.0521
Word=water	-0.0464	0.0307	-1.5126	0.1304
Word=year	0.1298	0.0311	4.1712	$<$ **0.0001**
B. smooth terms	edf	Ref.df	F-value	p-value
s(Sample)	8.9700	8.9934	235.5282	< 0.0001
s(Sample):Word=dark	8.9684	8.9936	159.7072	< 0.0001
s(Sample):Word=greasy	8.9446	8.9886	231.4914	< 0.0001
s(Sample):Word=oily	8.9562	8.9911	122.4444	< 0.0001
s(Sample):Word=rag	8.9534	8.9904	110.0716	< 0.0001
s(Sample):Word=wash	8.9625	8.9924	165.8256	< 0.0001
s(Sample):Word=water	8.9309	8.9855	71.5510	< 0.0001
s(Sample):Word=year	8.9613	8.9922	140.2448	< 0.0001
s(Sample,UniqueWord)	5327.3250	9595.0000	4.8275	< 0.0001

Table: Regression estimates of the GAMM model.

Entire TIMIT



Conclusions

- **A technique to interpret and visualize intermediate convolutional layers** when networks learn to classify words from unlabeled data in an unsupervised manner without ever having access to the actual training data
- **Introducing inferential statistics** — generalized additive mixed models — to infer underlying distributions of word representations.
- This allows inferential statistical tests of both **absolute values** and **shapes** of word representations at each convolutional layer.

Future directions

- Any acoustic contrast can be tested and compared
- The technique has the potential to serve as a diagnostic for detecting layers at which speech contrasts (such as phonemes) fail to get encoded

References I

Beguš, G., 2021. CiwGAN and fiwGAN: Encoding information in acoustic data to model lexical learning with Generative Adversarial Networks. Neural Networks 139, 305–325. URL https://www.sciencedirect.com/science/article/pii/S0893608021001052

Beguš, G., Zhou, A., 2021. Interpreting intermediate convolutional layers of cnns trained on raw speech. CoRR abs/2104.09489.
URL https://arxiv.org/abs/2104.09489

Donahue, C., McAuley, J. J., Puckette, M. S., 2019. Adversarial audio synthesis. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.
URL https://openreview.net/forum?id=ByMVTsR5KQ

Garofolo, J. S., Lamel, L., M Fisher, W., Fiscus, J., S. Pallett, D., L. Dahlgren, N., Zue, V., 11 1993. Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680.
URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

References II

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Thank you!

: {begus,azhou314}@berkeley.edu
: @BerkeleySClab