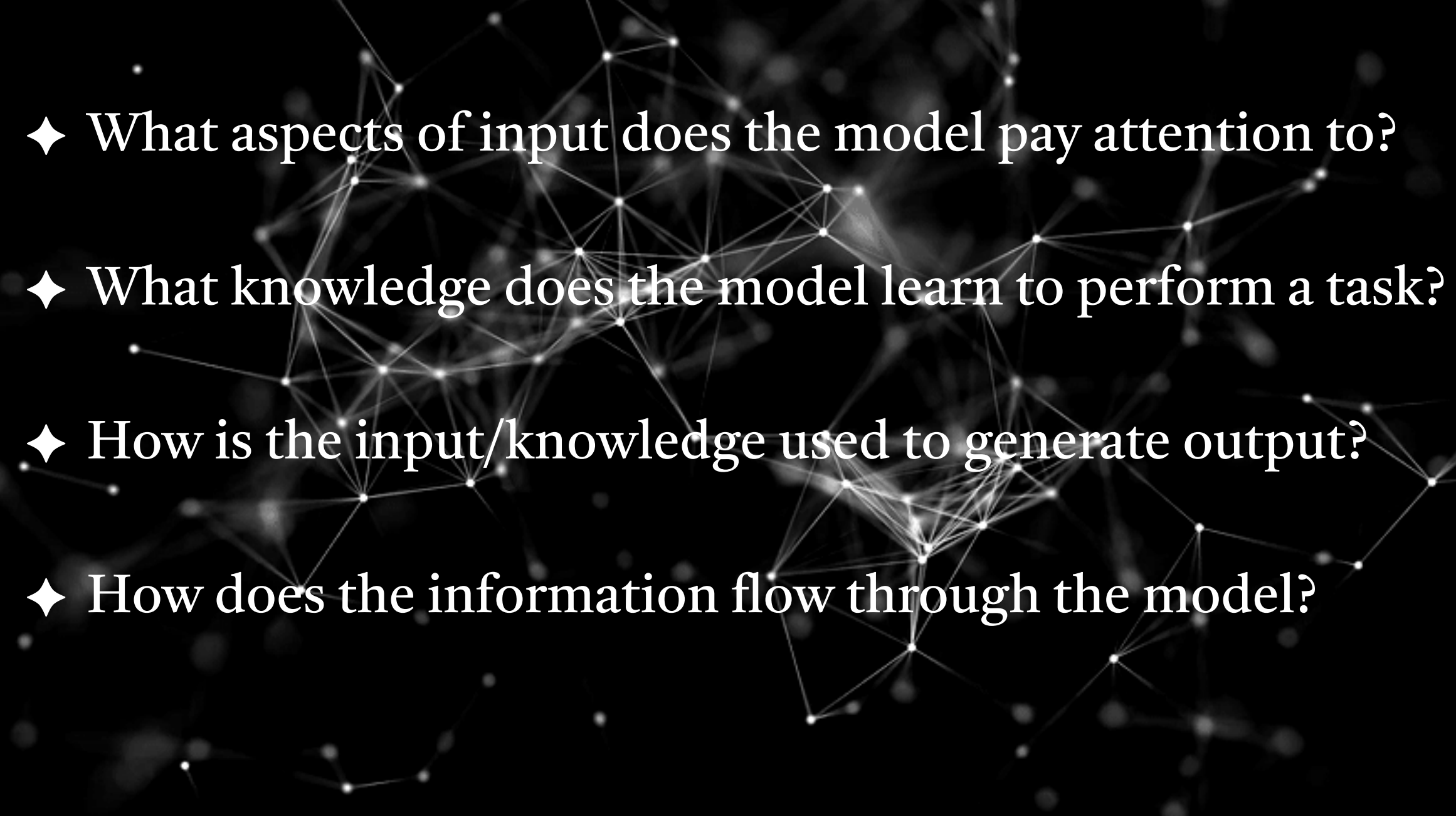


# Analysing and Interpreting Deep Neural Models of Language

Afra Alishahi

Cognitive Science & AI, Tilburg University

# Large language models are powerful but hard to understand.

- 
- ◆ What aspects of input does the model pay attention to?
  - ◆ What knowledge does the model learn to perform a task?
  - ◆ How is the input/knowledge used to generate output?
  - ◆ How does the information flow through the model?

# What's the Plan?

- A bit of History
- A bird's-eye view
- (Selected) interpretability techniques
- Evaluation of interpretations

# A Bit of History

COGNITIVE SCIENCE **14**, 179-211 (1990)

## Finding Structure in Time

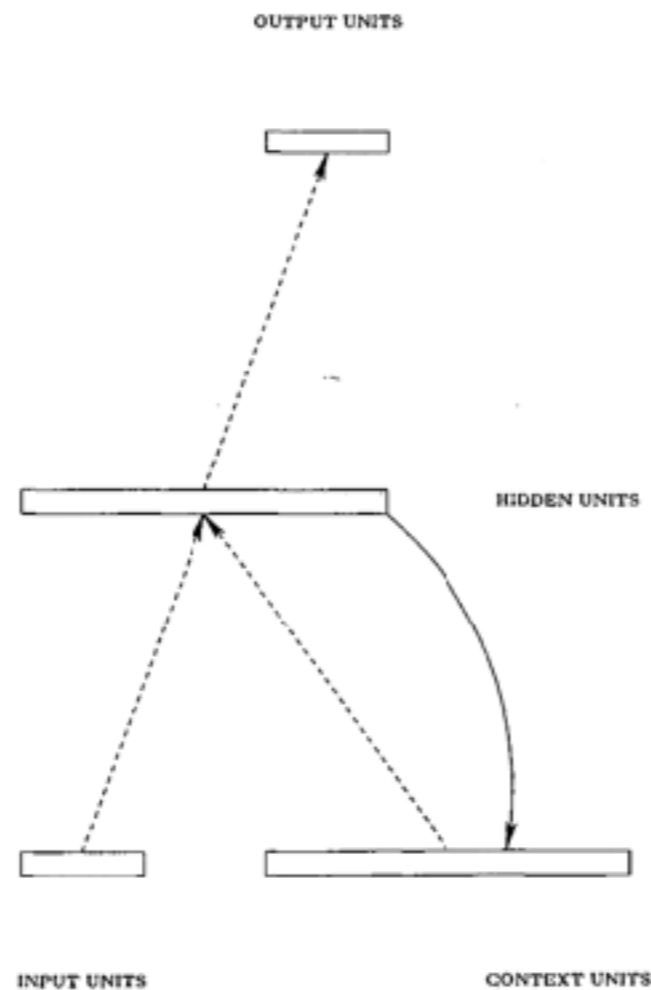
JEFFREY L. ELMAN

*University of California, San Diego*

Time underlies many interesting human behaviors. Thus, the question of how to represent time in connectionist models is very important. One approach is to represent time implicitly by its effects on processing rather than explicitly (as in a spatial representation). The current report develops a proposal along these lines first described by Jordan (1986) which involves the use of recurrent links in order to provide networks with a dynamic memory. In this approach, hidden unit patterns are fed back to themselves; the internal representations which develop

# Elman (1990): Simple RNN's

- He applied the model to a “language modelling” task, among others



[https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402\\_1](https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1)

Categories of Lexical Items Used in Sentence Simulation

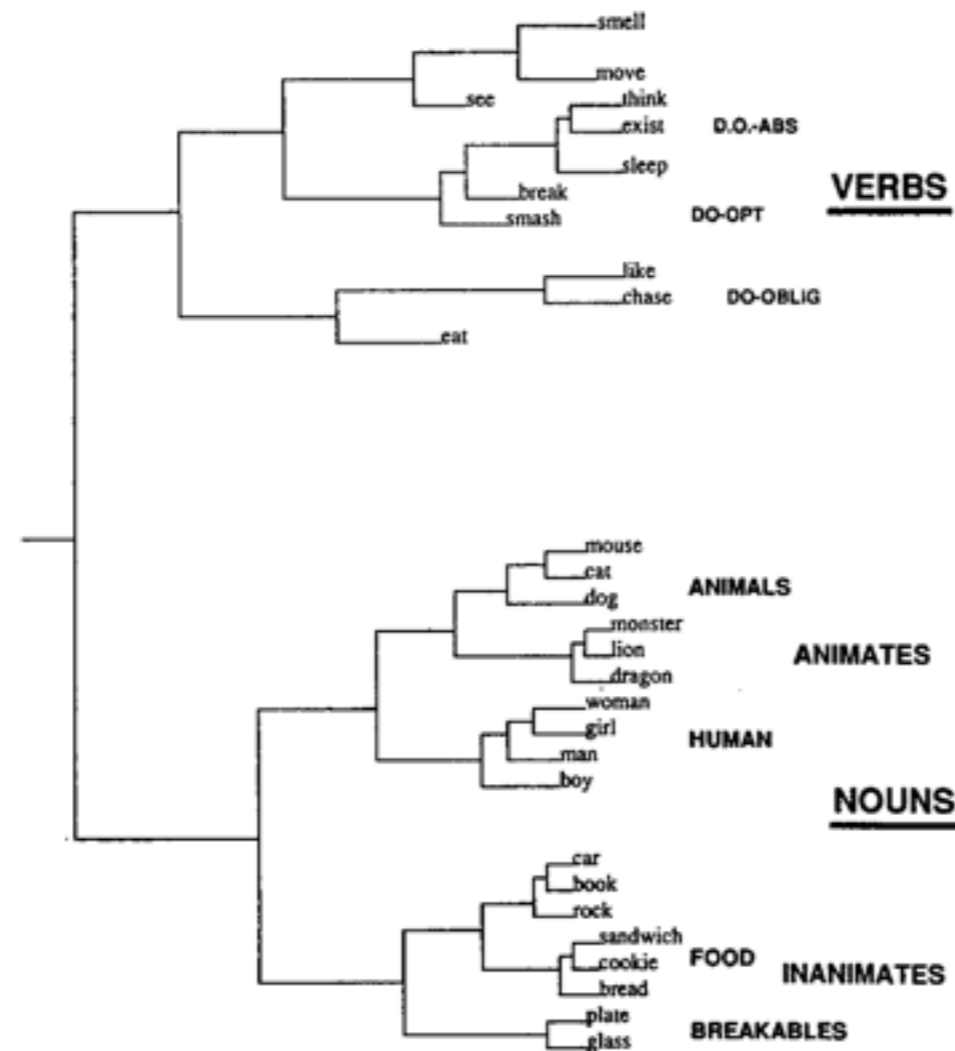
Category	Examples
NOUN-HUM	man, woman
NOUN-ANIM	cat, mouse
NOUN-INANIM	book, rock
NOUN-AGRESS	dragon, monster
NOUN-FRAG	glass, plate
NOUN-FOOD	cookie, break
VERB-INTRAN	think, sleep
VERB-TRAN	see, chase
VERB-AGPAT	move, break
VERB-PERCEPT	smell, see
VERB-DESTROY	break, smash
VERB-EAT	eat

TABLE 4  
Templates for Sentence Generator

WORD 1	WORD 2	WORD 3
NOUN-HUM	VERB-EAT	NOUN-FOOD
NOUN-HUM	VERB-PERCEPT	NOUN-INANIM
NOUN-HUM	VERB-DESTROY	NOUN-FRAG
NOUN-HUM	VERB-INTRAN	
NOUN-HUM	VERB-TRAN	NOUN-HUM
NOUN-HUM	VERB-AGPAT	NOUN-INANIM
NOUN-HUM	VERB-AGPAT	
NOUN-ANIM	VERB-EAT	NOUN-FOOD
NOUN-ANIM	VERB-TRAN	NOUN-ANIM
NOUN-ANIM	VERB-AGPAT	NOUN-INANIM
NOUN-ANIM	VERB-AGPAT	
NOUN-INANIM	VERB-AGPAT	
NOUN-AGRESS	VERB-DESTROY	NOUN-FRAG
NOUN-AGRESS	VERB-EAT	NOUN-HUM
NOUN-AGRESS	VERB-EAT	NOUN-ANIM
NOUN-AGRESS	VERB-EAT	NOUN-FOOD

# What Does the Model Learn?

- Apply hierarchical clustering on extracted word embeddings



[https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402\\_1](https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1)



# Analysing Deep Models of Vision

- 2012-2014: looking for “feature detectors” in early deep models of object classification (e.g. AlexNet, GoogLeNet )

# Analysing Deep Models of Vision

## BUILDING HIGH-LEVEL FEATURES USING LARGE SCALE UNSUPERVISED LEARNING

*Quoc V. Le*

Google Inc., USA

### ABSTRACT

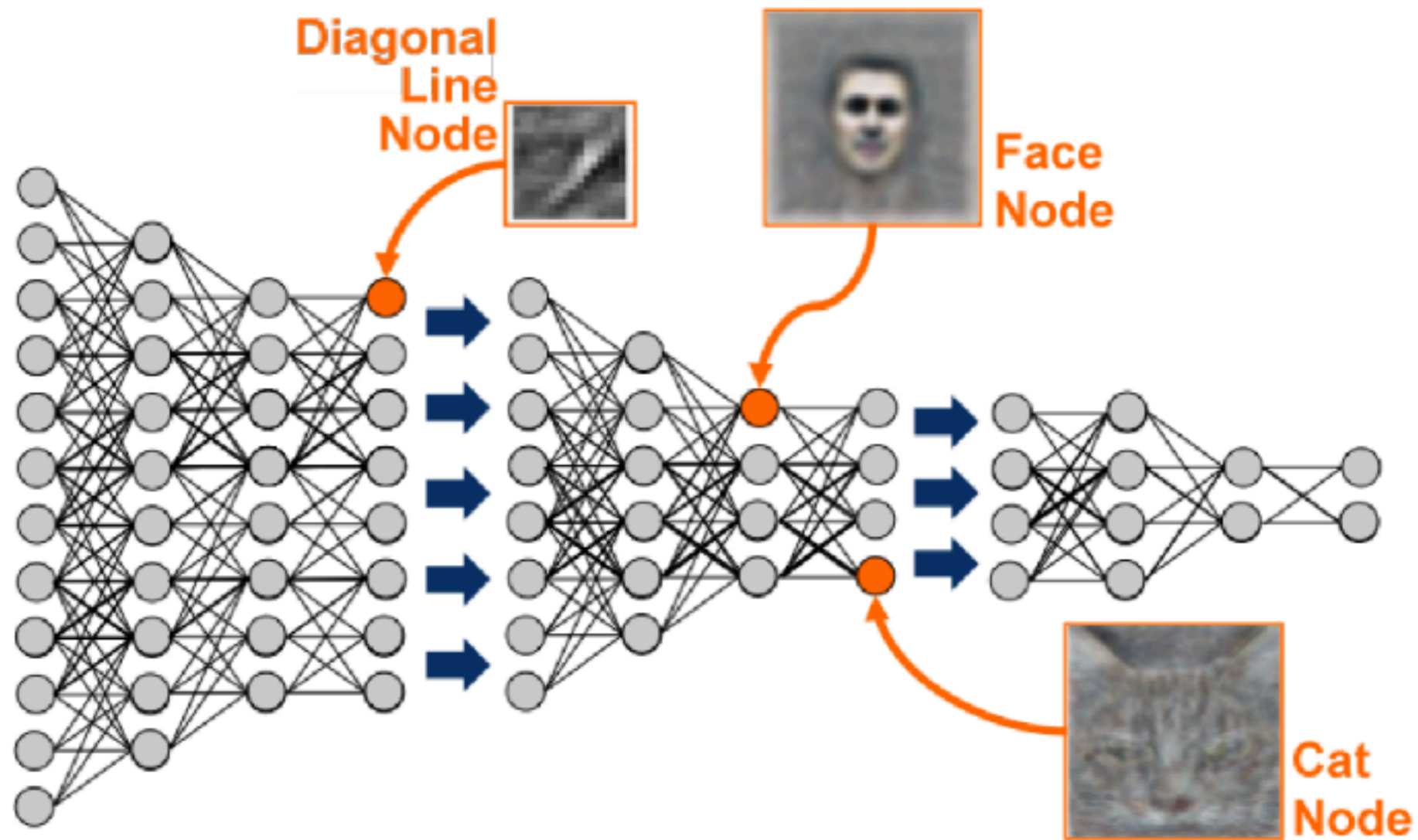
We consider the problem of building high-level, class-specific feature detectors from only unlabeled data. For example, is it possible to learn a face detector using only unlabeled images? To answer this, we train a deep sparse autoencoder on a large dataset of images (the model has 1 billion connections, the dataset has 10 million 200x200 pixel images downloaded from the Internet). We train this network using model parallelism and asynchronous SGD on a cluster with 1,000 machines (16,000 cores) for three days. Contrary to what appears to be a widely-held intuition, our experimental results reveal that it is possible to train a face detector without having to label images as containing a face or not. Control ex-

periments that make use of inexpensive unlabeled data are often preferred, they have not been shown to work well for building high-level features.

This work investigates the feasibility of building high-level features from only *unlabeled* data. A positive answer to this question will give rise to two significant results. Practically, this provides an inexpensive way to develop features from unlabeled data. But perhaps more importantly, it answers an intriguing question as to whether the specificity of the “grandmother neuron” could possibly be learned from unlabeled data. Informally, this would suggest that it is at least in principle possible that a baby learns to group faces into one class because it has seen many of them and not because it is guided by supervision or rewards.

[https://ieeexplore.ieee.org/abstract/document/6639343?casa\\_token=CYuG5JIZBGMAAAAA:Sh4MQcFPbyIveM3Z4kN\\_UIBrULYG-senEtSCHX5CpUTCsrBEOBcXecQeiBnTARDBBHBwHQEQ0XyqmQ](https://ieeexplore.ieee.org/abstract/document/6639343?casa_token=CYuG5JIZBGMAAAAA:Sh4MQcFPbyIveM3Z4kN_UIBrULYG-senEtSCHX5CpUTCsrBEOBcXecQeiBnTARDBBHBwHQEQ0XyqmQ)

# Analysing Deep Models of Vision



[https://ieeexplore.ieee.org/abstract/document/6639343?casa\\_token=CYuG5JIZBGMAAAAA:Sh4MQcFPbyIveM3Z4kN\\_UIBrULYG-senEtSCHX5CpUTCsrBEOBcXecQeiBnTARDBBHBwHQQEQ0XyqmQ](https://ieeexplore.ieee.org/abstract/document/6639343?casa_token=CYuG5JIZBGMAAAAA:Sh4MQcFPbyIveM3Z4kN_UIBrULYG-senEtSCHX5CpUTCsrBEOBcXecQeiBnTARDBBHBwHQQEQ0XyqmQ)

# Early Interpretability Attempts in NLP

## Diagnostic classifiers: revealing how neural networks process hierarchical structure

Sara Veldhoen, Dieuwke Hupkes and Willem Zuidema  
ILLC, University of Amsterdam  
P.O.Box 94242, 1090 CE Amsterdam, Netherlands  
{s.f.veldhoen, d.hupkes, zuidema}@uva.nl

### Abstract

We investigate how neural networks can be used for hierarchical, compositional semantics. To this end, we define the simple but nontrivial artificial task of processing nested arithmetic expressions and study whether different types of neural networks can learn to add and subtract. We find that recursive neural networks can implement a generalising solution, and we visualise the intermediate steps: projection, summation and squashing. We also show that gated recurrent neural networks, which process the expressions incrementally, perform surprisingly well

## Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

Tal Linzen<sup>1,2</sup> Emmanuel Dupoux<sup>1</sup>  
LSCP<sup>1</sup> & IJN<sup>2</sup>, CNRS,  
EHESS and ENS, PSL Research University  
{tal.linzen,  
emmanuel.dupoux}@ens.fr

Yoav Goldberg  
Computer Science Department  
Bar Ilan University  
yoav.goldberg@gmail.com

### Abstract

The success of long short-term memory (LSTM) neural networks in language processing is typically attributed to their ability to capture long-distance statistical regularities. Linguistic regularities are often sensitive to syntactic structure; can such dependencies be captured by LSTMs, which do not have explicit structural representations? We begin ad-

(Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Cho et al., 2014), has led to significant gains in language modeling (Mikolov et al., 2010; Sundermeyer et al., 2012), parsing (Vinyals et al., 2015; Kiprwasser and Goldberg, 2016; Dyer et al., 2016), machine translation (Bahdanau et al., 2015) and other tasks.

The effectiveness of RNNs<sup>1</sup> is attributed to their ability to capture statistical contingencies that may

Published as a conference paper at ICLR 2017

## FINE-GRAINED ANALYSIS OF SENTENCE EMBEDDINGS USING AUXILIARY PREDICTION TASKS

Yossi Adi<sup>1,2</sup>, Einat Kerner<sup>2</sup>, Yonatan Belinkov<sup>3</sup>, Ofer Lavi<sup>2</sup>, Yoav Goldberg<sup>1</sup>

<sup>1</sup>Bar-Ilan University, Ramat-Gan, Israel  
{yoav.goldberg, yossiadicdrum}@gmail.com

<sup>2</sup>IBM Haifa Research Lab, Haifa, Israel  
{einatke, oferl}@il.ibm.com

<sup>3</sup>MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA  
belinkov@mit.edu

## Representation of Linguistic Form and Function in Recurrent Neural Networks

Ákos Kádár\*  
Tilburg University

Grzegorz Chrupala\*  
Tilburg University

Afra Alishahi\*  
Tilburg University

*We present novel methods for analyzing the activation patterns of recurrent neural networks from a linguistic point of view and explore the types of linguistic structure they learn. As a case study, we use a standard standalone language model, and a multi-task gated recurrent network architecture consisting of two parallel pathways with shared word embeddings: The VISUAL pathway is trained on predicting the representations of the visual scene corresponding to an input*

# First BlackboxNLP @EMNLP'2018

## Analyzing and interpreting neural networks for NLP

Revealing the content of the neural black box: workshop on the analysis and interpretation of neural networks for Natural Language Processing.

 [View On GitHub](#)

This project is maintained by [blackboxnlp](#)

Hosted on [GitHub Pages](#) using the [Dinky theme](#)

## Venue

The workshop will be collocated with [EMNLP 2018](#) in Brussels.

## Important dates

- July 19. Submission deadline (11:59pm Pacific Daylight Savings Time, UTC-7h).
- August 3. Notification of acceptance.
- August 30. Camera ready (11:59pm Pacific Daylight Savings Time, UTC-7h).
- November 1. Workshop.

## Proceedings

The workshop proceedings are available via ACL Anthology: [proceedings](#)

## Workshop program

Time	Program item
09:00-09:10	Opening remarks
09:10-10:00	Invited talk 1: <a href="#">Yoav Goldberg</a>
10:00-11:00	Poster session 1 (10:30-11 tea break)
11:00-12:30	Oral presentation session 1 (6 x 15 minutes)
12:30-14:00	Lunch
14:00-14:50	Invited talk 2: <a href="#">Graham Neubig</a>
14:50-16:00	Poster session 2 (15:30-16 tea break)
16:00-16:50	Invited talk 3: <a href="#">Lila Wernke</a>
16:50-17:20	Oral presentation session 2 (2 x 15 minutes)
17:20-17:30	Best paper announcement and closing remarks

ARTICLE

## Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop

Afra Alishahi<sup>1,\*</sup>, Grzegorz Chrupała<sup>1</sup> and Tal Linzen<sup>2</sup>

<sup>1</sup>Department of Cognitive Science and Artificial Intelligence, Tilburg University, The Netherlands and <sup>2</sup>Department of Cognitive Science, Johns Hopkins University, Baltimore, United States

\*Corresponding author. Email: [A.Alishahi@uvt.nl](mailto:A.Alishahi@uvt.nl)

### Abstract

The Empirical Methods in Natural Language Processing (EMNLP) 2018 workshop BlackboxNLP was dedicated to resources and techniques specifically developed for analyzing and understanding the inner-workings and representations acquired by neural models of language. Approaches included: systematic manipulation of input to neural networks and investigating the impact on their performance, testing whether interpretable knowledge can be decoded from intermediate representations acquired by neural networks, proposing modifications to neural network architectures to make their knowledge state or generated output more explainable, and examining the performance of networks on simplified or formal languages. Here we review a number of representative studies in each category.

**Keywords:** neural networks; interpretability; natural language processing

ARTICLE

Analyzing and interpreting neural networks for NLP:  
A report on the first BlackboxNLP workshop

Afra Alishahi<sup>1,\*</sup>, Grzegorz Stasiński<sup>1</sup>, Grzegorz Stasiński<sup>1</sup> and Tal Linzen<sup>2</sup>

<sup>1</sup>Department of Cognitive Science and Artificial Intelligence, Tilburg University, The Netherlands and <sup>2</sup>Department of Cognitive Science, Johns Hopkins University, Baltimore, United States

\*Corresponding author. Email: [afra@uvt.nl](mailto:afra@uvt.nl)

Abstract

The Empirical Methods in Natural Language Processing (EMNLP) 2019 workshop was dedicated to resources and techniques for analyzing and interpreting neural networks and representations acquired by neural networks. This workshop focused on the manipulation of input to neural networks and investigating the impact on their output, whether interpretable knowledge can be decoded from intermediate representations of neural networks, proposing modifications to neural network architectures to make their knowledge generated output more explainable, and examining the performance of networks on simple languages. Here we review a number of representative studies in each category.

**Keywords:** neural networks; interpretability; natural language processing

*Much has  
happened since!*

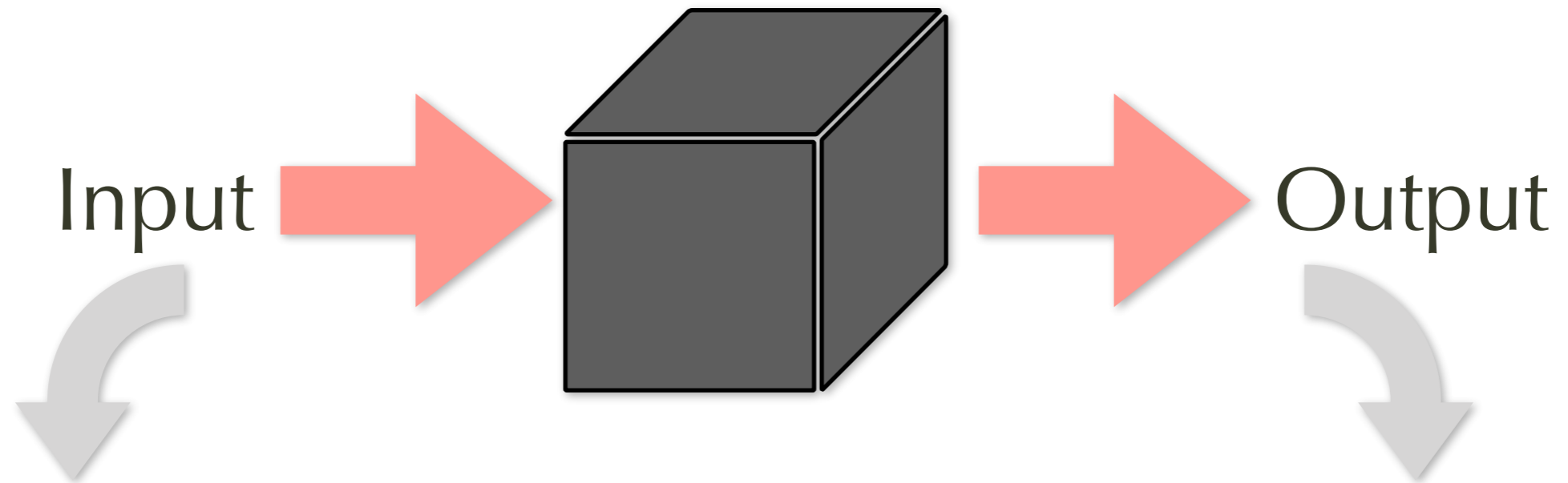
# A Bird's-eye View



# How to Analyse a Model?

- Blackbox versus whitebox approaches
- Hypothesis-driven versus data-driven approaches
- Modality-specific approaches

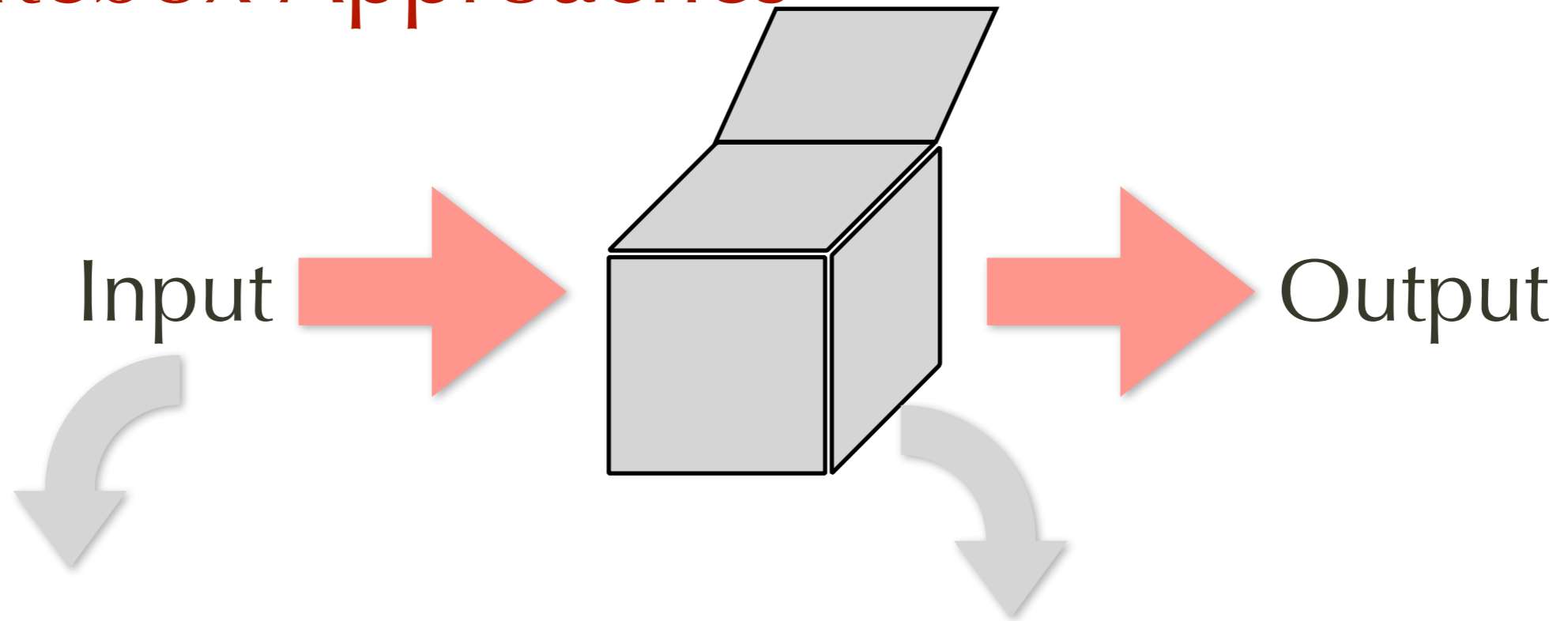
# Blackbox Approaches



Systematically manipulate input, observe how it affects output

- Behavioural tests
- Input manipulation/perturbation
- Adversarial attacks/investigations

# Whitebox Approaches



Systematically manipulate input, look inside the model

- Hidden-layer representations, embeddings, ...
- Feature attribution, context mixing, ...
- Information flow, gradients, ...

# How to Analyse a Model?

- Blackbox versus whitebox approaches

- Hypothesis-driven versus data-driven approaches

- Modality-specific approaches

# Hypothesis-driven: What do Models Learn?

Hypothesise that knowledge  $X$  is needed for performing task  $Y$



Look for evidence that a model trained on  $Y$  has encoded  $X$

- Probing/diagnostic classifiers or regressors
- Correlation analyses (e.g. CCA)
- Information-theoretic measures (e.g. Mutual Information)
- Representational Similarity Analysis (RSA)

# Data-driven: How do Models Work?

Feed the model with controlled input



Analyse information processing that leads to the output

- Input perturbation
- Feature attribution and context mixing
- Relevance propagation analysis
- Inducing explainable architectures
- Mechanistic interpretability

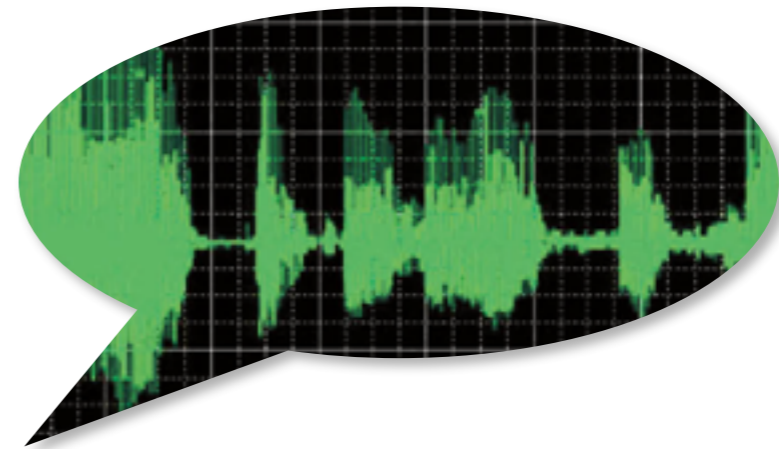
# How to Analyse a Model?

- Blackbox versus whitebox approaches
- Hypothesis-driven versus data-driven approaches
- Modality-specific approaches

# Modality-specific Approaches

Birds are flying  
over the pond

Text



Speech

Multimodal

Birds are flying  
over the pond



Multilingual

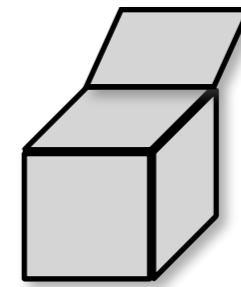
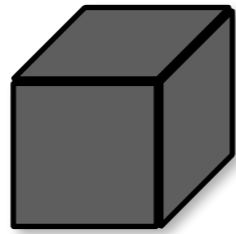
Birds are flying  
over the pond

پرندگان روی برکه  
پرواز می کنند



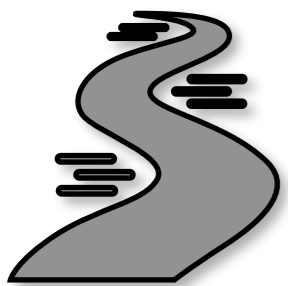
# Interpretability Techniques

- Probing/diagnostic classifiers or regressors
- Correlation analyses (e.g. CCA)
- Information-theoretic measures (e.g. Mutual Information)
- Representational Similarity Analysis (RSA)
- Input perturbation
- Feature attribution and context mixing
- Relevance propagation analysis
- Inducing explainable architectures
- Mechanistic analysis
- ...



- Behavioural Tests
- Input Perturbation

- Probing Classifiers
- Correlation Analyses (e.g. CCA)
- Mutual Information
- Representational Similarity Analysis (RSA)
- Feature Attribution



- Input Perturbation
- Inducing Explainable Architectures

- Feature Attribution
- Relevance Propagation Analysis
- Mechanistic Analysis

# Analysis Methods in Neural Language Processing: A Survey

Yonatan Belinkov<sup>1,2</sup> and James Glass<sup>1</sup>

<sup>1</sup>MIT Computer Science and Artificial Intelligence Laboratory

<sup>2</sup>Harvard School of Engineering and Applied Sciences

Cambridge, MA, USA

{belinkov, glass}@mit.edu

## Abstract

The field of natural language processing has seen impressive progress in recent years, with neural network models replacing many of the traditional systems. A plethora of new mod-

the networks in different ways.<sup>1</sup> Others strive to better understand how NLP models work. This theme of analyzing neural networks has connections to the broader work on interpretability in machine learning, along with specific characteristics of the NLP field.

# Selected Interpretability Techniques

# Interpretability Techniques

1. Input perturbation
2. Probing/diagnostic classifiers or regressors
3. Representational Similarity Analysis (RSA)
4. Attribution and context mixing scores
5. Neuron-level interpretation
6. Inducing explainable architectures

# Interpretability Techniques

1. Input perturbation
2. Probing/diagnostic classifiers or regressors
3. Representational Similarity Analysis (RSA)
4. Attribution and context mixing scores
5. Neuron-level interpretation
6. Inducing explainable architectures

# Input Perturbation

- Systematically manipulate aspects of input and monitor model output
  - Example: manipulating “distractors” in verb-subject agreement test cases:

1. The **roses** in the vase by the door **are** red.
2. The **roses** in the vase by the chairs **are** red.

- Example: inserting “not” in NLI test cases:

1. The robin is a bird.
2. The robin is **not** a bird.

# Input Perturbation: Apply a "Behavioural Test" to an LLM

- Use a small-scale, controlled test set
  - ... often borrowed from a published psycholinguistic study
- Subject an LLM to these tests
- Make conclusions about
  - the nature of linguistic knowledge learned by the LLM
  - the processing mechanisms employed by the LLM
- *Compare their performance to that of human subjects*



# Targeted Syntactic Evaluation of Language Models

**Rebecca Marvin**

Department of Computer Science  
Johns Hopkins University  
becky@jhu.edu

**Tal Linzen**

Department of Cognitive Science  
Johns Hopkins University  
tal.linzen@jhu.edu

## Abstract

We present a dataset for evaluating the grammaticality of the predictions of a language model. We automatically construct a large number of minimally different pairs of En-

glish sentences that differ only in syntactic cations, semantics, pragmatics, syntax, and so on. The quality of the **syntactic** predictions made by the LM is arguably particularly difficult to measure using perplexity: since most sentences are grammatically simple and most words can be pre-

## Targeted Syntactic Evaluation of Language Models

		RNN	Multitask	<i>n</i> -gram	Humans	# sents
<b>Rebecca</b> Department of Johns Hopkins University Computer Science and Artificial Intelligence Institute Baltimore, MD 21218 rebecka@cs.jhu.edu	<b>SUBJECT-VERB AGREEMENT:</b>					
	Simple	0.94	1.00	0.79	0.96	280
	In a sentential complement	0.99	0.93	0.79	0.93	3360
	Short VP coordination	0.90	0.90	0.51	0.94	1680
	Long VP coordination	0.61	0.81	0.50	0.82	800
	Across a prepositional phrase	0.57	0.69	0.50	0.85	44800
	Across a subject relative clause	0.56	0.74	0.50	0.88	22400
	Across an object relative clause	0.50	0.57	0.50	0.85	44800
	Across an object relative (no <i>that</i> )	0.52	0.52	0.50	0.82	44800
	In an object relative clause	0.84	0.89	0.50	0.78	44800
	In an object relative (no <i>that</i> )	0.71	0.81	0.50	0.79	44800
	<b>REFLEXIVE ANAPHORA:</b>					
	Simple	0.83	0.86	0.50	0.96	560
In a sentential complement	0.86	0.83	0.50	0.91	6720	
Across a relative clause	0.55	0.56	0.50	0.87	44800	
<b>NEGATIVE POLARITY ITEMS:</b>						
Simple	0.40	0.48	0.06	0.98	792	
Across a relative clause	0.41	0.73	0.60	0.81	31680	

We present a dataset for the evaluation of the automaticity of the prediction model. We automatically generate a number of minimally

Table 1: Overall accuracies for the LSTMs, *n*-gram model and humans on each test case.

<https://aclanthology.org/D18-1151/>

# Input Perturbation: Grounded Language Learning

## Representation of Linguistic Form and Function in Recurrent Neural Networks

Ákos Kádár\*  
Tilburg University

Grzegorz Chrupała\*  
Tilburg University

Afra Alishahi\*  
Tilburg University

*We present novel methods for analyzing the activation patterns of recurrent neural networks from a linguistic point of view and explore the types of linguistic structure they learn. As a case study, we use a standard standalone language model, and a multi-task gated recurrent network architecture consisting of two parallel pathways with shared word embeddings: The VISUAL pathway is trained on predicting the representations of the visual scene corresponding to an input*

<https://direct.mit.edu/coli/article/43/4/761/1583/Representation-of-Linguistic-Form-and-Function-in>



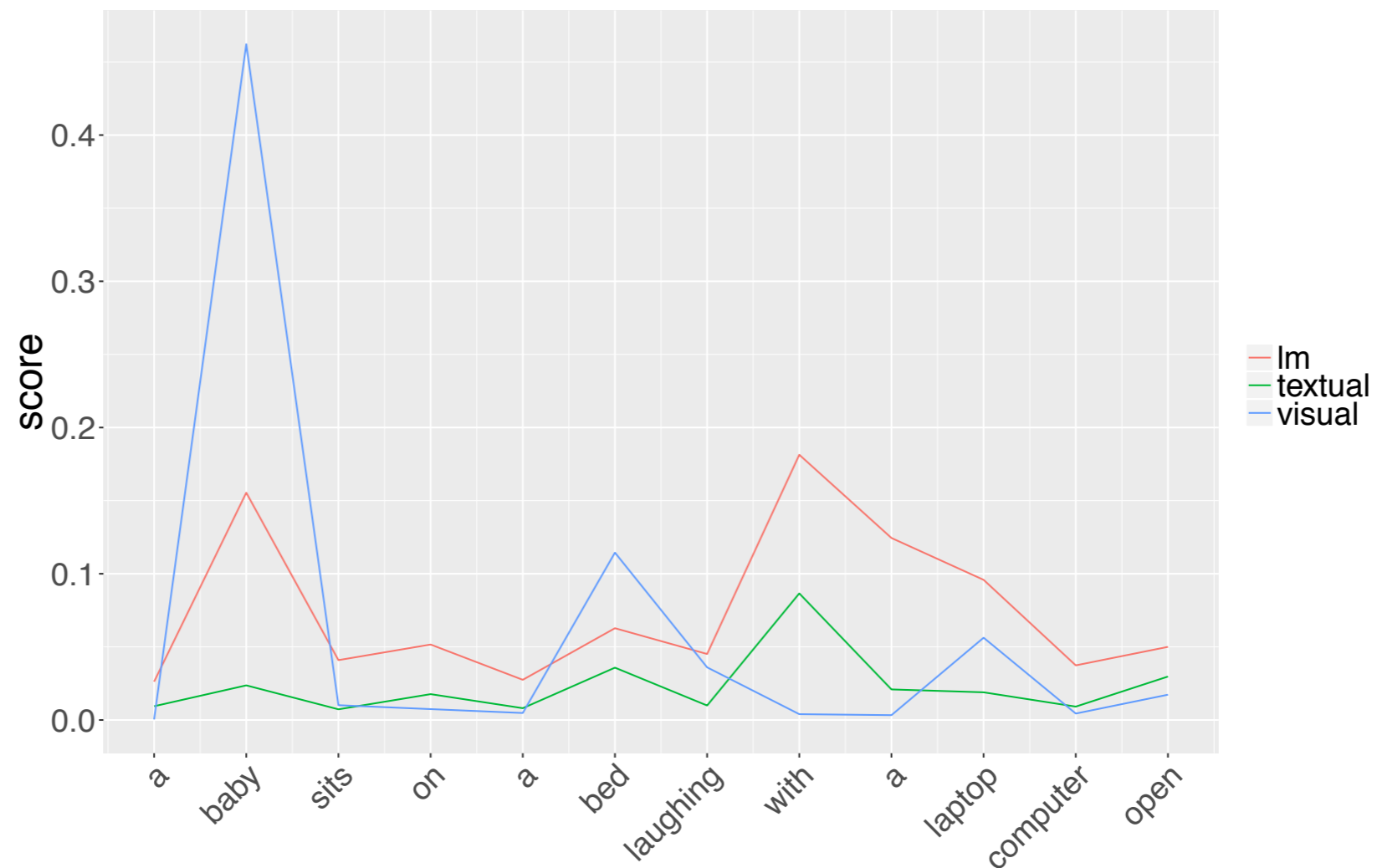
A baby sits on a bed laughing with a laptop computer open



A sits on a bed laughing with a laptop computer open

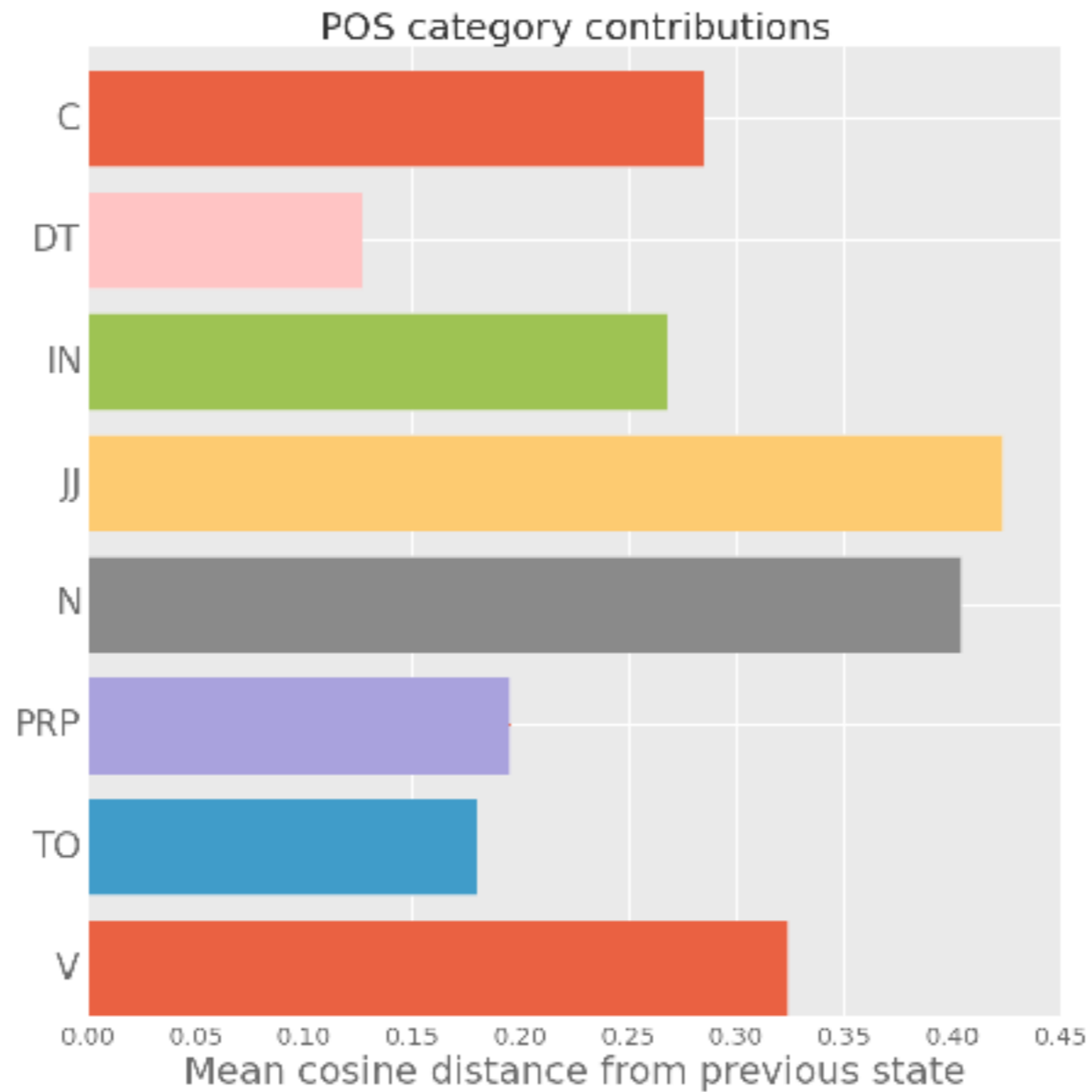
# Quantifying Contribution of Each Word

$$\text{omission}(i, S) = 1 - \text{cosine}(\mathbf{h}_{\text{end}}(S), \mathbf{h}_{\text{end}}(S_{\setminus i}))$$



<https://direct.mit.edu/coli/article/43/4/761/1583/Representation-of-Linguistic-Form-and-Function-in>

# Accumulating Omission Scores



<https://direct.mit.edu/coli/article/43/4/761/1583/Representation-of-Linguistic-Form-and-Function-in>

# Looking for Linguistic Structure

**Input sentence:**

*“a brown teddy bear lying on top of a dry grass covered ground .”*



- Does the model learn and use any knowledge about sentence structure?



# Looking for Linguistic Structure

## Input sentence:

*“a brown teddy bear lying on top of a dry grass covered ground .”*

## Scrambled input sentence:

*“a a of covered laying bear on brown grass top teddy ground . dry”*



# BLiMP: The Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt<sup>1</sup>, Alicia Parrish<sup>1</sup>, Haokun Liu<sup>2</sup>, Anhad Mohananey<sup>2</sup>,  
Wei Peng<sup>2</sup>, Sheng-Fu Wang<sup>1</sup>, Samuel R. Bowman<sup>1,2,3</sup>

<sup>1</sup>Department of Linguistics    <sup>2</sup>Department of Computer Science    <sup>3</sup>Center for Data Science  
New York University                      New York University                      New York University

{warstadt, alicia.v.parrish, haokunliu, anhad,  
weipeng, shengfu.wang, bowman}@nyu.edu

## Abstract

We introduce The Benchmark of Linguistic Minimal Pairs (BLiMP),<sup>1</sup> a challenge set for evaluating the linguistic knowledge of language models (LMs) on major grammatical phenomena in English. BLiMP consists of 67 individual datasets, each containing 1,000

of these studies uses a different set of metrics, and focuses on a small set of linguistic paradigms, severely limiting any possible big-picture conclusions.

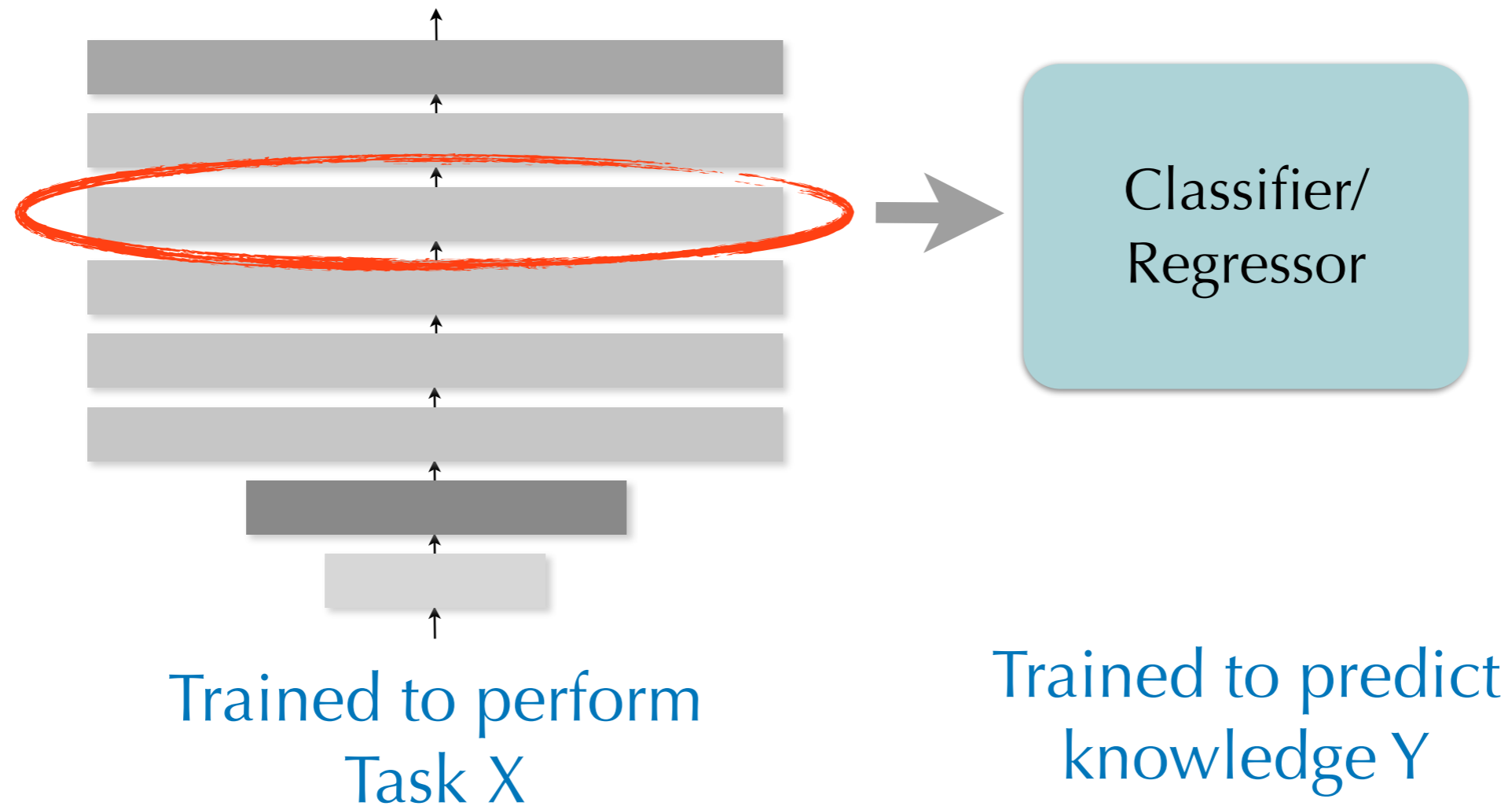
- (1) a. The cats annoy Tim. (*grammatical*)  
b. \*The cats annoys Tim. (*ungrammatical*)

# Interpretability Techniques

1. Input perturbation
2. Probing/diagnostic classifiers or regressors
3. Representational Similarity Analysis (RSA)
4. Attribution and context mixing scores
5. Neuron-level interpretation
6. Inducing explainable architectures

# Probing Classifiers/Regressors

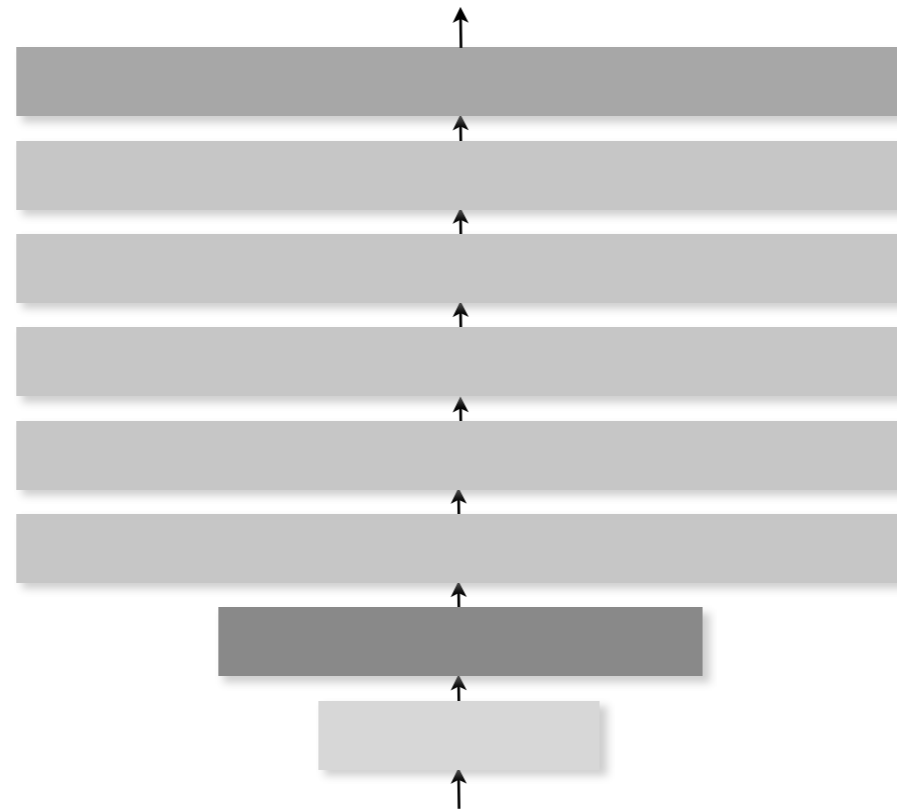
- Also diagnostic classifiers, auxiliary tasks, ...



# Sample Auxiliary Tasks

- Utterance length
- Presence of specific words
- Part of speech
- Dependency labels
- Similarity judgment
- Homonym/synonym detection
- Semantic roles
- ...

# Analyse Internal Representations



- What aspects of language does the model encode?
  - Does the model encode linguistic form, and where?
  - Does the model encode meaning, and where?

## Encoding of phonology in a recurrent neural model of grounded speech

**Afra Alishahi**  
Tilburg University  
a.alishahi@uvt.nl

**Marie Barking**  
Tilburg University  
m.barking@uvt.nl

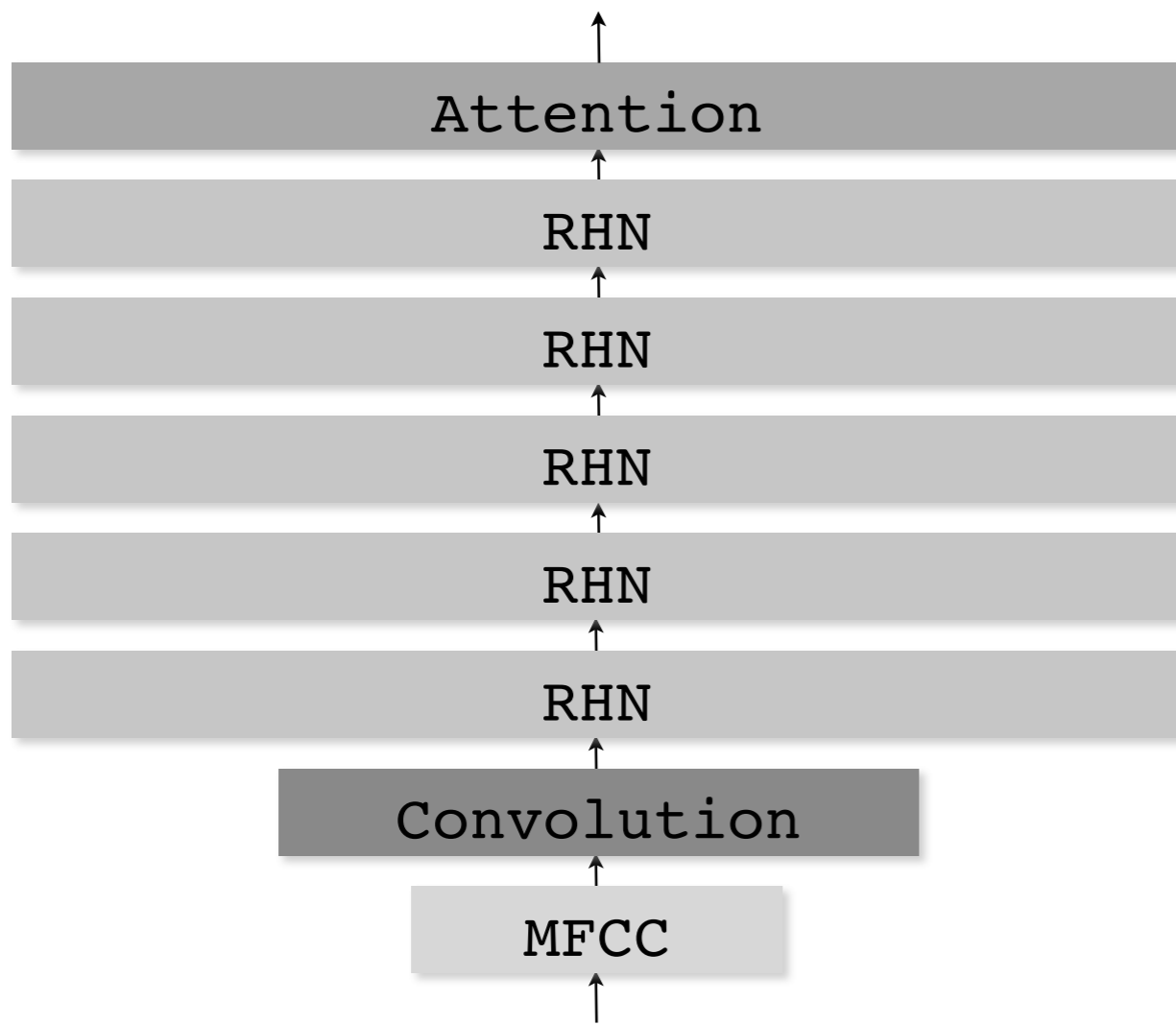
**Grzegorz Chrupała**  
Tilburg University  
g.chrupala@uvt.nl

### Abstract

We study the representation and encoding of phonemes in a recurrent neural network model of grounded speech. We use a model which processes images and their spoken descriptions, and projects the visual and auditory representations into

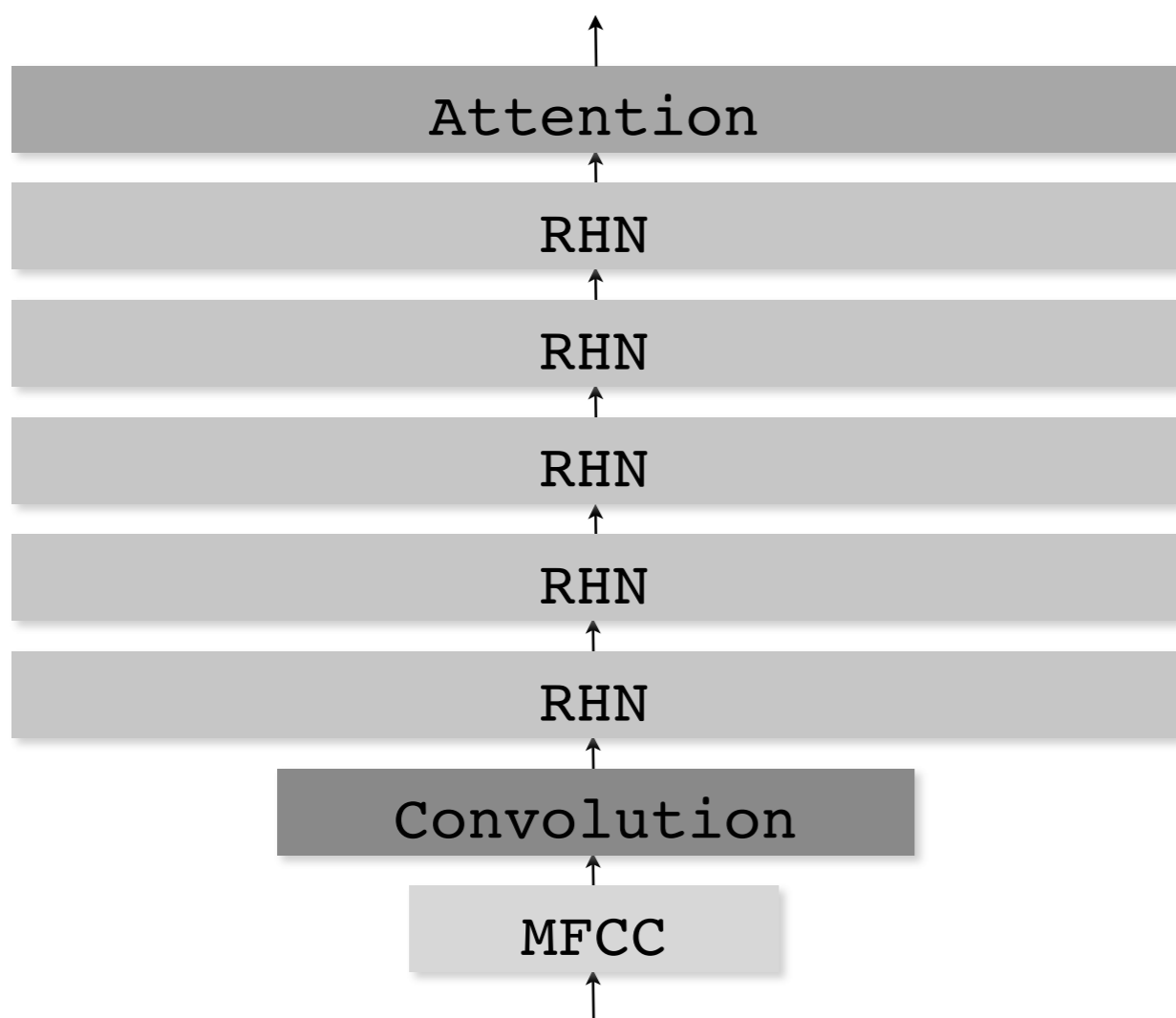
commonly via the analysis of neuro-imaging data of participants exposed to simplified, highly controlled inputs. More recently, naturalistic data has been used and patterns in the brain have been correlated with patterns in the input (e.g. [Wehbe et al., 2014](#); [Khalighinejad et al., 2017](#)).

This type of approach is relevant also when the goal is the understanding of the dynamics in com-



Utterance Length	Word Presence	Edit Similarity	Meaning Similarity	Homonym Detection	Synonym Detection
				*	
	*		*	*	
		*			*
*					
*					





Utterance Length  
 Word Presence  
 Edit Similarity  
 Meaning Similarity  
 Homonym Detection  
 Synonym Detection

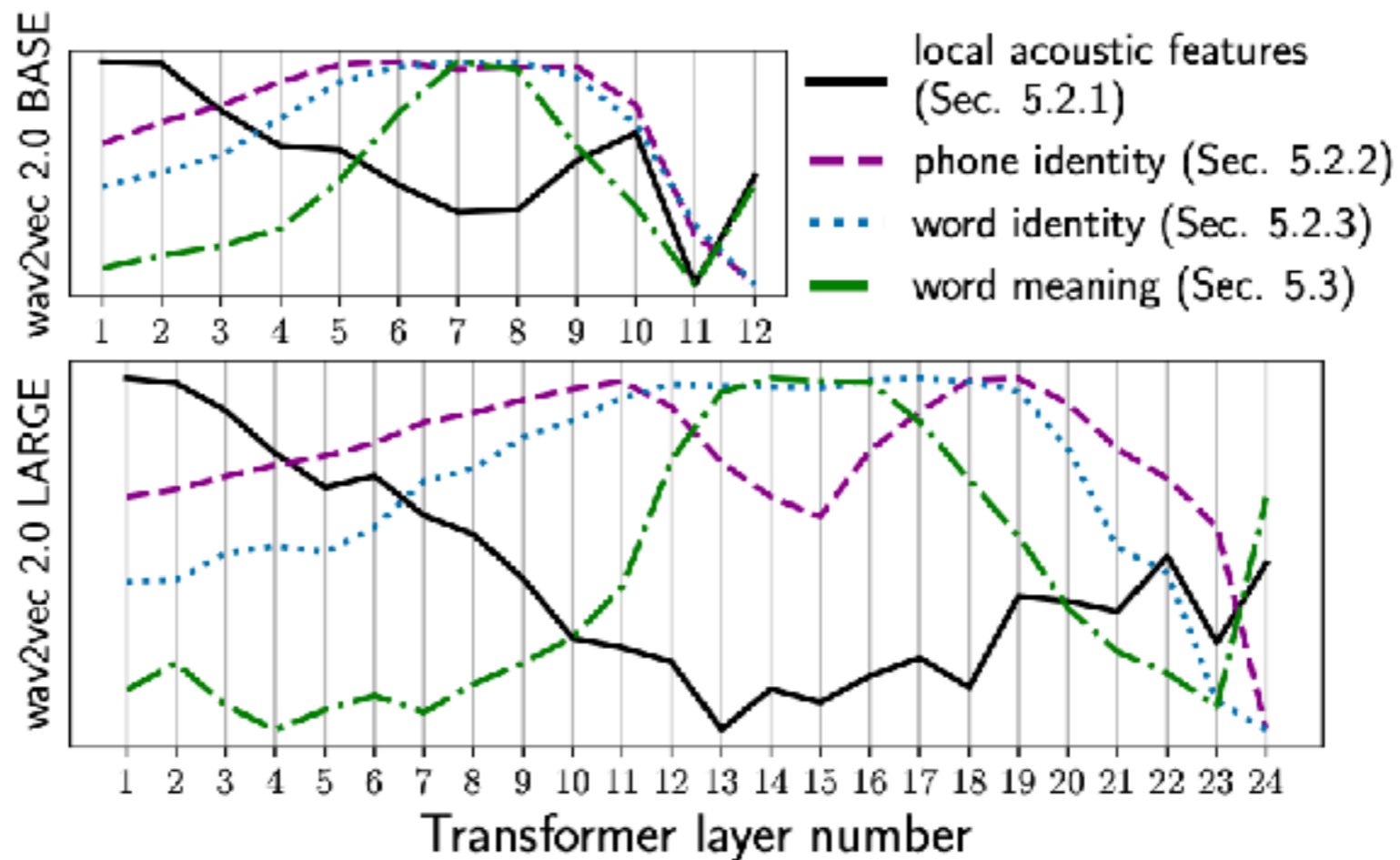
				*	
	*		*	*	
*		*			*
*					

# LAYER-WISE ANALYSIS OF A SELF-SUPERVISED SPEECH REPRESENTATION MODEL

Ankita Pasad, Ju-Chieh Chou, Karen Livescu

Toyota Technological Institute at Chicago

{ankitap, jcchou, klivescu}@ttic.edu



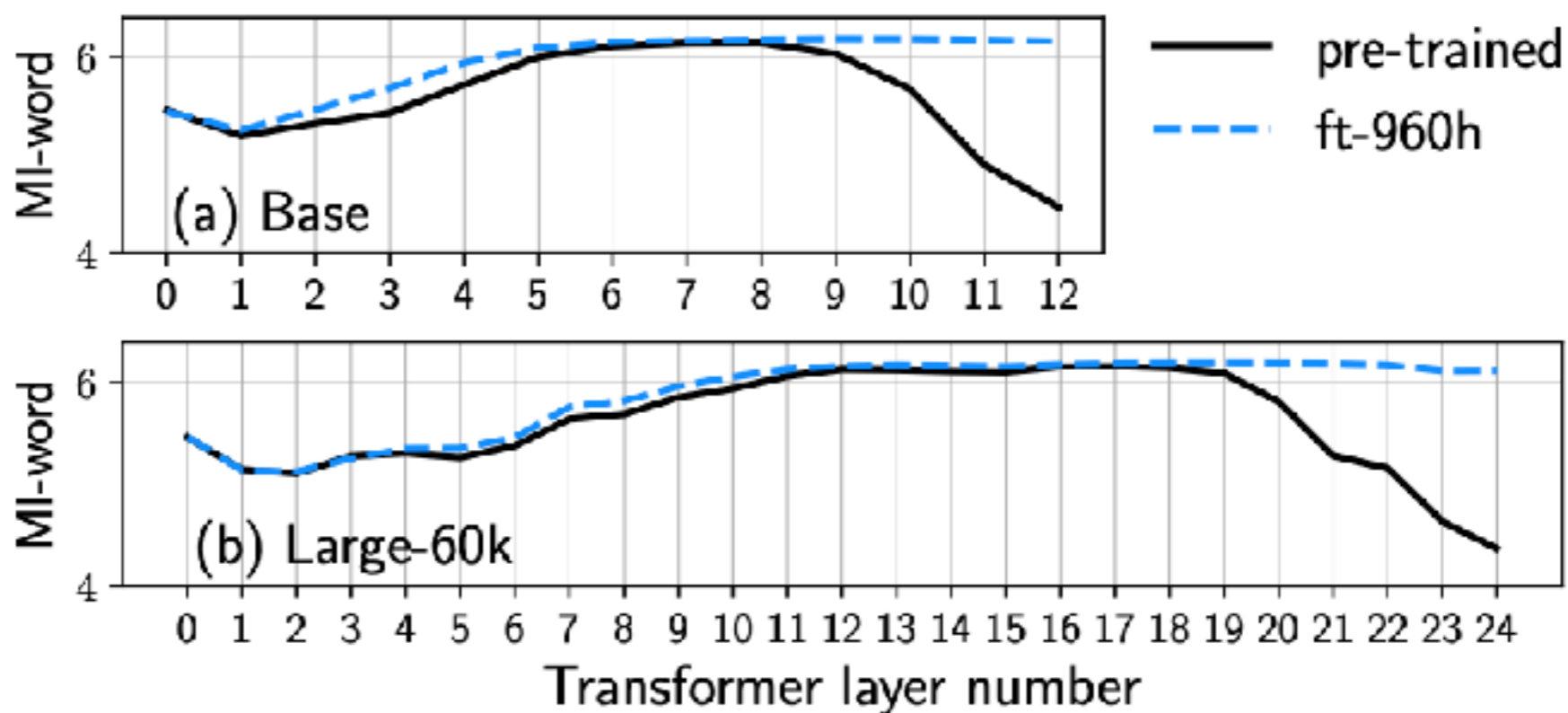
[https://ieeexplore.ieee.org/abstract/document/9688093?casa\\_token=9i9SkqxnXM0AAAAA:GSL5eW\\_VCmvYFjvpdF\\_Ju1vvJM\\_5crGci2R4zm6EgUIV\\_RK-kwVZqOwuYiX4wsZLCJGLNaeJRrMlgw](https://ieeexplore.ieee.org/abstract/document/9688093?casa_token=9i9SkqxnXM0AAAAA:GSL5eW_VCmvYFjvpdF_Ju1vvJM_5crGci2R4zm6EgUIV_RK-kwVZqOwuYiX4wsZLCJGLNaeJRrMlgw)

# LAYER-WISE ANALYSIS OF A SELF-SUPERVISED SPEECH REPRESENTATION MODEL

Ankita Pasad, Ju-Chieh Chou, Karen Livescu

Toyota Technological Institute at Chicago

{ankitap, jcchou, klivescu}@ttic.edu



**Fig. 6.** *MI with word labels (max: 6.2).*

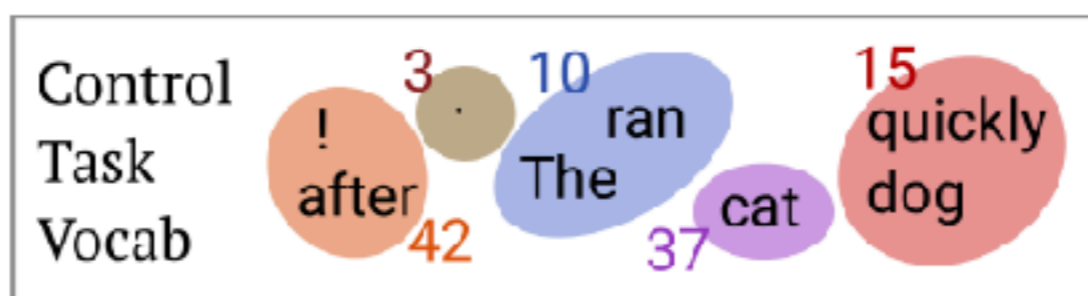
# Limitations of Probing

- Probes' negative results are hard to interpret
  - Lack of knowledge in the original model?
  - Wrong probe or insufficient probing data?
- Probes' positive results can be accidental
  - Prone to false positives : they can learn beyond the original model's encoded knowledge
  - Often not reliable when applied to complex, structured knowledge

## Designing and Interpreting Probes with Control Tasks

**John Hewitt**  
 Stanford University  
 johnhew@stanford.edu

**Percy Liang**  
 Stanford University  
 pliang@cs.stanford.edu



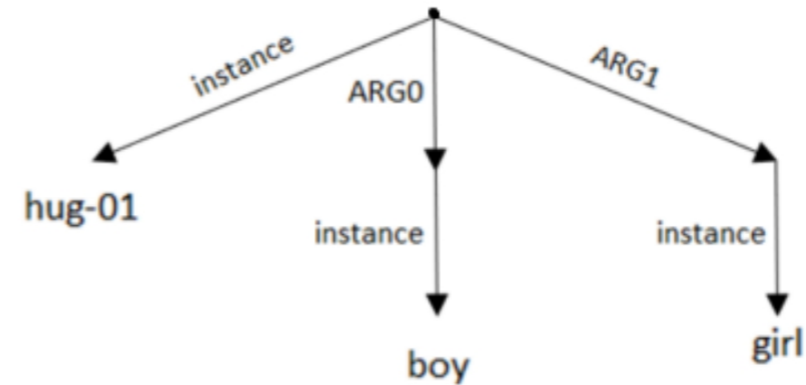
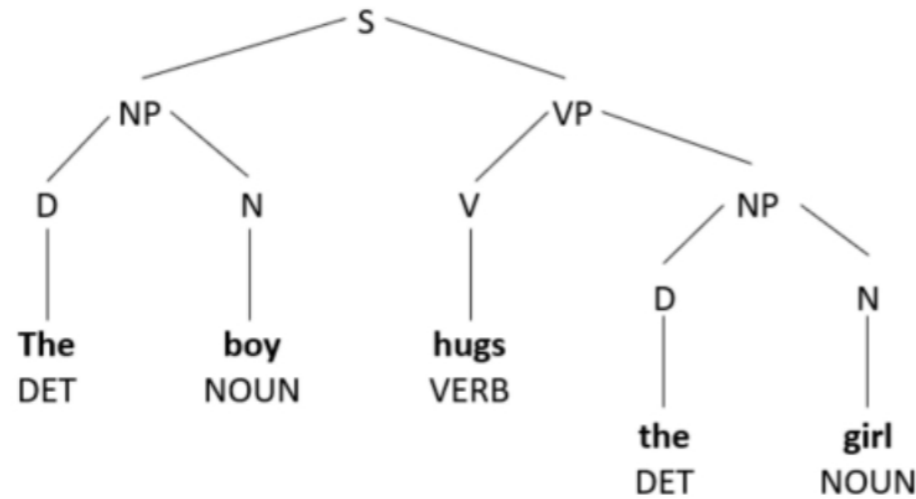
Sentence 1	The	cat	ran	quickly	.
<b>Part-of-speech</b>	DT	NN	VBD	RB	.
<b>Control task</b>	<b>10</b>	<b>37</b>	<b>10</b>	<b>15</b>	<b>3</b>
Sentence 2	The	dog	ran	after	!
<b>Part-of-speech</b>	DT	NN	VBD	IN	.
<b>Control task</b>	<b>10</b>	<b>15</b>	<b>10</b>	<b>42</b>	<b>42</b>

<https://aclanthology.org/D19-1275.pdf>

# Interpretability Techniques

1. Input perturbation
2. Probing/diagnostic classifiers or regressors
3. Representational Similarity Analysis (RSA)
4. Attribution and context mixing scores
5. Neuron-level interpretation
6. Inducing explainable architectures

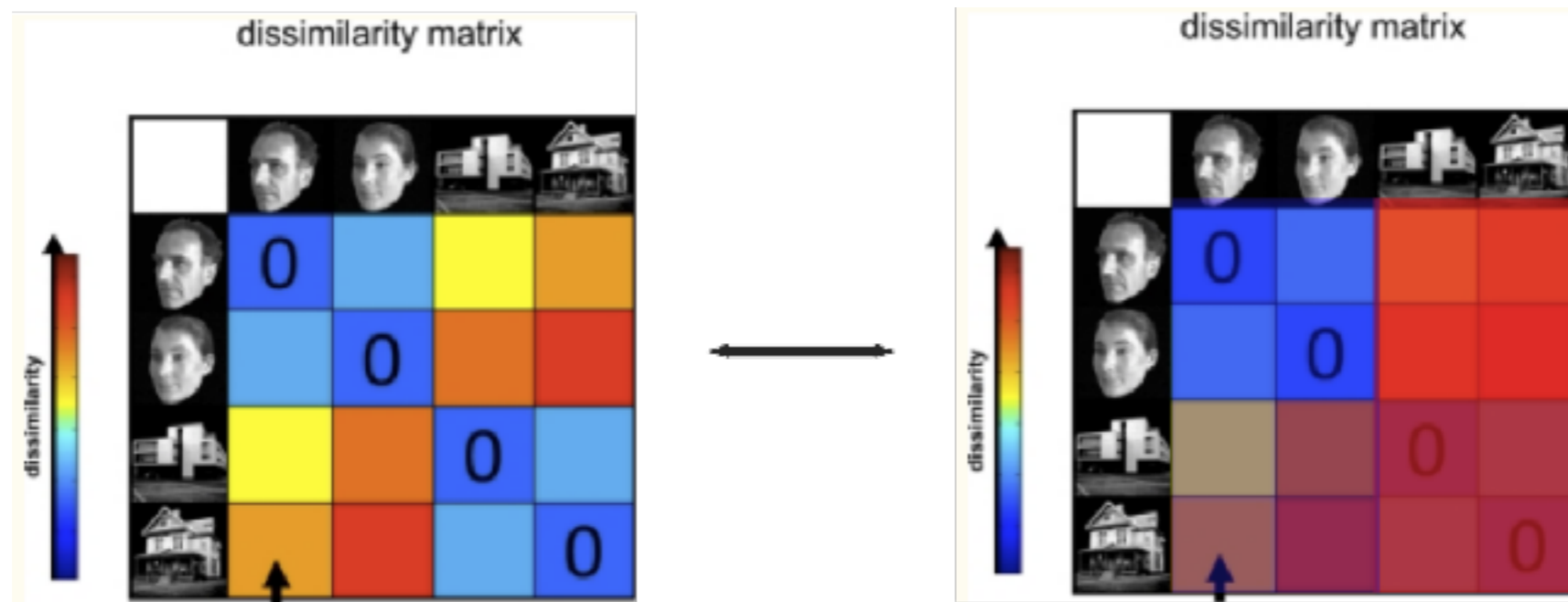
# Looking for Structured Knowledge



$\exists h, b, g:$   
instance(a, hug-01)  $\wedge$   
instance(b, boy)  $\wedge$   
instance(c, girl)  $\wedge$   
ARG0(a, b)  $\wedge$   
ARG1(a, c)

# Representational Similarity Analysis (RSA)

- RSA (Kriegeskorte et al., 2008): Measuring correlations between representations A and B in a similarity space
  - Compute representation (dis)similarity matrices in two spaces
  - Measure correlations between upper triangles





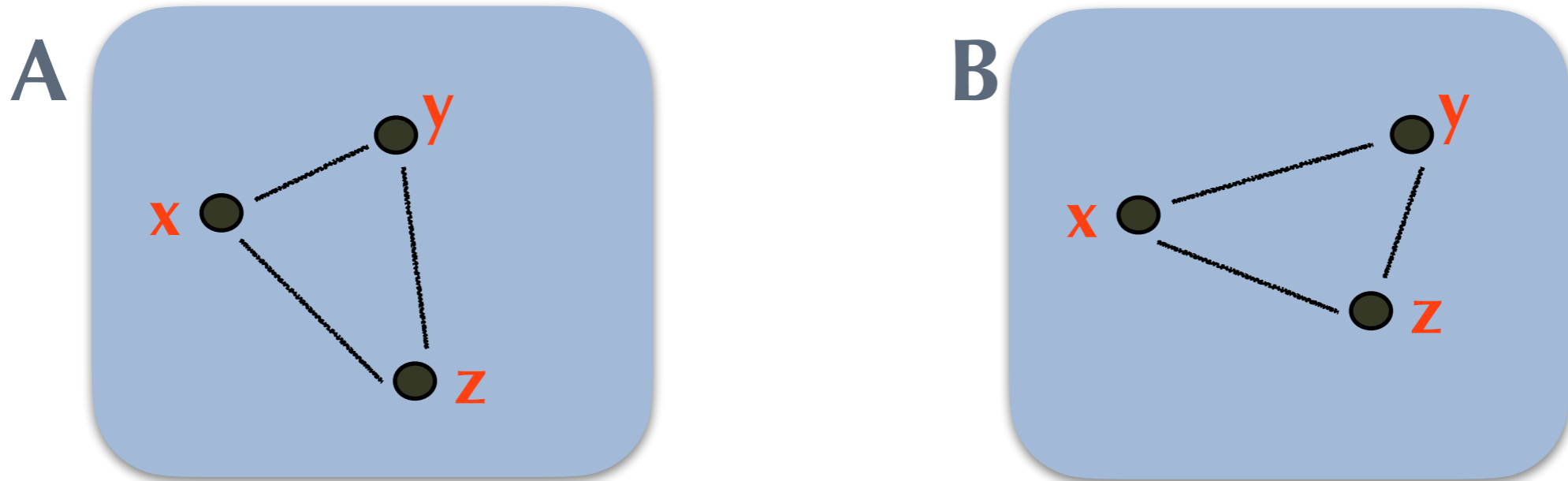
# RSA: An Example

- Sentence similarity according to
  - Sim A: human judgment
  - Sim B: estimated by a model

Stimulus 1	Stimulus 2	Sim A	Sim B
A slice of pizza	A bowl of salad	7.0	6.2
Two dogs run	A kitty running	8.0	9.0
A yellow and white bird	A kitty running	1.0	4.5

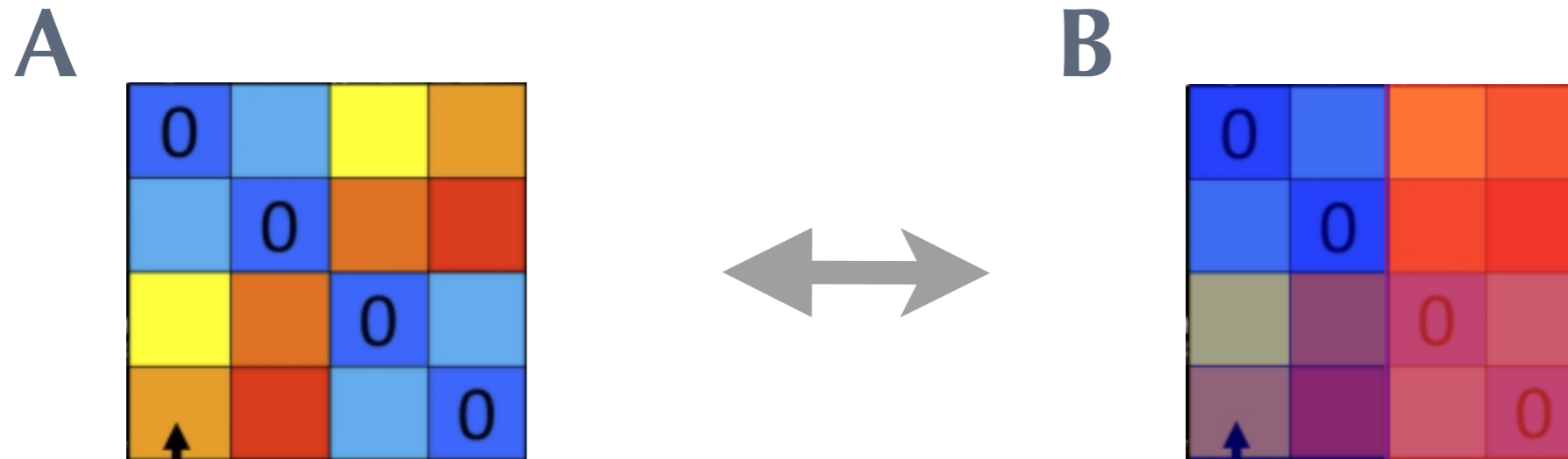
- RSA score: correlation between Sim A and Sim B

# Applying RSA to Language



- What we need: a similarity metric within two spaces A & B
  - Eg., A is a vector space, B is a space of trees or graphs
- What we do not need: a mapping between space A & B

# Applying RSA to Language



- What we need: a similarity metric within two spaces A & B
  - Eg., A is a vector space, B is a space of trees or graphs
- What we do not need: a mapping between space A & B

# Correlating neural and symbolic representations of language

**Grzegorz Chrupala**

Tilburg University  
g.chrupala@uvt.nl

**Afra Alishahi**

Tilburg University  
a.alishahi@uvt.nl

## Abstract

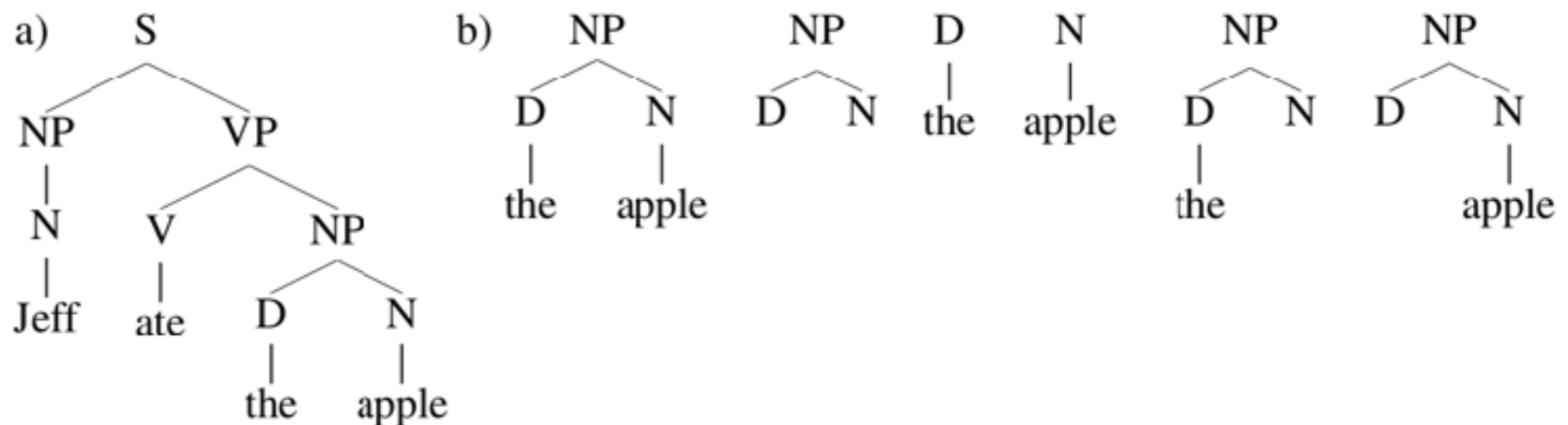
Analysis methods which enable us to better understand the representations and functioning of neural models of language are increasingly needed as deep learning becomes the dominant approach in NLP. Here we present two methods based on Representational Similarity Analysis (RSA) and Tree Kernels (TK) which allow us to directly quantify how strongly the

structure prediction algorithms, running the risk that the analytic method becomes no simpler than the actual neural model.

Here we introduce an alternative approach based on correlating neural representations of sentences and structured symbolic representations commonly used in linguistics. Crucially, the correlation is in similarity space rather than in the original representation space, preserving most of

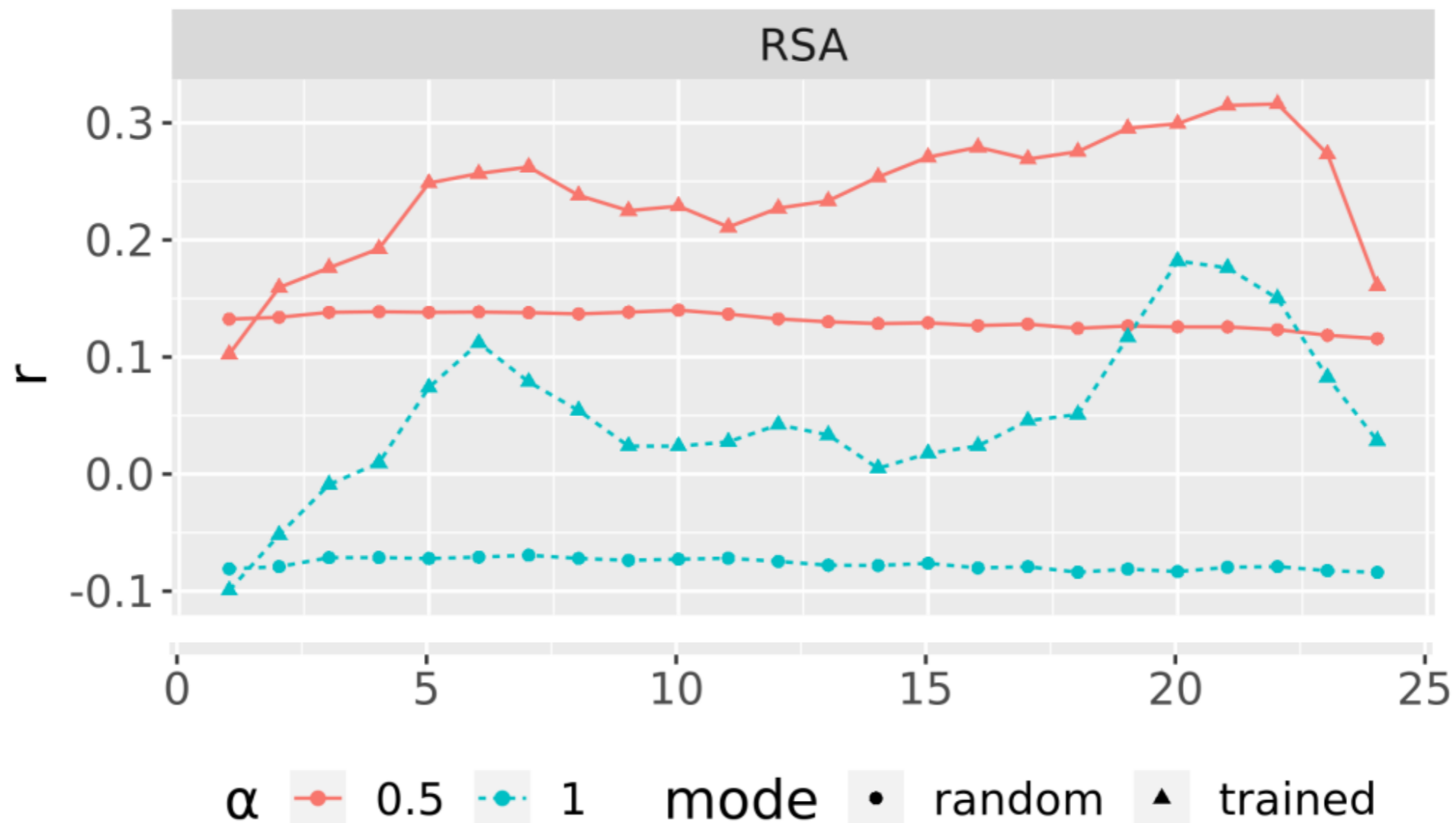
# Tree Kernels

- Measuring the similarity between two syntactic trees: count their overlapping subtrees



<https://aclanthology.org/P19-1283/>

# Applying RSA to BERT



<https://aclanthology.org/P19-1283/>

# RSA for Stability Analysis

## **Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains**

**Samira Abnar   Lisa Beinborn   Rochelle Choenni   Willem Zuidema**

Institute for Logic, Language and Computation  
University of Amsterdam

{abnar, l.beinborn}@uva.nl, rochelle.choenni@student.uva.nl, zuidema@uva.nl

### **Abstract**

In this paper, we define and apply *representational stability analysis* (ReStA), an intuitive way of analyzing neural language models. ReStA is a variant of the popular *representational similarity analysis* (RSA) in cognitive neuroscience. While RSA can be used

is simple: instead of directly trying to map models to brains, we first construct two similarity matrices that record how similar brain responses are to each other for different stimuli, and how similar the computational model's representations for each stimulus are to each other. The representational similarity score is then defined as the simi-

# Interpretability Techniques

1. Input perturbation
2. Probing/diagnostic classifiers or regressors
3. Representational Similarity Analysis (RSA)
4. Attribution and context mixing scores
5. Neuron-level interpretation
6. Inducing explainable architectures



# Feature Attribution & Context Mixing

- **Feature attribution:** the contribution of each input token to the model's output
- **Context mixing:** the contribution of each input token to the representation of other tokens

Review

# Explainable AI: A Review of Machine Learning Interpretability Methods

Pantelis Linardatos \* , Vasilis Papastefanopoulos  and Sotiris Kotsiantis 

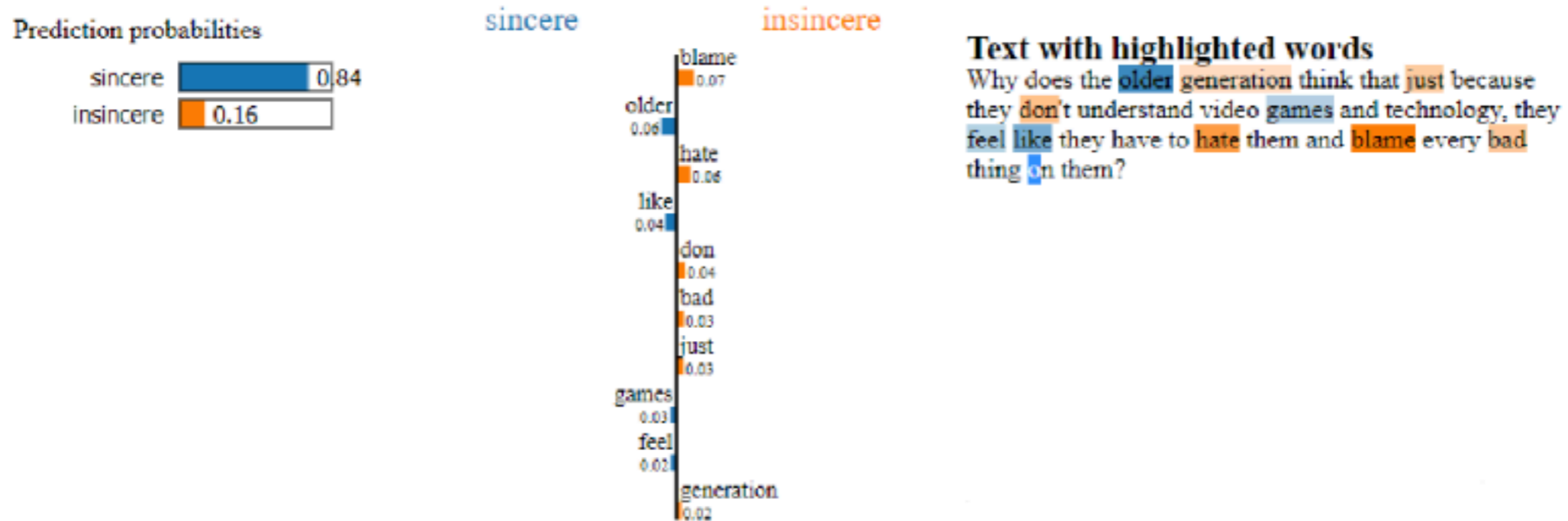
Department of Mathematics, University of Patras, 26504 Patras, Greece;  
vasileios.papastefanopoulos@upatras.gr (V.P.); sotos@math.upatras.gr (S.K.)

\* Correspondence: p.linardatos@upnet.gr

**Abstract:** Recent advances in artificial intelligence (AI) have led to its widespread industrial adoption, with machine learning systems demonstrating superhuman performance in a significant number of tasks. However, this surge in performance, has often been achieved through increased model complexity, turning such systems into “black box” approaches and causing uncertainty regarding the way they operate and, ultimately, the way that they come to decisions. This ambiguity has made it problematic for machine learning systems to be adopted in sensitive yet critical domains, where their value could be immense, such as healthcare. As a result, scientific interest in the field of Explainable Artificial Intelligence (XAI), a field that is concerned with the development of new methods that explain and interpret machine learning models, has been tremendously reignited over recent years. This study focuses on machine learning interpretability methods; more specifically, a literature review and taxonomy of these methods are presented, as well as links to their programming implementations, in the hope that this survey would serve as a reference point for both theorists and practitioners.

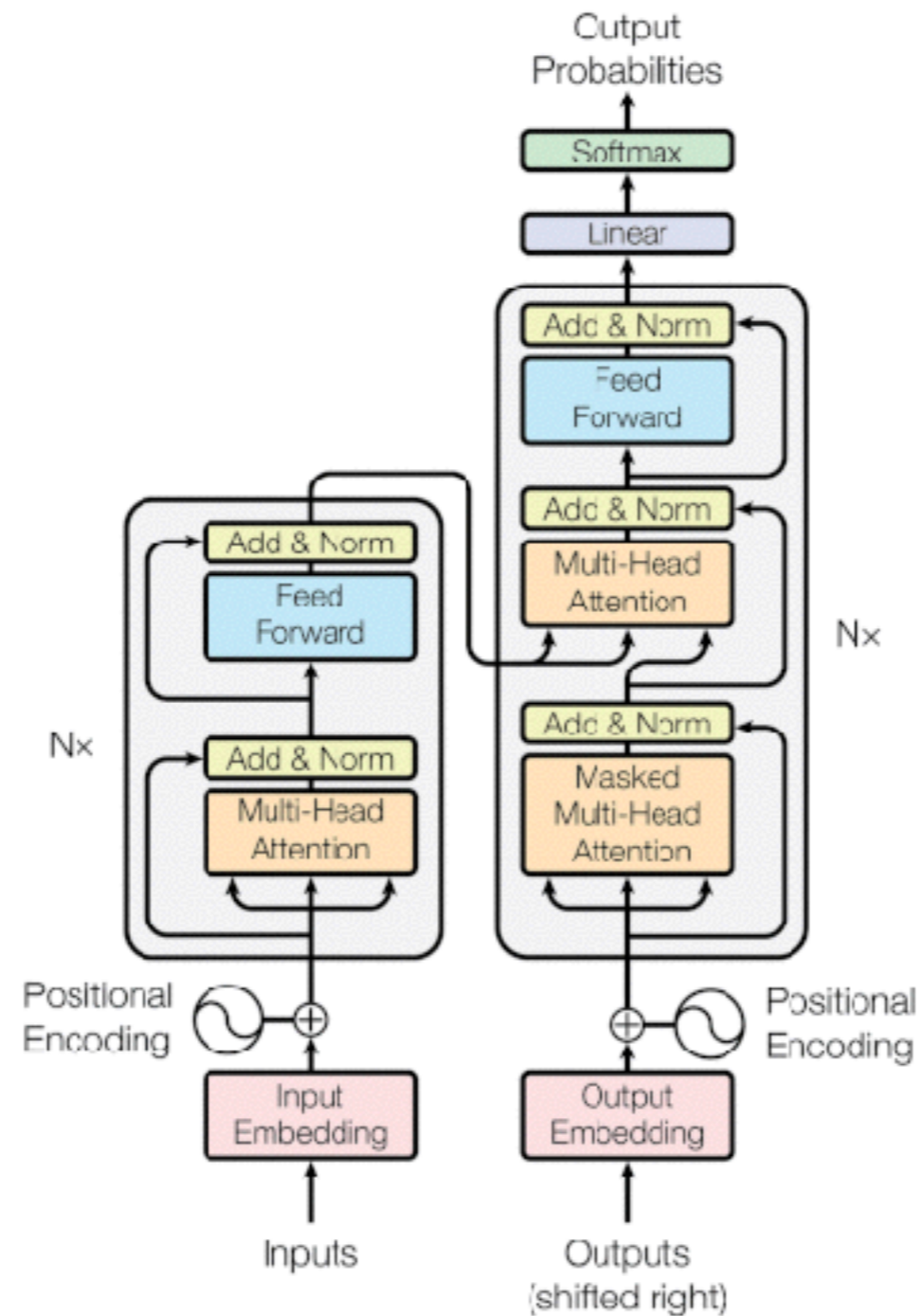
**Keywords:** xai; machine learning; explainability; interpretability; fairness; sensitivity; black-box

# Feature Attribution

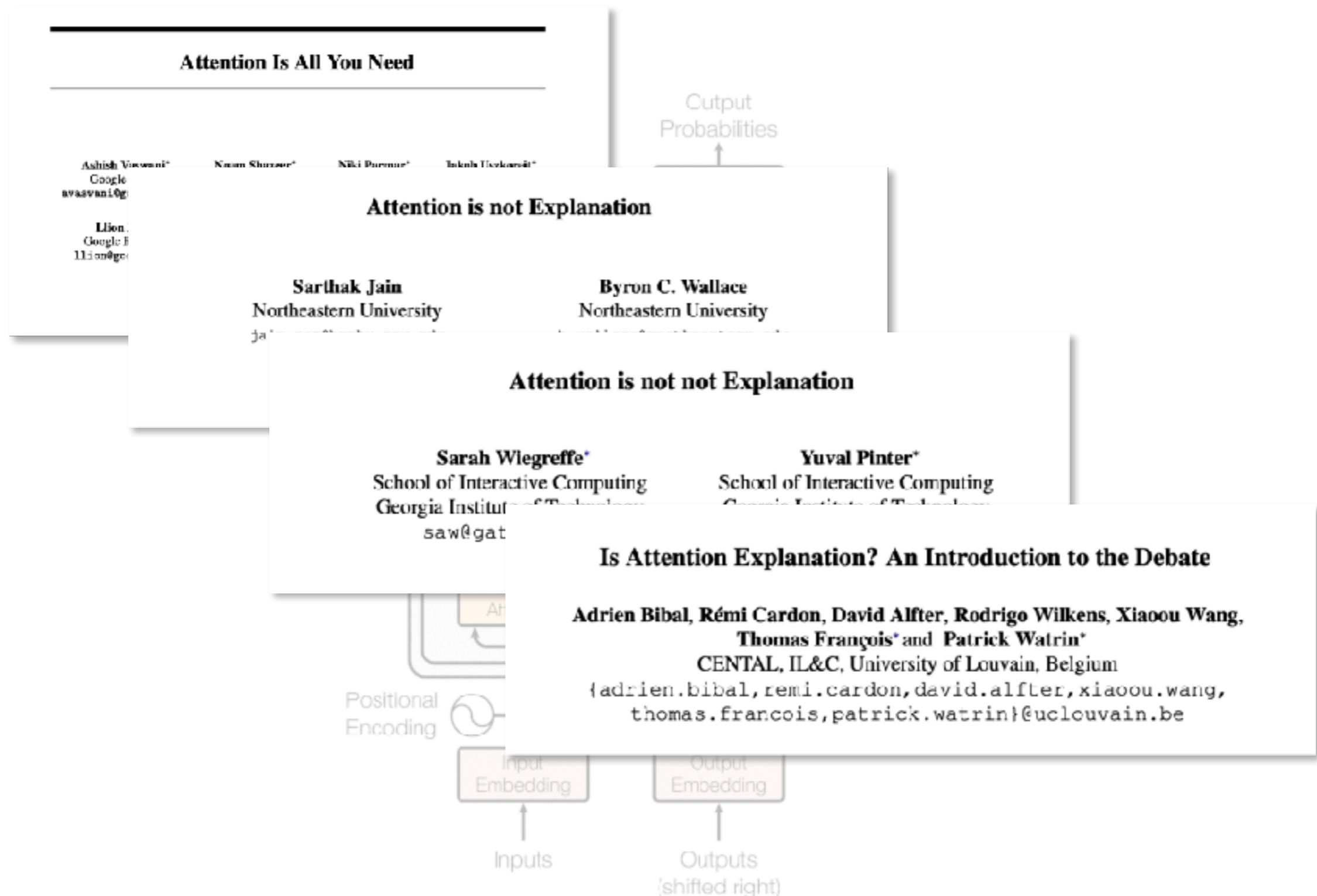


**Figure 4.** Local interpretable model-agnostic explanations (LIME) is used to explain the rationale behind the classification of an instance of the Quora Insincere Questions Dataset.

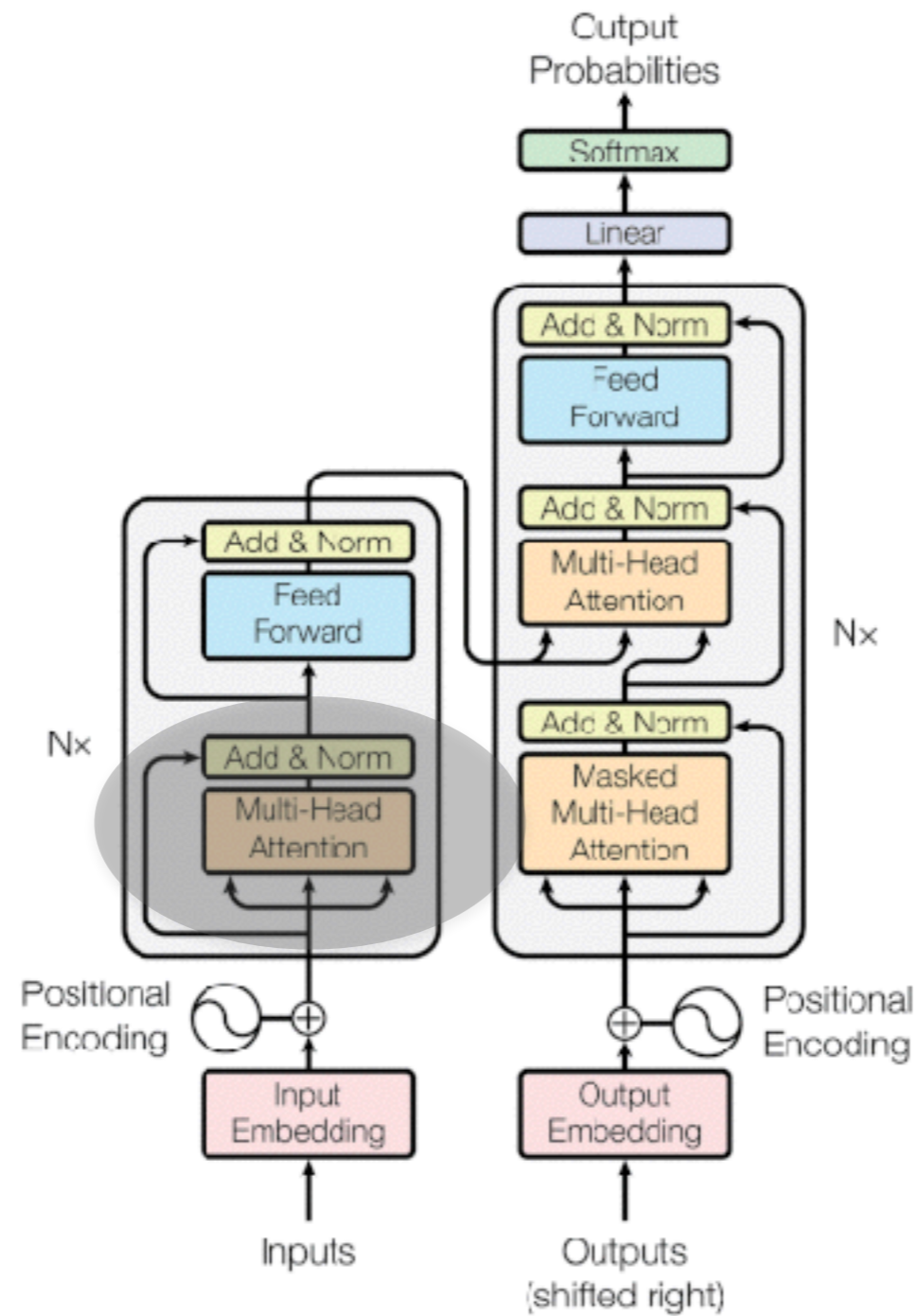
# Where to Look?



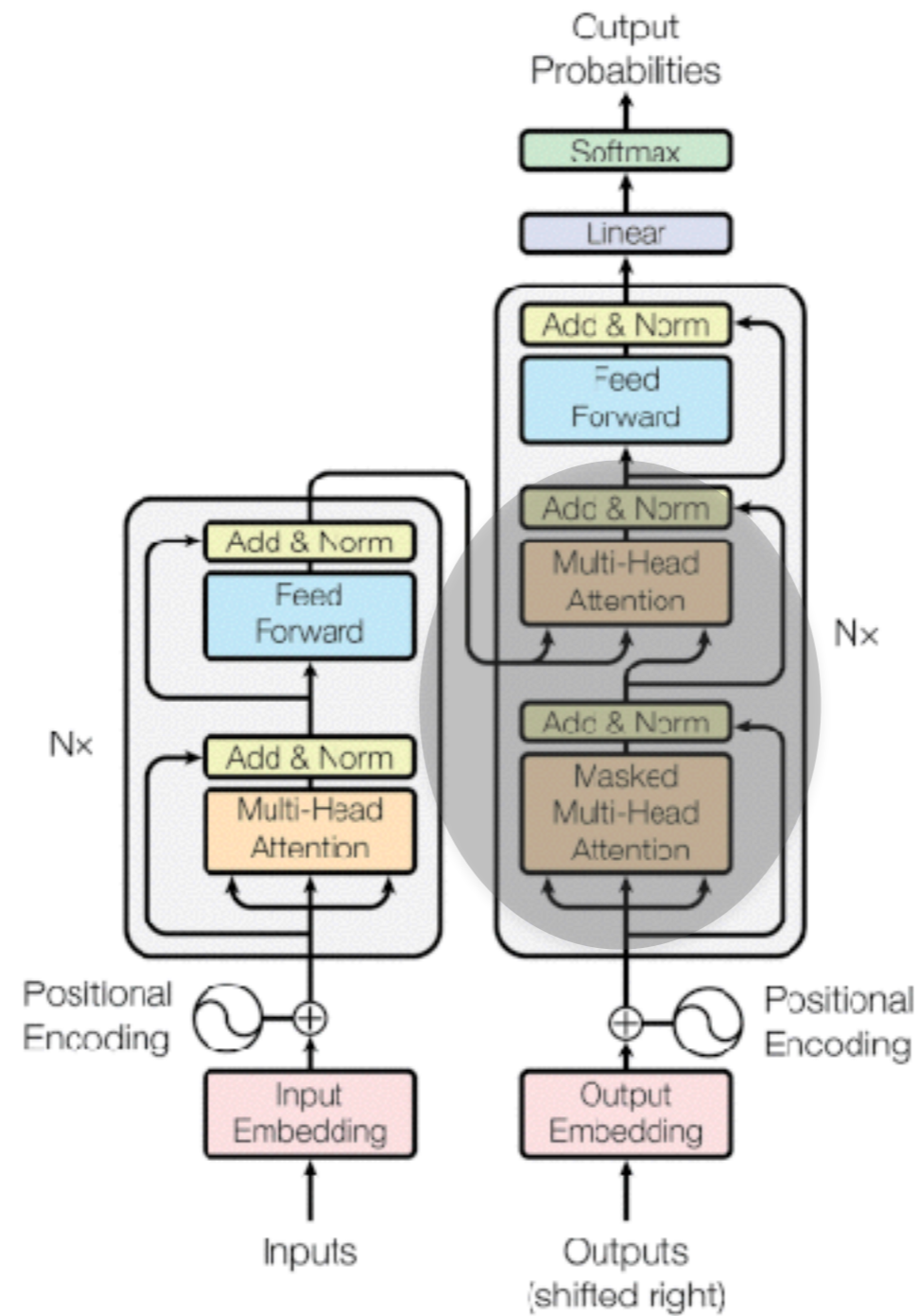
# Where to Look?



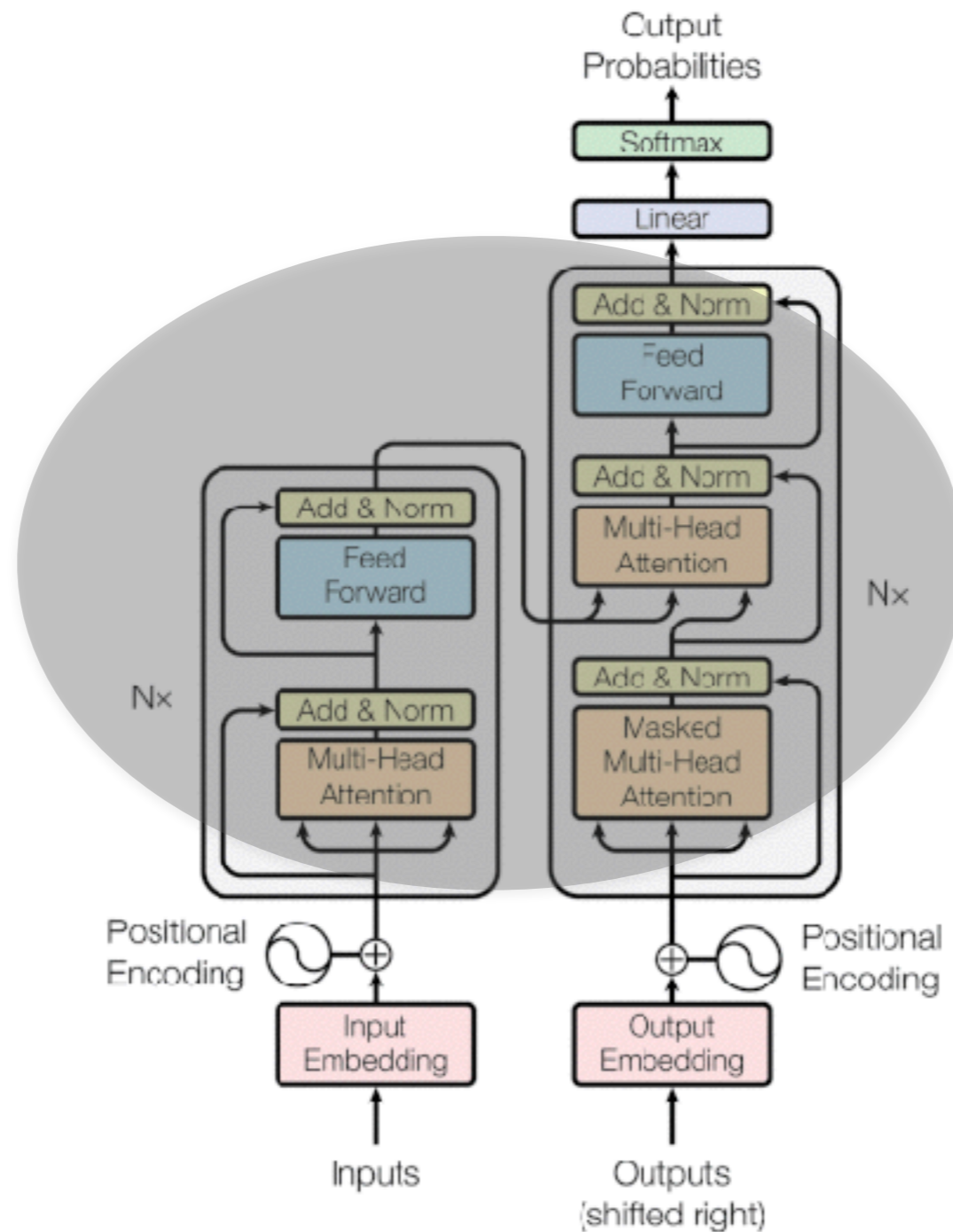
# Where to Look?



# Where to Look?



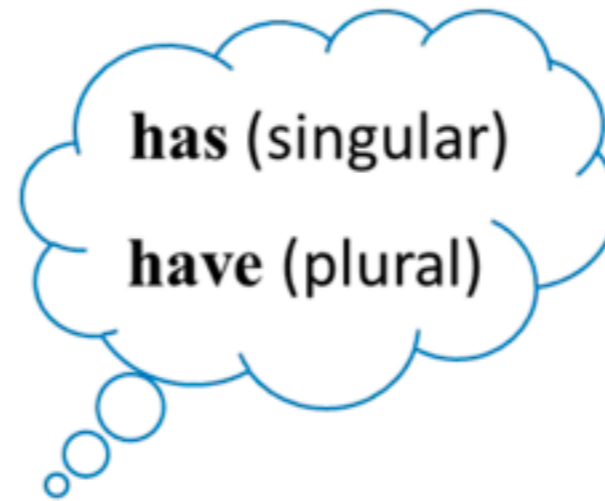
# Where to Look?





# Context Mixing

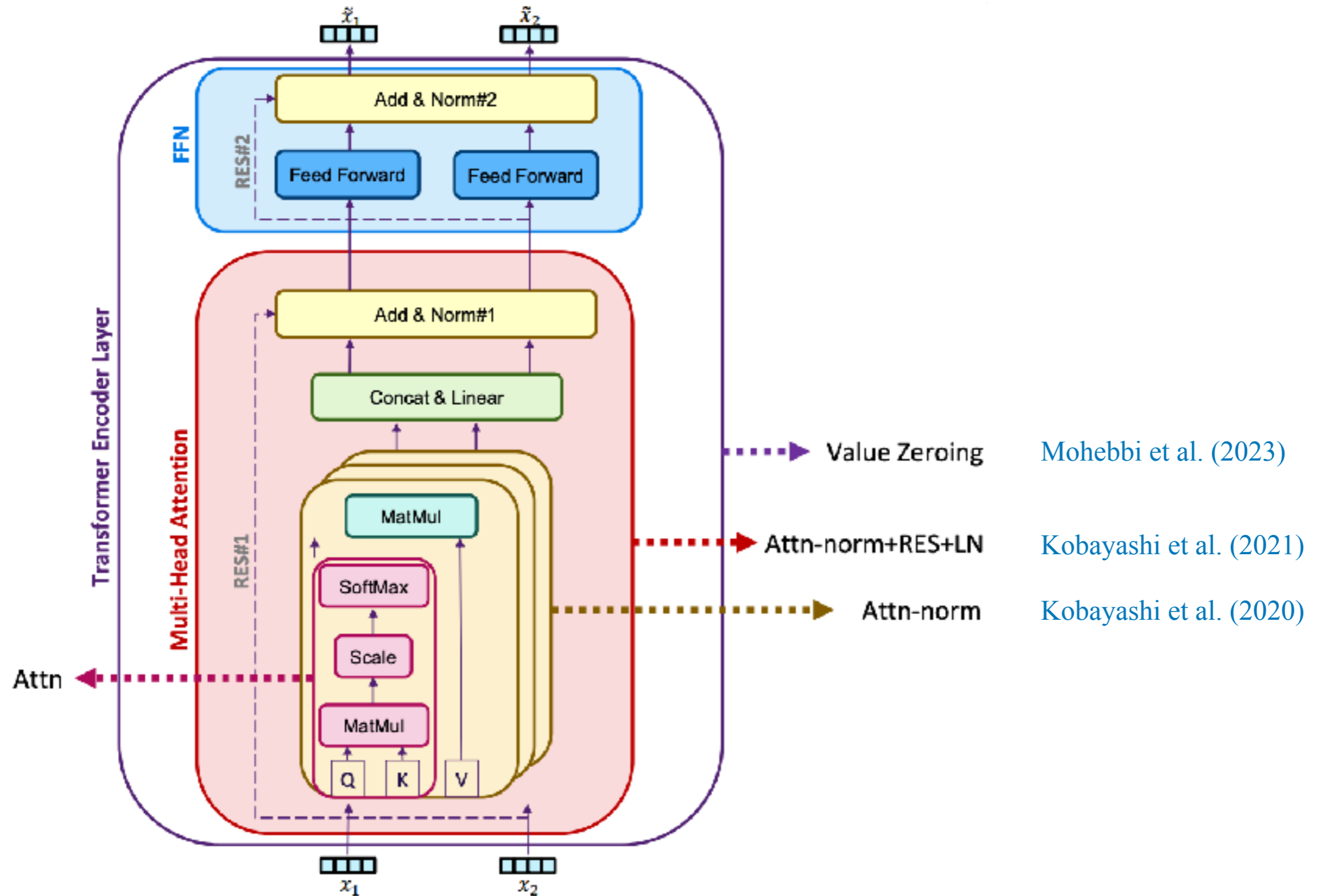
Target



My friend [MASK] fixed this **chair**.



# Context Mixing: Current Landscape



# Quantifying Context Mixing in Transformers

Hosein Mohebbi<sup>1</sup> Willem Zuidema<sup>2</sup> Grzegorz Chrupała<sup>1</sup> Afra Alishahi<sup>1</sup>

<sup>1</sup> CSAI, Tilburg University <sup>2</sup> ILLC, University of Amsterdam

{h.mohebbi, a.alishahi}@tilburguniversity.edu

w.h.zuidema@uva.nl

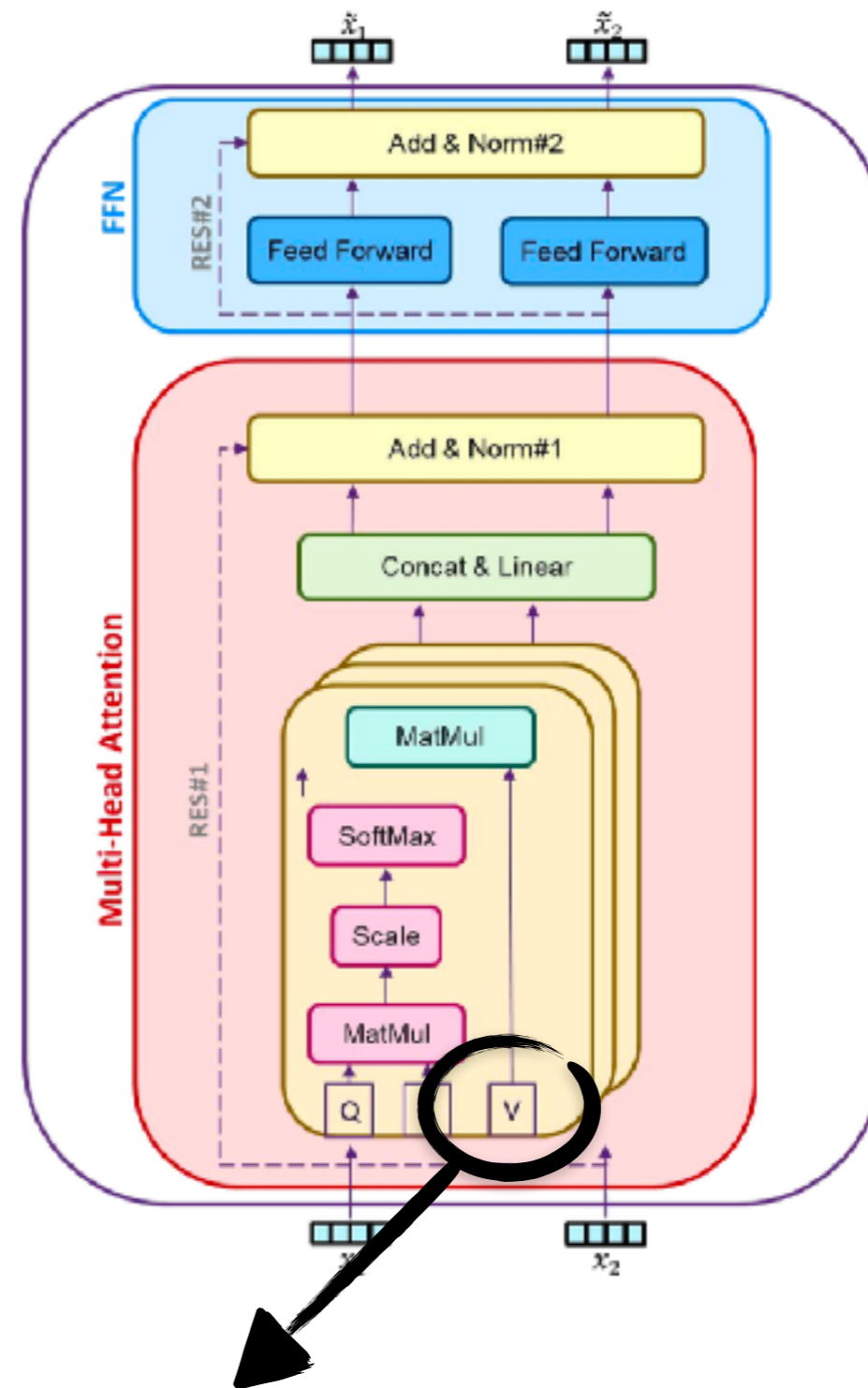
grzegorz@chrupala.me

## Abstract

Self-attention weights and their transformed variants have been the main source of information for analyzing token-to-token interactions in Transformer-based models. But despite their ease of interpretation, these weights are not faithful to the models' decisions as they are only one part of an encoder, and other com-

point for understanding this flow, and these weights ('raw attention') have been used in many studies (Clark et al., 2019; Kovaleva et al., 2019; Reif et al., 2019; Htut et al., 2019a, *inter alia*). However, the reliability and usefulness of raw attention weights has also been questioned (Jain and Wallace, 2019; Bibal et al., 2022). In particular, attention weights tend to concentrate on uninformative tokens in the

# Value Zeroing



Set the value vector of a **cue** token to zero, then measure impact on the **target** token

# Value Zeroing

Dot Product

Layer	Rand	Attn	Attn-norm	Attn-norm + RES	Attn-norm + RES + LN	GlobEnc	ALTI	GradXinput	IG	DL	Value Zeroing
12	0.13	0.04	0.17	0.03	0.02	0.04	0.02	0.00	0.00	0.00	0.21
11	0.13	0.09	0.19	0.06	0.06	0.05	0.03	0.05	0.03	0.03	0.26
10	0.12	0.09	0.18	0.06	0.06	0.05	0.04	0.09	0.05	0.07	0.21
9	0.12	0.12	0.22	0.10	0.09	0.09	0.06	0.10	0.06	0.09	0.28
8	0.12	0.12	0.19	0.09	0.09	0.08	0.07	0.12	0.08	0.10	0.21
7	0.12	0.09	0.17	0.09	0.07	0.08	0.06	0.14	0.09	0.11	0.17
6	0.12	0.08	0.16	0.08	0.07	0.07	0.06	0.13	0.07	0.12	0.17
5	0.12	0.10	0.19	0.10	0.08	0.09	0.08	0.15	0.06	0.13	0.24
4	0.12	0.10	0.18	0.10	0.07	0.08	0.07	0.16	0.08	0.14	0.21
3	0.12	0.09	0.15	0.08	0.05	0.07	0.07	0.17	0.09	0.16	0.19
2	0.12	0.08	0.13	0.08	0.04	0.07	0.06	0.18	0.12	0.18	0.14
1	0.12	0.06	0.11	0.07	0.04	0.06	0.04	0.19	0.12	0.19	0.10

# Information Rollout & Flow

## Quantifying Attention Flow in Transformers

**Samira Abnar**

ILLC, University of Amsterdam  
s.abnar@uva.nl

**Willem Zuidema**

ILLC, University of Amsterdam  
w.h.zuidema@uva.nl

### Abstract

In the Transformer model, “self-attention” combines information from attended embeddings into the representation of the focal embedding in the next layer. Thus, across layers of the Transformer, information originating from different tokens gets increasingly mixed.

We propose two simple but effective methods to compute attention scores to input tokens (i.e., *token attention*) at each layer, by taking raw attentions (i.e., *embedding attention*) of that layer as well as those from the precedent layers. These methods are based on modelling the information flow in the network with a *DAG* (Directed Acyclic Graph), in

# Information Rollout & Flow

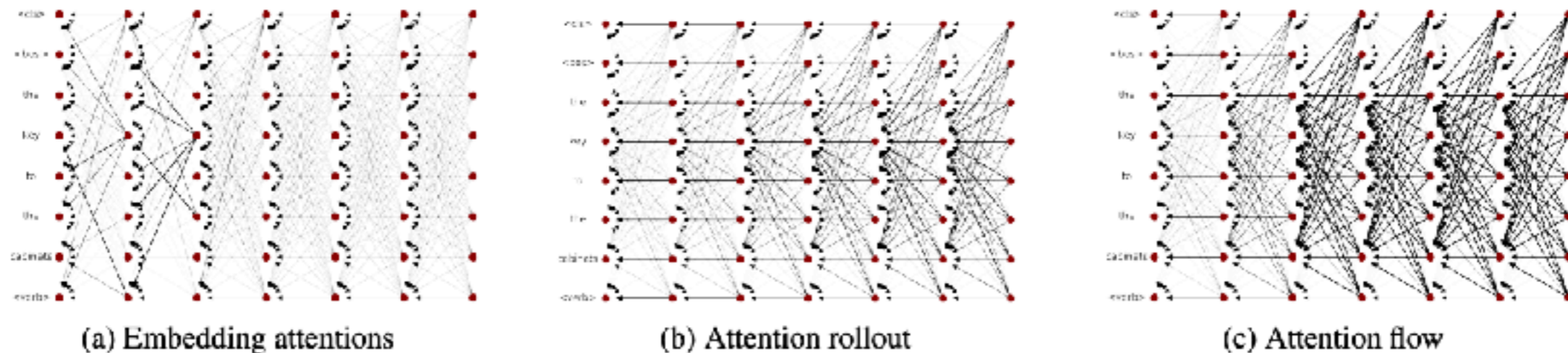


Figure 1: Visualisation of attention weights.

<https://www.mdpi.com/1099-4300/23/1/18>

- The same techniques can be applied to more complex attribution and context-mixing scores

# Interpretability Techniques

1. Input perturbation
2. Probing/diagnostic classifiers or regressors
3. Representational Similarity Analysis (RSA)
4. Attribution and context mixing scores
5. Neuron-level interpretation
6. Inducing explainable architectures



# Neuron-level Interpretation of Deep NLP Models: A Survey

Hassan Sajjad♣\* Nadir Durrani♣\* Fahim Dalvi♣\*

♣Faculty of Computer Science, Dalhousie University, Canada†

♣Qatar Computing Research Institute, HBKU, Doha, Qatar

hsajjad@dal.ca, {ndurrani, faimaduddin}@hbku.edu.qa

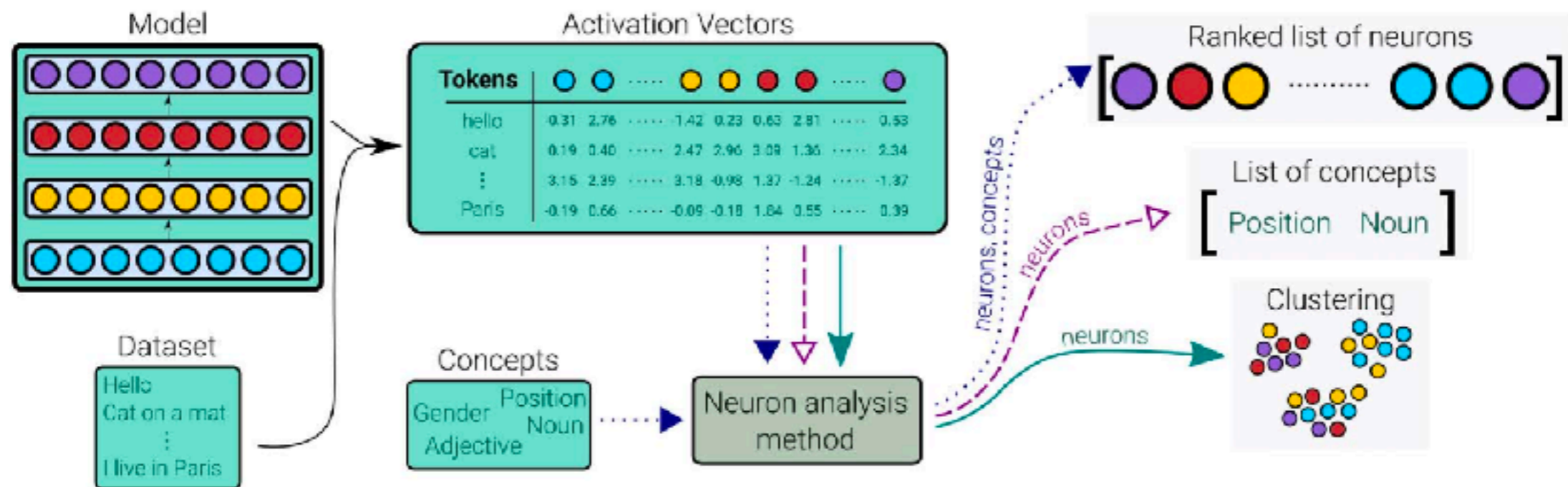
## Abstract

The proliferation of Deep Neural Networks in various domains has seen an increased need for interpretability of these models. Preliminary work done along this line, and papers that surveyed such, are focused on high-level representation analysis. However, a recent branch of work has concentrated on interpretability at a more granular level of analyzing neurons within these models. In this paper, we survey the work done on neuron analysis including: i)

models and to answer one question in particular: *What knowledge is learned within representations?* We term this work as the *Representation Analysis*.

Representation Analysis thrives on post-hoc decomposability, where we analyze the embeddings to uncover linguistic (and non-linguistic) concepts<sup>1</sup> that are captured as the network is trained towards an NLP task (Adi et al., 2016; Belinkov et al., 2017a; Conneau et al., 2018; Liu et al., 2019; Tenney et al., 2019). A majority of

# Overview of Neuron Analysis Methods



[https://direct.mit.edu/tacl/article/doi/10.1162/tacl\\_a\\_00519/113852/Neuron-level-Interpretation-of-Deep-NLP-Models-A](https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00519/113852/Neuron-level-Interpretation-of-Deep-NLP-Models-A)

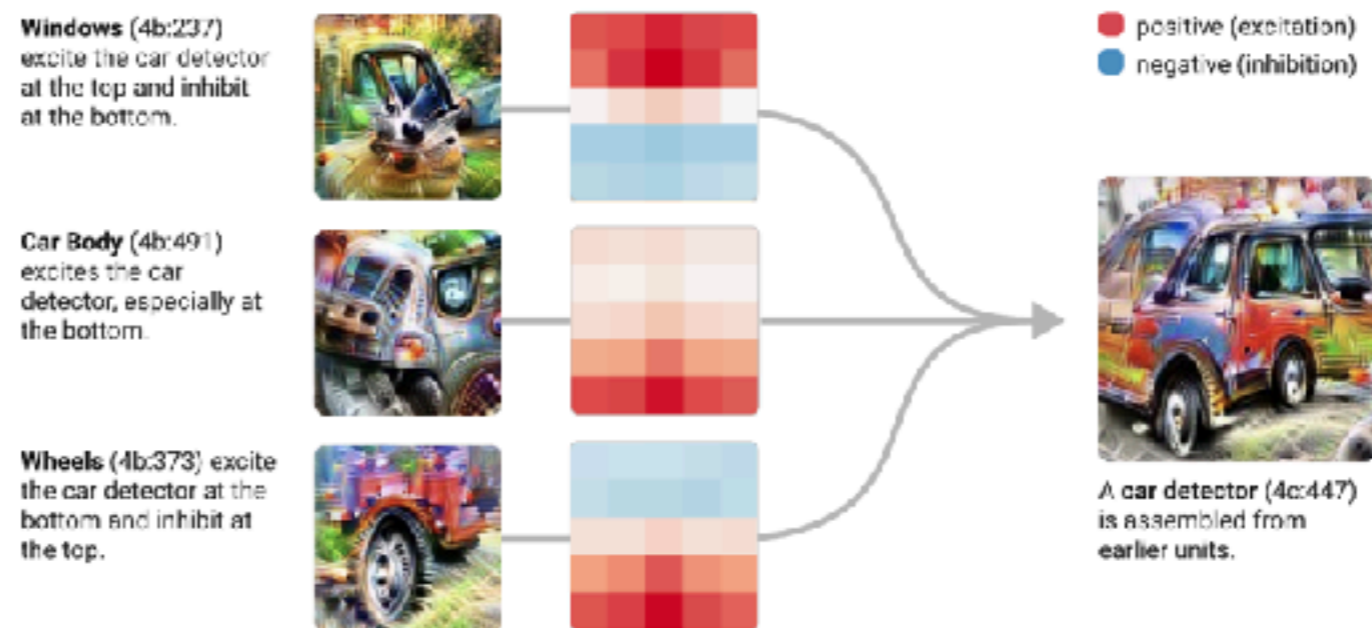
	Scope	Input	Output	Scalability	HITL	Supervision	Causation
<b>Visualization</b>							
Karpathy et al. (2015)	local	neuron	concept	low	yes	no	no
<b>Corpus-based methods</b>							
Concept Search							
Kádár et al. (2017)	global	neuron	concept	low	yes	no	no
Na et al. (2019)	global	neuron	concept	high	no	no	no
Neuron Search							
Mu and Andreas (2020); Suau et al. (2020); Antverg and Belinkov (2022)	global	concept	neurons	high	no	yes	no
<b>Probing-based methods</b>							
Linear (Dalvi et al., 2019)	global	concept	neurons	high	no	yes	no
Gaussian (Hennigen et al., 2020)	global	concept	neurons	high	no	yes	no
<b>Causation-based methods</b>							
Ablation (Lakretz et al., 2019)	both	concept/ class	neurons	medium	no	no	yes
Knowledge attribution (Dai et al., 2021)	local	concept/ class	neurons	high	no	no	yes
<b>Miscellaneous methods</b>							
Corpus generation (Poerner et al., 2018)	global	neuron	concept	low	yes	no	no
Matrix factorization (Alammar, 2020)	local	neurons	neurons	low	yes	no	no
Clustering (Dalvi et al., 2020)	global	neurons	neurons	high	yes	no	no
Multi model search (Bau et al., 2019)	global	neurons	neurons	high	yes	no	no

[https://direct.mit.edu/tacl/article/doi/10.1162/tacl\\_a\\_00519/113852/Neuron-level-Interpretation-of-Deep-NLP-Models-A](https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00519/113852/Neuron-level-Interpretation-of-Deep-NLP-Models-A)

# Mechanistic Interpretability

## Zoom In: An Introduction to Circuits

By studying the connections between neurons, we can find meaningful algorithms in the weights of neural networks.



### AUTHORS

Chris Olah  
Nick Cammarata  
Ludwig Schubert  
Gabriel Goh  
Michael Petrov  
Shan Carter

### AFFILIATIONS

OpenAI  
OpenAI  
OpenAI  
OpenAI  
OpenAI  
OpenAI

### PUBLISHED

March 10, 2020

### DOI

10.23915/distill.00024.001

# Mechanistic Interpretability

- Aim: discovering algorithmic explanations of the inner-mechanisms of deep neural models
  - Causal interpretability
  - Sub-layer levels (individual attention heads, single neurons)
- How is it different from earlier methods?

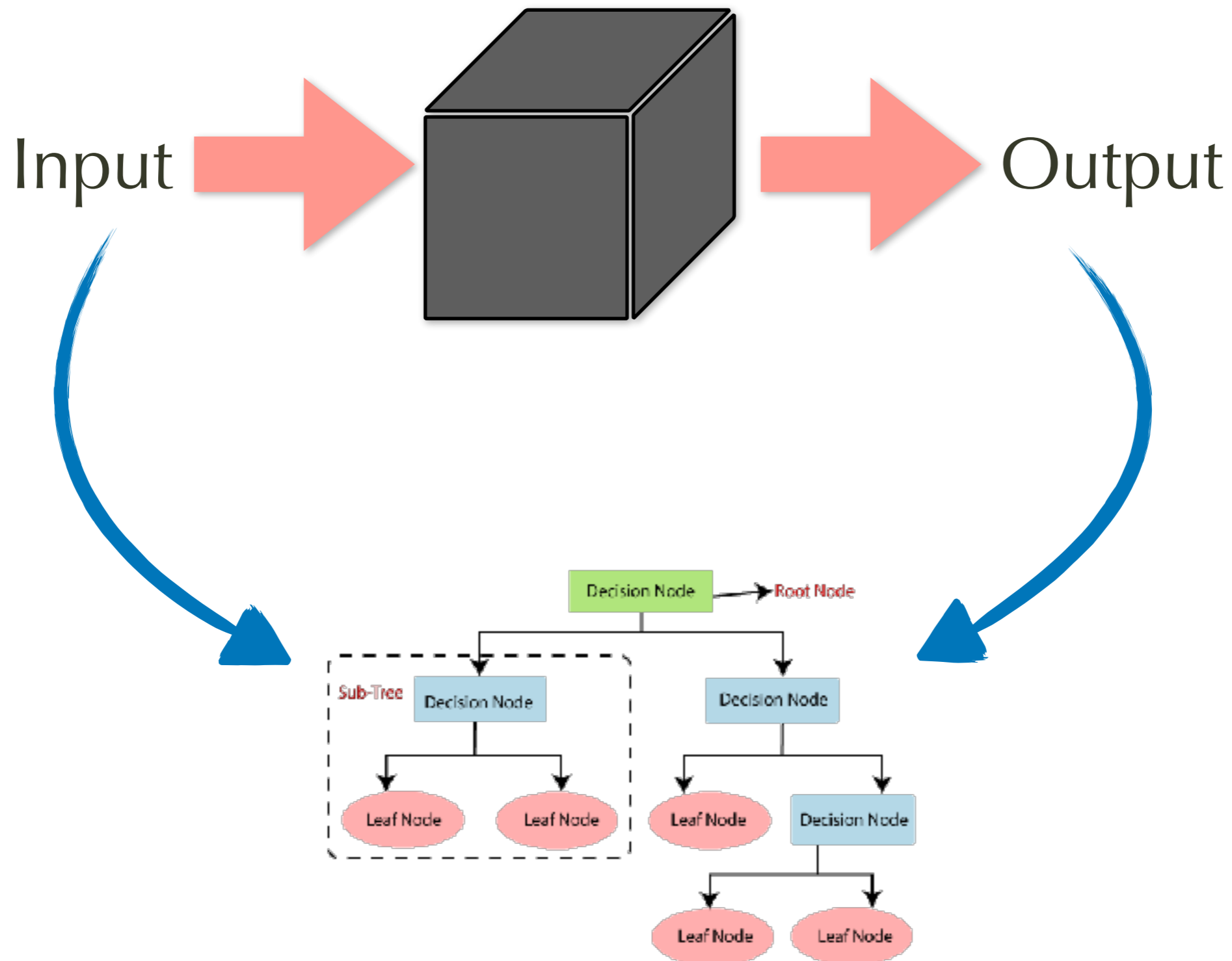
# Mechanistic Interpretability

- A framework for finding task-specific **circuits**
  - **Minimal computational subgraphs**: minimal set of components (nodes, edges) in a network sufficient for performing a task
  - If all other components are **ablated**, the task-specific loss is not affected
  - Various methods are proposed for ablation/intervention
  - Often computationally expensive (work in progress)

# Interpretability Techniques

1. Input perturbation
2. Probing/diagnostic classifiers or regressors
3. Representational Similarity Analysis (RSA)
4. Attribution and context mixing scores
5. Neuron-level interpretation
6. Inducing explainable architectures

# Student-Teacher Knowledge Distillation

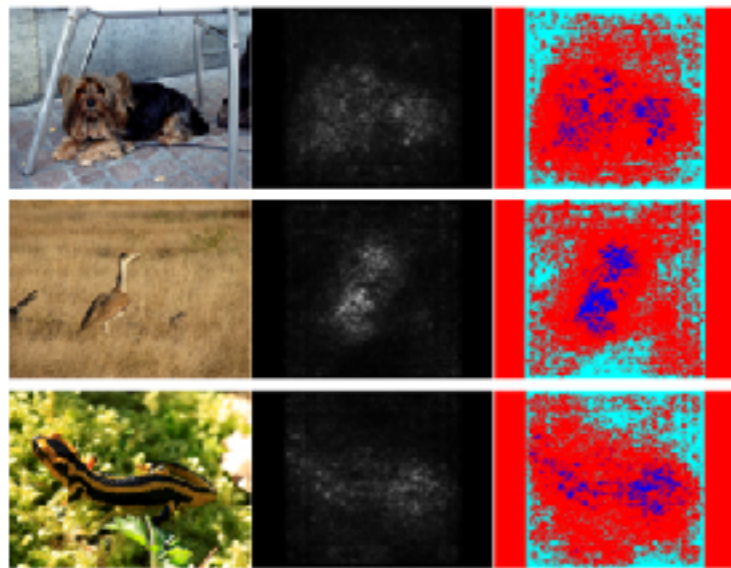




# Evaluation of Interpretability Methods

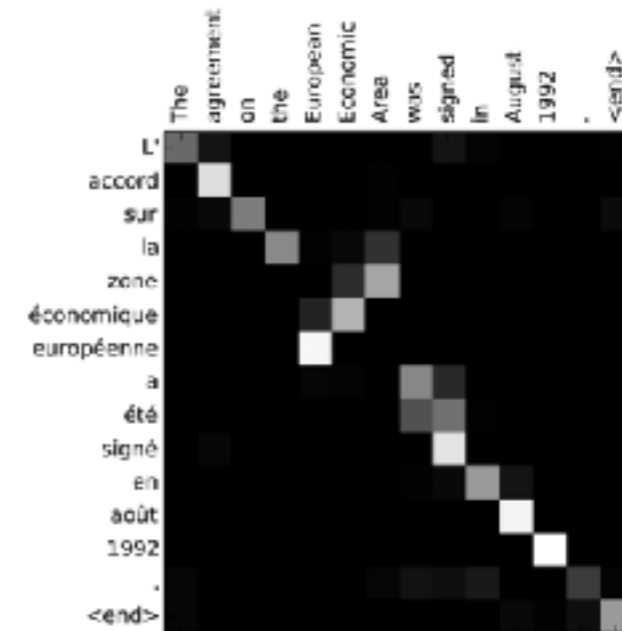
# Are Interpretations Plausible?

Saliency maps in an image classification model



<https://arxiv.org/pdf/1312.6034.pdf>

Attention weights in an NMT model

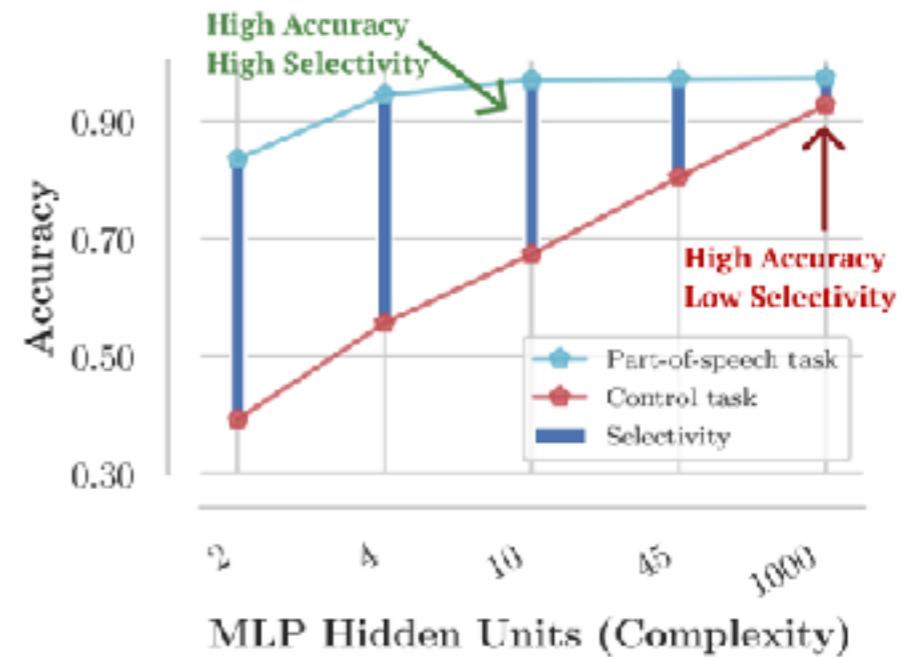


<https://arxiv.org/abs/1409.0473>

- How much do you trust what your interpretability technique tells you?
  - It makes sense to me! → *Plausibility*

# Are Interpretations Reliable?

Control Task Vocab	!	3	10	15	
	after	.	ran	quickly	dog
	42	37			
Sentence 1	The	cat	ran	quickly	.
Part-of-speech	DT	NN	VBD	RB	.
Control task	10	37	10	15	3
Sentence 2	The	dog	ran	after	!
Part-of-speech	DT	NN	VBD	IN	.
Control task	10	15	10	42	42



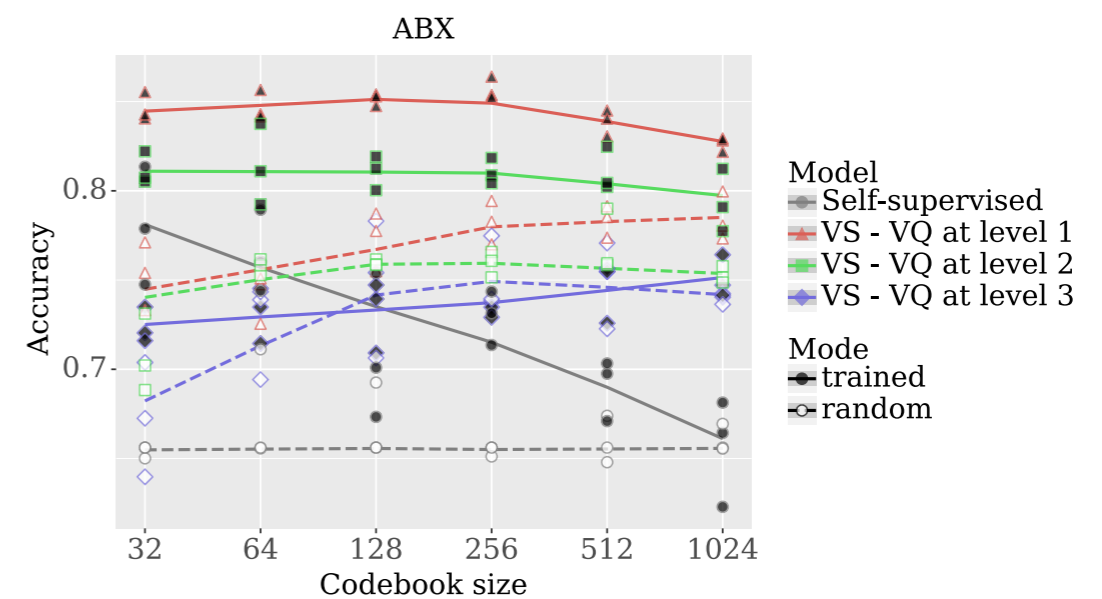
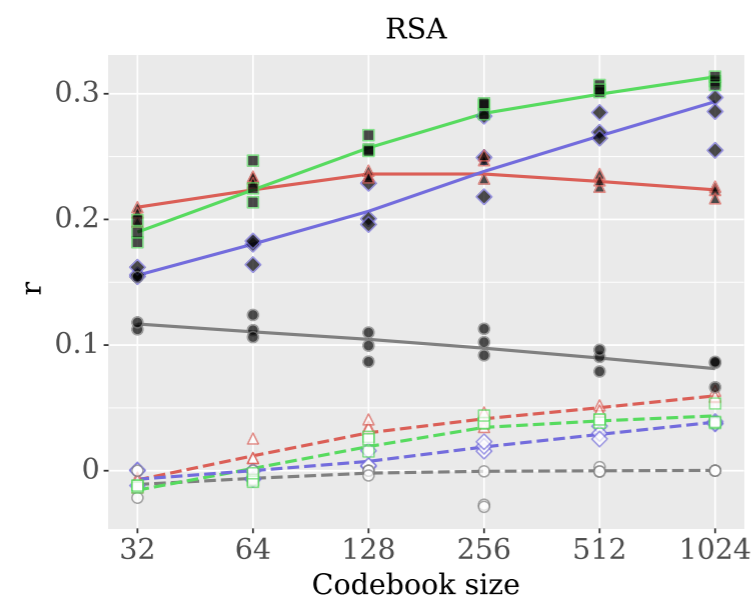
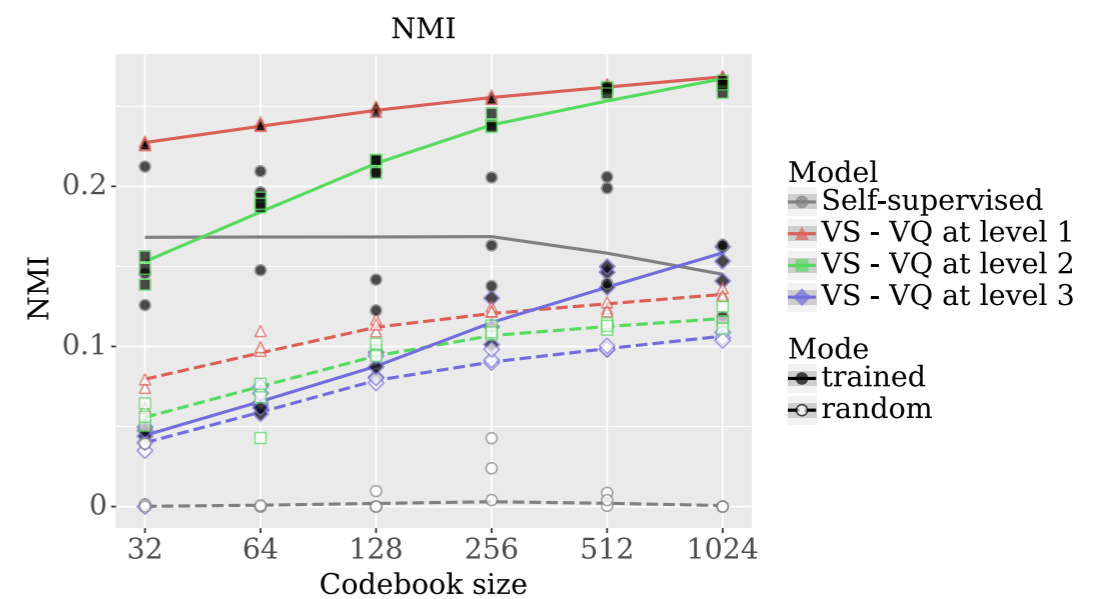
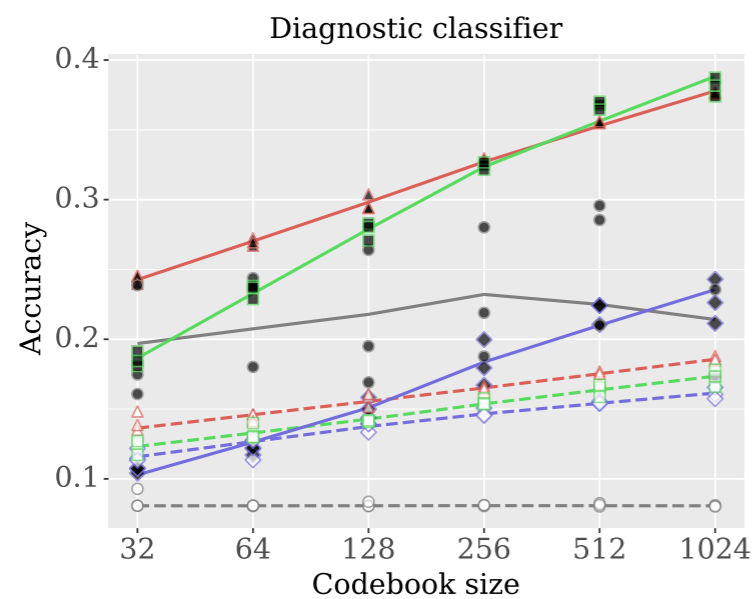
<https://aclanthology.org/D19-1275.pdf>

- Measure the difference between the accuracy of the *linguistic* and the *control* tasks

➔ *Selectivity*

# (In)consistency of Predictions

- Encoding of phonemes according to different methods:



# (In)consistency of Predictions

- Context mixing for Subject-Verb Agreement according to different scores:

12	0.33	0.24	0.43	0.27	0.27	0.27	0.28	0.12	0.12	0.12	0.44
11	0.33	0.29	0.54	0.32	0.31	0.32	0.33	0.23	0.16	0.18	0.54
10	0.31	0.28	0.45	0.29	0.29	0.29	0.29	0.29	0.17	0.21	0.45
9	0.32	0.32	0.57	0.34	0.34	0.34	0.33	0.26	0.17	0.22	0.57
8	0.32	0.29	0.46	0.30	0.30	0.30	0.30	0.28	0.19	0.23	0.47
7	0.32	0.27	0.43	0.29	0.29	0.29	0.29	0.34	0.19	0.25	0.43
6	0.32	0.25	0.41	0.28	0.28	0.28	0.29	0.34	0.18	0.27	0.41
5	0.33	0.27	0.53	0.33	0.33	0.33	0.32	0.40	0.18	0.30	0.54
4	0.31	0.21	0.51	0.31	0.31	0.31	0.29	0.45	0.20	0.36	0.48
3	0.33	0.22	0.39	0.25	0.24	0.25	0.26	0.49	0.23	0.51	0.37
2	0.32	0.18	0.39	0.25	0.24	0.24	0.26	0.52	0.26	0.53	0.36
1	0.32	0.16	0.34	0.23	0.22	0.22	0.22	0.54	0.29	0.56	0.32
	Rand	Attn	Attn-norm	Attn-norm + RES	Attn-norm + RES + LN	GlobEnc	ALTI	GradXinput	IG	DL	Value Zeroing

<https://aclanthology.org/2023.eacl-main.245/>

# Using Controlled Case Studies

- A carefully controlled experimental setup provides strong hypotheses and expectations
- Example: synthetic languages or arithmetic expressions

Syntax	Meaning	<a href="https://arxiv.org/pdf/1905.06401.pdf">https://arxiv.org/pdf/1905.06401.pdf</a>
$E \rightarrow L E_1 O E_2 R$	$[E] = [O]([E_1], [E_2])$	
$E \rightarrow D$	$[E] = [D]$	
$O \rightarrow +$	$[O] = \lambda x, y. x + y \text{ mod } 10$	
$O \rightarrow -$	$[O] = \lambda x, y. x - y \text{ mod } 10$	
$L \rightarrow ($		
$R \rightarrow )$		
$D \rightarrow 0$	$[D] = 0$	
$\vdots$	$\vdots$	
$D \rightarrow 9$	$[D] = 9$	

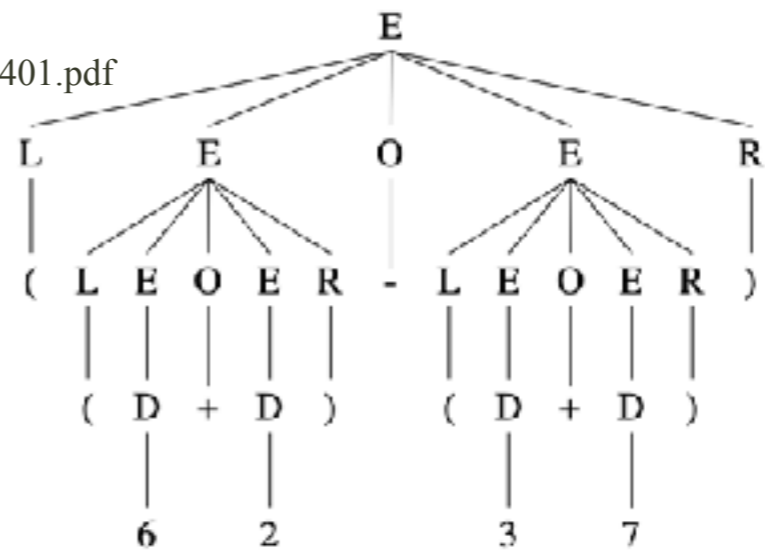


Figure 4: Syntax tree of the expression  $((6+2)-(3+7))$ .

<https://arxiv.org/pdf/1905.06401.pdf>

# Are Interpretations Faithful?

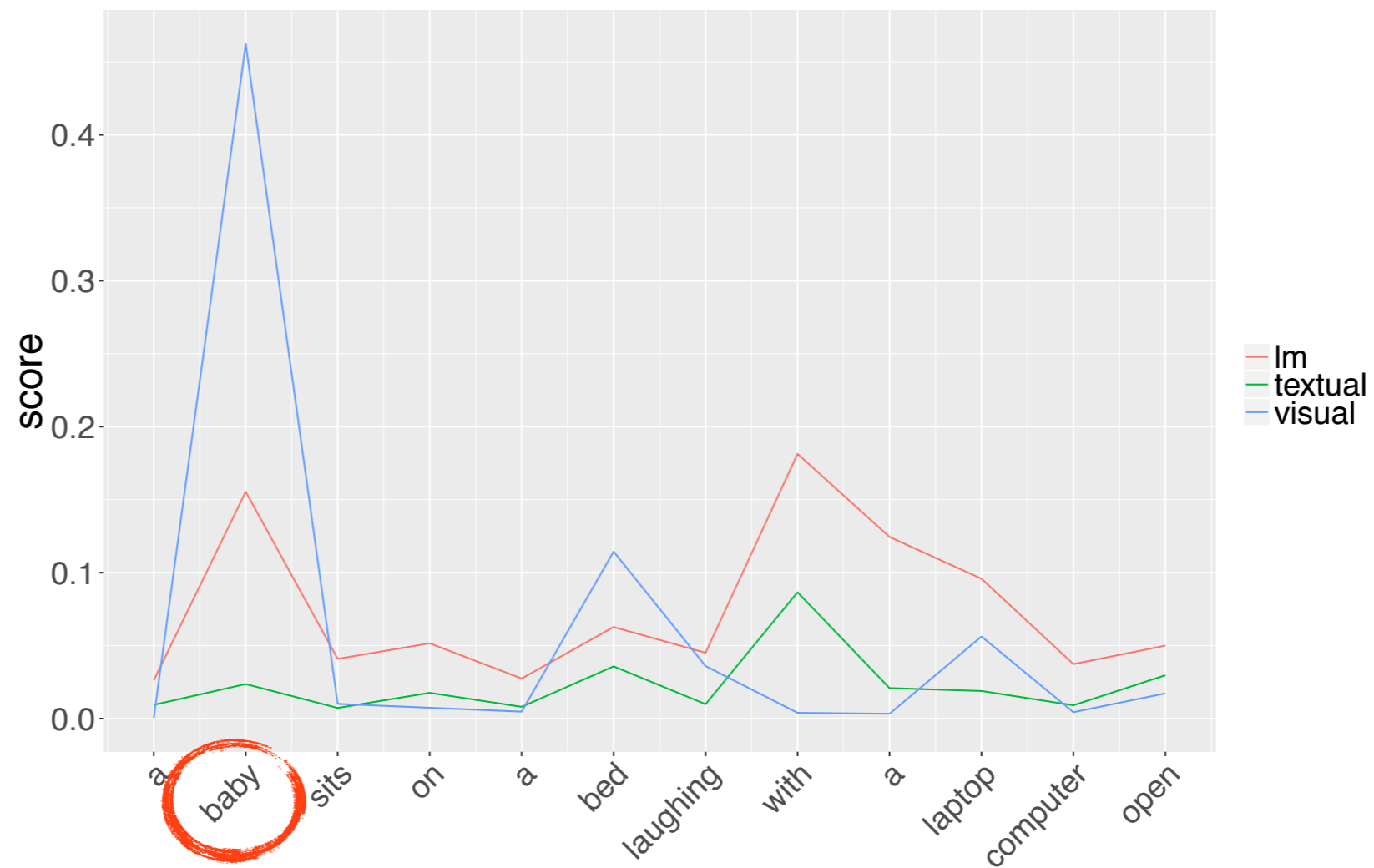
- Faithfulness: an interpretation must accurately describe how the model behaves
  - E.g. removing features hypothesised as salient must hurt model performance
- Plausibility and faithfulness are not always compatible
  - Model might pick up on data artefacts that are effective for the target task but not plausible



A baby sits on a bed laughing with a laptop computer open



$$\text{omission}(i, S) = 1 - \text{cosine}(\mathbf{h}_{\text{end}}(S), \mathbf{h}_{\text{end}}(S_{\setminus i}))$$



<https://direct.mit.edu/coli/article/43/4/761/1583/Representation-of-Linguistic-Form-and-Function-in>



A ~~boy~~ sits on a bed laughing with a laptop computer open

blank-out:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
Attn:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
Attn-norm:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
Attn-norm+RES:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
Attn-norm+RES+LN:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
GlobEnc:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
ALTI:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
GradXinput:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
IG:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
DL:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]
Value Zeroing:	[CLS] the pictures of some hat [MASK] scar ##ing marcus . [SEP]

<https://aclanthology.org/2023.eacl-main.245/>

# **Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?**

**Alon Jacovi**

Bar Ilan University

alonjacovi@gmail.com

**Yoav Goldberg**

Bar Ilan University and Allen Institute for AI

yoav.goldberg@gmail.com

## **Abstract**

With the growing popularity of deep-learning based NLP models, comes a need for interpretable systems. But what is interpretability, and what constitutes a high-quality interpretation? In this opinion piece we reflect on the

One such pain is the challenge of defining—and evaluating—what constitutes a quality interpretation. Current approaches define interpretation in a rather ad-hoc manner, motivated by practical use-cases and applications. However, this view often fails to distinguish between distinct aspects of the interpretation’s quality, such as readability, plausi-

# Towards Faithful Model Explanation in NLP: A Survey

Qing Lyu

University of Pennsylvania

lyuqing@sas.upenn.edu

Marianna Apidianaki

University of Pennsylvania

marapi@seas.upenn.edu

Chris Callison-Burch

University of Pennsylvania

ccb@seas.upenn.edu

*End-to-end neural Natural Language Processing (NLP) models are notoriously difficult to understand. This has given rise to numerous efforts towards model explainability in recent years. One desideratum of model explanation is faithfulness, i.e. an explanation should accurately represent the reasoning process behind the model's prediction. In this survey, we review over 110 model explanation methods in NLP through the lens of faithfulness. We first*

# Interpretability vs. Explainability

- Who is the target audience?
  - Interpretability is aimed at researchers/developers
  - Explainability is aimed at users (e.g. explanation of MT systems for human translators/interpreters)
- Explainability requires human in the loop (e.g. experimental studies of understandability of explanations)



# INDEEP

## INTERPRETING DEEP LEARNING MODELS FOR TEXT AND SOUND



Radboud University



UNIVERSITY  
OF AMSTERDAM

TILBURG



UNIVERSITY



VRIJE  
UNIVERSITEIT  
AMSTERDAM



university of  
 groningen



**Grzegorz Chrupała**



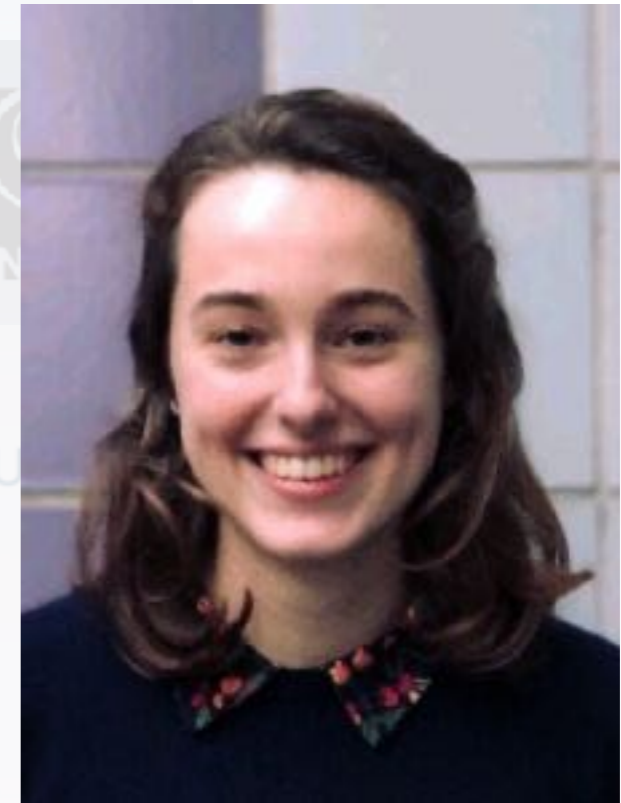
**Willem Zuidema**



**Hosein Mohebbi**



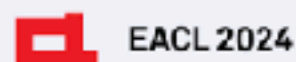
**Gaofei Shen**



**Charlotte Pauw**



# Upcoming Event

[CALLS](#)[COMMITTEES](#)[VENUE](#)[REGISTRATION](#)[PROGRAM](#)[SPONSORS](#)

## PROGRAM

EACL 2024 Program of Events

Workshops

**Tutorials**

Main Conference Accepted Papers

Findings Accepted Papers

## Accepted Tutorials

Tutorial	Date	Time
<a href="#">Transformer-specific Interpretability</a> Hosein Mohebbi, Jaap Jumelet, Michael Hanna, Afra Alishahi and Willem Zuidema	March 21	14:00 - 17:30
<a href="#">LLMs for Low Resource Languages in Multilingual, Multimodal and Dialectal Settings</a> Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel and Maram Hasanain	March 21	14:00 - 17:30
<a href="#">Item Response Theory for Natural Language Processing</a> João Sedoc, John P. Lalor, Pedro Rodriguez and Jose Hernandez-Orallo	March 21	9:00 - 12:30