BlackboxNLP 2022

BlackboxNLP Analyzing and Interpreting Neural Networks for NLP

Proceedings of the Workshop

December 8, 2022

The BlackboxNLP organizers gratefully acknowledge the support from the following sponsors.

Main Sponsors

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-959429-05-0

Introduction

BlackboxNLP is the fifth workshop on analyzing and interpreting neural networks for NLP, hosted by the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022) in Abu Dhabi, United Arab Emirates, and online (hybrid).

Many recent performance improvements in NLP have come at the cost of understanding of the systems. How do we assess what representations and computations models learn? How do we formalize desirable properties of interpretable models, and measure the extent to which existing models achieve them? How can we build models that better encode these properties? What can new or existing tools tell us about the inductive biases of systems?

The goal of this workshop is to bring together researchers focused on interpreting and explaining NLP models by taking inspiration from machine learning, psychology, linguistics, and neuroscience. We hope the workshop will serve as an interdisciplinary meetup that allows for cross-collaboration.

The topics of the workshop include, but are not limited to: Explanation methods such as saliency, attribution, free-text explanations, or explanations with structured properties; Probing methods for testing whether models have acquired or represent certain linguistic properties; Applying analysis techniques from other disciplines (e.g., neuroscience or computer vision); Examining model performance on simplified or formal languages; More interpretable model architectures; Open-source tools for analysis, visualization, or explanation; Evaluation of explanation methods; Opinion pieces about the state of explainable NLP.

We received an impressive number of 76 submissions (including both archival papers and extended abstracts), suggesting that the issue of interpretability of neural networks remains important within the NLP community. The final program contains three keynote talks, four oral presentations and 74 posters (33 archival papers, 13 extended abstracts, and 28 Findings papers). We hope this workshop provides a venue for bringing together ideas and stimulate new ways of building methods and resources for facilitating better analysis and understanding of the inner-dynamics of neural networks for NLP.

BlackboxNLP would not have been possible without the dedication of its program committee. We would like to thank them for their invaluable effort in providing high-quality reviews in a very short period of time. We are also grateful to our invited speakers, Lena Voita, Catherine Olsson and David Bau, for contributing to our program. Finally, we are very thankful to our sponsor, Google, that made it possible for some of our participants to attend the workshop.

Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, Sarah Wiegreffe

Organizing Committee

Organizers

Jasmijn Bastings, Google Research
Yonatan Belinkov, Technion
Yanai Elazar, Allen Institute for AI
Dieuwke Hupkes, Meta AI
Naomi Saphra, NYU
Sarah Wiegreffe, Allen Institute for AI

Program Committee

Reviewers

Heike Adel, Sachin Agarwal, Leila Arras, Pepa Atanasova, Tania Avgustinova

Parsa Bagherzadeh, Vinayshekhar Bannihatti, Jonathan Brophy, Lisa Bylinina

Rahma Chaabouni, Stergios Chatzikyriakidis, Hanjie Chen

Arya D., Verna Dankers, Anubrata Das

Michael Eric

Ian F., Nils Feldhus, Ghazi Felhi, Richard Futrell

Mario Giulianelli, Seraphina Goldfarb-Tarrant

David Harwath, John Hewitt

Alon Jacovi, Carolyn Jane, Robin Jia, Jinhang Jiang, Zhiying Jiang, Richard Johansson, Jaap Jumelet

Sayantan Kumar, Jenny Kunz

Cassandra L, Minghan Li, Yuchen Lian, Sheng Liang, Jindřich Libovický, Tomasz Limisiewicz, Zhiyu Lin

Badr M, Valentin Malykh, Kate McCurdy, Francois Meyer, Janusz Milewski, Koji Mineshima, Marius Mosbach

Anmol Nayak, Joakim Nivre

Xiang Pan, Swetasudha Panda, Bhargavi Paranjape, Lis Pereira, Yuval Pinter, Jacob Portes, Aman Priyanshu

Shauli Ravfogel, Abhilasha Ravichander, Rudolf Rosa, Sara rajaee

Hassan Sajjad, Hendrik Schuff, Rico Sennrich, Mattia Setzu, Tatiana Shavrina, Victor Siemen, Sanchit Sinha, Pia Sommerauer, Shane Steinert-Threlkeld, Vinitra Swamy

Marc Tanti, Jörg Tiedemann

Jithendra Vepa

Eric Wallace, Alex Warstadt, Adina Williams

Lan Xiao, Zhouhang Xie

Dylan Z, Yichu Zhou

Keynote Talk: The Two Viewpoints on the NMT Training Process

Lena Voita
Facebook AI Research

Abstract: In this talk, I illustrate how the same process (in this case, NMT training process) can be viewed from different perspectives: from the inside of the model and from the outside, i.e. in a black-box manner. In the first view, we look at the model's inner workings and try to understand how NMT balances two different types of context, the source and the prefix of the target sentence. In the second view, we look at model outputs (i.e. generated translations) at different steps during training and evaluate how the model acquires different competences. We find that NMT training consists of the stages where it focuses on the competences mirroring three core SMT components: target-side language modeling, lexical translation and reordering. Most importantly, the two views show the same process, and we will see how this process is reflected in these two types of analysis.

Bio: Elena (Lena) Voita is a Research Scientist joining Facebook AI Research. She is mostly interested in understanding what and how neural models learn. Her analysis works so far include looking at model components, adapting attribution methods to NLP models, black-box analysis of model outputs, as well as information-theoretic view on analysis (e.g., probing). Previously, she was a PhD student at the University of Edinburgh supervised by Ivan Titov and Rico Sennrich, was awarded Facebook PhD Fellowship, worked as a Research Scientist at Yandex Research side by side with the Yandex Translate team. She enjoys writing blog posts and teaching; a public version of (a part of) her NLP course is available at lena-voita.github.io/nlp_course.html.

Keynote Talk: In-Context Learning and Induction Heads

Catherine Olsson
Anthropic AI

Abstract: "Induction heads" are attention heads that implement a simple algorithm to complete token sequences like [A][B] ... [A] $\rightarrow$ [B]. In this work, we present preliminary and indirect evidence for a hypothesis that induction heads might constitute the mechanism for the majority of all "in-context learning" in large transformer models (i.e. decreasing loss at increasing token indices). We find that induction heads develop at precisely the same point as a sudden sharp increase in in-context learning ability, visible as a bump in the training loss. We present six complementary lines of evidence, arguing that induction heads may be the mechanistic source of general in-context learning in transformer models of any size. For small attention-only models, we present strong, causal evidence; for larger models with MLPs, we present correlational evidence.

Bio: Catherine Olsson is a research engineer at Anthropic, and the lead author on the recent mechanistic interpretability paper In-context Learning and Induction Heads. She has previously worked in technical research roles at Google Brain and OpenAI, and as a grantmaker at Open Philanthropy Project funding academic research in ML robustness.

Keynote Talk: Direct Model Editing

David Bau
Northeastern Khoury College

Abstract: Can we understand large deep networks well enough to reprogram them by changing their parameters directly? In this talk I will talk about Direct Model Editing: how to modify the weights of a large model directly by understanding its structure. We will consider examples in computer vision and NLP: how to probe and rewrite computations within an image synthesis model to alter compositional rules that govern rendering of realistic images, and how the ROME method can edit specific factual memories within a large language model, directly tracing and modifying parameters that store associations within GPT. I will talk about how causal mediation analysis can serve as a key to unlock the secrets of a huge model; the specificity-generalization trade-off when evaluating knowledge changes in a large model; and how recent results in our MEMIT work suggest that direct editing in huge models may scale orders-of-magnitudes better than traditional opaque fine-tuning.

Bio: David Bau is Assistant Professor at the Northeastern University Khoury College of Computer Science. He received his PhD from MIT and AB from Harvard. He is known for his network dissection studies of individual neurons in deep networks and has published research on the interpretable structure of large models in PNAS, CVPR, NeurIPS, and SIGGRAPH. Prof. Bau is also coauthor of the textbook, Numerical Linear Algebra.

Table of Contents

Program

Thursday, December 8, 2022

08:00 - 09:00 *Poster Session 0 - Virtual*

A Minimal Model for Compositional Generalization on gSCAN
Alice Hein and Klaus Diepold

Sparse Interventions in Language Models with Differentiable Masking
Nicola De Cao, Leon Schmid, Dieuwke Hupkes and Ivan Titov

Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit
Mughilan Muthupari, Samrat Halder, Asad B. Sayeed and Yuval Marton

Sentence Ambiguity, Grammaticality and Complexity Probes
Sunit Bhattacharya, Vilém Zouhar and Ondrej Bojar

Post-Hoc Interpretation of Transformer Hyperparameters with Explainable Boosting Machines
Kiron Deb, Xuan Zhang and Kevin Duh

Revisit Systematic Generalization via Meaningful Learning
Ning Shi, Boxin Wang, Wei Wang, Xiangyu Liu and Zhouhan Lin

Is It Smaller Than a Tennis Ball? Language Models Play the Game of Twenty Questions
Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann and Walter Daelemans

Post-hoc analysis of Arabic transformer models
Ahmed Abdelali, Nadir Durrani, Fahim Dalvi and Hassan Sajjad

Universal Evasion Attacks on Summarization Scoring
Wenchuan Mu and Kwan Hui Lim

How (Un)Faithful is Attention?
Hessam Amini and Leila Kosseim

Are Multilingual Sentiment Models Equally Right for the Right Reasons?
Rasmus Kær Jørgensen, Fiammetta Caccavale, Christian Igel and Anders Søgaard

Probing for Understanding of English Verb Classes and Alternations in Large Pre-trained Language Models
David K Yi, James V. Bruno, Jiayu Han, Peter Zukerman and Shane Steinert-Threlkeld

Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of a NMT System
Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier and François Yvon

Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions
Jenny Kunz, Martin Jirenius, Oskar Holmström and Marco Kuhlmann

Analyzing the Representational Geometry of Acoustic Word Embeddings
Badr M. Abdullah and Dietrich Klakow

Understanding Domain Learning in Language Models Through Subpopulation Analysis
Zheng Zhao, Yftah Ziser and Shay B Cohen

Intermediate Entity-based Sparse Interpretable Representation Learning
Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh and Byron C Wallace

Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information
Isar Nejadgholi, Esma Balkir, Kathleen C. Fraser and Svetlana Kiritchenko

Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?
Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa

It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark
Teemu Vahtola, Mathias Creutz and Jörg Tiedemann

Controlling for Stereotypes in Multimodal Language Model Evaluation
Manuj Malik and Richard Johansson

On the Compositional Generalization Gap of In-Context Learning
Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni and Aaron Courville

Explaining Translationese: why are Neural Classifiers Better and what do they Learn?
Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Van Genabith and Cristina España-Bonet

Probing GPT-3's Linguistic Knowledge on Semantic Tasks
Lining Zhang, Mengchen Wang, Liben Chen and Wenxin Zhang

Garden Path Traversal in GPT-2
William Jurayj, William Rudman and Carsten Eickhof

Testing Pre-trained Language Models' Understanding of Distributivity via Causal Mediation Analysis
Pangbo Ban, Yifan Jiang, Tianran Liu and Shane Steinert-Threlkeld

Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence
Jingcheng Niu, Wenjie Lu, Eric Corlett and Gerald Penn

DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models
Royi Rassin, Shauli Ravfogel and Yoav Goldberg

Practical Benefits of Feature Feedback Under Distribution Shift
Anurag Katakkar, Clay H. Yoo, Weiqin Wang, Zachary Chase Lipton and Divyansh Kaushik

Identifying the Source of Vulnerability in Explanation Discrepancy: A Case Study in Neural Text Classification
Ruixuan Tang, Hanjie Chen and Yangfeng Ji

Probing Pretrained Models of Source Codes
Sergey Troshin and Nadezhda Chirkova

Probing the representations of named entities in Transformer-based Language Models
Stefan Frederik Schouten, Peter Bloem and Piek Vossen

Decomposing Natural Logic Inferences for Neural NLI
Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino and Andre Freitas

Probing with Noise: Unpicking the Warp and Weft of Embeddings
Filip Klubicka and John D. Kelleher

Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering
Kazutoshi Shinoda, Saku Sugawara and Akiko Aizawa

A Continuum of Generation Tasks for Investigating Length Bias and Degenerate Repetition
Darcey Riley and David Chiang

Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation
Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova and Tatiana Shavrina

The Rediscovery Hypothesis: Language Models Need to Meet Linguistics
Vassilina Nikoulina, Maxat Tezekbayev, Nuradil Kozhakhmet, Madina Babazhanova, Matthias Gallé and Zhenisbek Assylbekov

The Solvability of Interpretability Evaluation Metrics
Yilun Zhou and Julie Shah

Discontinuous Constituency and BERT: A Case Study of Dutch
Konstantinos Kogkalidis and Gijs Wijnholds

The BERT Walked Down the Garden Path Assigned Semantic Roles
Tovah Irwin, Kyra Wilson and Alec Marantz

Analyzing Transformers in Embedding Space
Guy Dar, Mor Geva, Ankit Gupta and Jonathan Berant

FIDAM-Eval: A Framework for Evaluating Feature Interaction Detection and Attribution Methods
Jaap Jumelet and Willem H. Zuidema

Faithful, Interpretable Model Explanations via Causal Abstraction
Atticus Geiger, Zhengxuan Wu, Karel D'Oosterlinck, Elisa Kreiss, Noah Goodman, Thomas Icard and Christopher Potts

Behavioral Testing of Knowledge Graph Embedding Models for Link Prediction
Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert and Naoaki Okazaki

Do LSTMs See Gender? Probing the Ability of LSTMs to Learn Abstract Syntactic Rules
Priyanka Sukumaran, Conor Houghton and Nina Kazanina

FRAME: Evaluating Rationale-Label Consistency Metrics for Free-Text Rationales
Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi and Xiang Ren

Tracing and Manipulating Intermediate Results in Neural Math Problem Solvers
Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa and Kentaro Inui

A Critical Look at Fine-tuning Generalization: Are Patterns All We Need?
Marius Mosbach, Shauli Ravfogel, Dietrich Klakow and Yanai Elazar

A Human-Centric Assessment Framework for AI
Sascha Saralajew, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Wiem Ben Rim, Jürgen Quittek and Carolin Lawrence

Do Language Models Understand Measurements?
Edward Choi, Seungwoo Ryu and Sungjin Park

Outlier Dimensions that Disrupt Transformers are Driven by Frequency
Felice Dell'Orletta, Aleksandr Drozd, Anna Rogers and Giovanni Puccetti

On the Impact of Temporal Concept Drift on Model Explanations
Nikolaos Aletras, Kalina Bontcheva, George Chrysostomou and Zhixue Zhao

Lexical Generalization Improves with Larger Models and Longer Training
Yanai Elazar, Yoav Goldberg and Elron Bandel

What Has Been Enhanced in my Knowledge-Enhanced Language Model?
Mrinmaya Sachan, Guoji Fu and Yifan Hou

SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings
Markus Strohmaier, Marlene Lutz, Sandipan Sikdar and Jan Engler

Identifying Human Strategies for Generating Word-Level Adversarial Examples
Lewis Griffin, Bennett Kleinberg and Maximilian Mozes

Transformer Language Models without Positional Encodings Still Learn Positional Information
Omer Levy, Peter Izsak, Ofir Press, Ori Ram and Adi Haviv

Probing Relational Knowledge in Language Models via Word Analogies
Jose Camacho-Collados and Kiamehr Rezaee

Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup
Oana-Maria Camburu, Thomas Lukasiewicz, Vid Kocijan and Yordan Yordanov

The Curious Case of Absolute Position Embeddings
Adina Williams, Dieuwke Hupkes, Joelle Pineau, Siva Reddy, Amirhossein Kazemnejad and Koustuv Sinha

Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining
Siva Reddy, Vaibhav Adlakha, Nicholas Meade and Andreas Madsen

Are Large Pre-Trained Language Models Leaking Your Personal Information?
Kevin Chen-Chuan Chang, Hanyin Shao and Jie Huang

Exploring The Landscape of Distributional Robustness for Question Answering Models
Ludwig Schmidt, Hannaneh Hajishirzi, Ian Magnusson, Sewon Min, Gabriel Ilharco, Mitchell Wortsman and Anas Awadalla

Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning
Sameer Singh, Matt Gardner, Robert L Logan IV and Yasaman Razeghi

What do Large Language Models Learn beyond Language?
Shashank Srivastava and Avinash Madasu

How Much Does Attention Actually Attend? Questioning the Importance of Attention in Pretrained Transformers
Roy Schwartz, Noah A. Smith, Ivan Montero, Jungo Kasai, Daniel Rotem, Hao Peng and Michael Hassid

Probing for Constituency Structure in Neural Language Models
Hassan Sajjad, Laura Kallmeyer, Younes Samih and David Arps

Recursive Neural Networks with Bottlenecks Diagnose (Non-)Compositionality
Ivan Titov and Verna Dankers

CAT-probing: A Metric-based Approach to Interpret How Pre-trained Models for Programming Language Attend Code Structure
Ming Gao, Xuesong Lu, Xiang Li, Renyu Zhu, Qiushi Sun and Nuo Chen

ER-Test: Evaluating Explanation Regularization Methods for Language Models
Xiang Ren, Hamed Firooz, Maziar Sanjabi, Shaoliang Nie, Ziyi Liu, Aaron Chan and Brihi Joshi

Can Language Models Serve as Temporal Knowledge Bases?
Hai Jin, Sixiao Zhang, Guandong Xu, Feng Zhao and Ruilin Zhao

Baked-in State Probing
Kevin Gimpel, Karen Livescu, Sam Wiseman and Shubham Toshniwal

Towards Tracing Knowledge in Language Models Back to the Training Data
Kelvin Guu, Jacob Andreas, Ian Tenney, Binbin Xiong, Frederick Liu, Tolga Bolukbasi and Ekin Akyurek

Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes
Mark Riedl, Sarah Wiegreffe and Kaige Xie

CodeExp: Explanatory Code Document Generation
Nan Duan, Jianfeng Gao, Bo Wang, Todd Mytkowicz, Jeevana Priya Inala, Junjie Huang, Chenglong Wang and Haotian Cui

Thursday, December 8, 2022 (continued)

Influence Functions for Sequence Tagging Models
Ani Nenkova, Byron Wallace, Varun Manjunatha and Sarthak Jain

CORE: A Retrieve-then-Edit Framework for Counterfactual Data Generation
Luke Zettlemoyer, Hannaneh Hajishirzi, Bhargavi Paranjape and Tanay Dixit

09:00 - 09:10 *Opening Remarks*

09:15 - 10:00 *Invited Talk 1 - Lena Voita*

10:00 - 10:30 *Oral Presentations 1 and 2*

Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions
Jenny Kunz, Martin Jirenius, Oskar Holmström and Marco Kuhlmann

Analyzing the Representational Geometry of Acoustic Word Embeddings
Badr M. Abdullah and Dietrich Klakow

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Poster Session 1 - Onsite*

Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit
Mughilan Muthupari, Samrat Halder, Asad B. Sayeed and Yuval Marton

Sentence Ambiguity, Grammaticality and Complexity Probes
Sunit Bhattacharya, Vilém Zouhar and Ondrej Bojar

The Rediscovery Hypothesis: Language Models Need to Meet Linguistics
Vassilina Nikoulina, Maxat Tezekbayev, Nuradil Kozhakhmet, Madina Babazhanova, Matthias Gallé and Zhenisbek Assylbekov

The Solvability of Interpretability Evaluation Metrics
Yilun Zhou and Julie Shah

Analyzing Transformers in Embedding Space
Guy Dar, Mor Geva, Ankit Gupta and Jonathan Berant

FIDAM-Eval: A Framework for Evaluating Feature Interaction Detection and Attribution Methods
Jaap Jumelet and Willem H. Zuidema

Understanding Domain Learning in Language Models Through Subpopulation Analysis
Zheng Zhao, Yftah Ziser and Shay B Cohen

Intermediate Entity-based Sparse Interpretable Representation Learning
Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh and Byron C Wallace

Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?
Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa

It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark
Teemu Vahtola, Mathias Creutz and Jörg Tiedemann

Controlling for Stereotypes in Multimodal Language Model Evaluation
Manuj Malik and Richard Johansson

Behavioral Testing of Knowledge Graph Embedding Models for Link Prediction
Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert and Naoaki Okazaki

Do LSTMs See Gender? Probing the Ability of LSTMs to Learn Abstract Syntactic Rules
Priyanka Sukumaran, Conor Houghton and Nina Kazanina

Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence
Jingcheng Niu, Wenjie Lu, Eric Corlett and Gerald Penn

DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models
Royi Rassin, Shauli Ravfogel and Yoav Goldberg

Probing the representations of named entities in Transformer-based Language Models
Stefan Frederik Schouten, Peter Bloem and Piek Vossen

Decomposing Natural Logic Inferences for Neural NLI
Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino and Andre Freitas

Tracing and Manipulating Intermediate Results in Neural Math Problem Solvers
Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa and Kentaro Inui

Probing with Noise: Unpicking the Warp and Weft of Embeddings
Filip Klubicka and John D. Kelleher

A Critical Look at Fine-tuning Generalization: Are Patterns All We Need?
Marius Mosbach, Shauli Ravfogel, Dietrich Klakow and Yanai Elazar

A Human-Centric Assessment Framework for AI
Sascha Saralajew, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Wiem Ben Rim, Jürgen Quittek and Carolin Lawrence

Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering
Kazutoshi Shinoda, Saku Sugawara and Akiko Aizawa

Do Language Models Understand Measurements?
Edward Choi, Seungwoo Ryu and Sungjin Park

Outlier Dimensions that Disrupt Transformers are Driven by Frequency
Felice Dell'Orletta, Aleksandr Drozd, Anna Rogers and Giovanni Puccetti

On the Impact of Temporal Concept Drift on Model Explanations
Nikolaos Aletras, Kalina Bontcheva, George Chrysostomou and Zhixue Zhao

Lexical Generalization Improves with Larger Models and Longer Training
Yanai Elazar, Yoav Goldberg and Elron Bandel

What Has Been Enhanced in my Knowledge-Enhanced Language Model?
Mrinmaya Sachan, Guoji Fu and Yifan Hou

Identifying Human Strategies for Generating Word-Level Adversarial Examples
Lewis Griffin, Bennett Kleinberg and Maximilian Mozes

Transformer Language Models without Positional Encodings Still Learn Positional Information
Omer Levy, Peter Izsak, Ofir Press, Ori Ram and Adi Haviv

Probing Relational Knowledge in Language Models via Word Analogies
Jose Camacho-Collados and Kiamehr Rezaee

The Curious Case of Absolute Position Embeddings
Adina Williams, Dieuwke Hupkes, Joelle Pineau, Siva Reddy, Amirhossein Kazemnejad and Koustuv Sinha

Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining
Siva Reddy, Vaibhav Adlakha, Nicholas Meade and Andreas Madsen

Exploring The Landscape of Distributional Robustness for Question Answering Models
Ludwig Schmidt, Hannaneh Hajishirzi, Ian Magnusson, Sewon Min, Gabriel Ilharco, Mitchell Wortsman and Anas Awadalla

Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning
Sameer Singh, Matt Gardner, Robert L Logan IV and Yasaman Razeghi

How Much Does Attention Actually Attend? Questioning the Importance of Attention in Pretrained Transformers
Roy Schwartz, Noah A. Smith, Ivan Montero, Jungo Kasai, Daniel Rotem, Hao Peng and Michael Hassid

Probing for Constituency Structure in Neural Language Models
Hassan Sajjad, Laura Kallmeyer, Younes Samih and David Arps

Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes
Mark Riedl, Sarah Wiegreffe and Kaige Xie

Thursday, December 8, 2022 (continued)

CORE: A Retrieve-then-Edit Framework for Counterfactual Data Generation
Luke Zettlemoyer, Hannaneh Hajishirzi, Bhargavi Paranjape and Tanay Dixit

SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings
Markus Strohmaier, Marlene Lutz, Sandipan Sikdar and Jan Engler

12:30 - 14:00 *Lunch Break*

14:00 - 14:45 *Invited Talk 2 - Catherine Olsson*

14:45 - 15:30 *Oral Presentations 3 and 4*

Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?
Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa

Probing with Noise: Unpicking the Warp and Weft of Embeddings
Filip Klubicka and John D. Kelleher

15:30 - 16:00 *Coffee Break*

16:00 - 17:30 *Poster Session 2 - Virtual*

A Minimal Model for Compositional Generalization on gSCAN
Alice Hein and Klaus Diepold

Sparse Interventions in Language Models with Differentiable Masking
Nicola De Cao, Leon Schmid, Dieuwke Hupkes and Ivan Titov

Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit
Mughilan Muthupari, Samrat Halder, Asad B. Sayeed and Yuval Marton

Sentence Ambiguity, Grammaticality and Complexity Probes
Sunit Bhattacharya, Vilém Zouhar and Ondrej Bojar

Post-Hoc Interpretation of Transformer Hyperparameters with Explainable Boosting Machines
Kiron Deb, Xuan Zhang and Kevin Duh

Revisit Systematic Generalization via Meaningful Learning
Ning Shi, Boxin Wang, Wei Wang, Xiangyu Liu and Zhouhan Lin

Is It Smaller Than a Tennis Ball? Language Models Play the Game of Twenty Questions
Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann and Walter Daelemans

Post-hoc analysis of Arabic transformer models
Ahmed Abdelali, Nadir Durrani, Fahim Dalvi and Hassan Sajjad

Universal Evasion Attacks on Summarization Scoring
Wenchuan Mu and Kwan Hui Lim

How (Un)Faithful is Attention?
Hessam Amini and Leila Kosseim

Are Multilingual Sentiment Models Equally Right for the Right Reasons?
Rasmus Kær Jørgensen, Fiammetta Caccavale, Christian Igel and Anders Søgaard

Probing for Understanding of English Verb Classes and Alternations in Large Pre-trained Language Models
David K Yi, James V. Bruno, Jiayu Han, Peter Zukerman and Shane Steinert-Threlkeld

Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of a NMT System
Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier and François Yvon

Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions
Jenny Kunz, Martin Jirenius, Oskar Holmström and Marco Kuhlmann

Analyzing the Representational Geometry of Acoustic Word Embeddings
Badr M. Abdullah and Dietrich Klakow

Understanding Domain Learning in Language Models Through Subpopulation Analysis
Zheng Zhao, Yftah Ziser and Shay B Cohen

Intermediate Entity-based Sparse Interpretable Representation Learning
Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh and Byron C Wallace

Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information
Isar Nejadgholi, Esma Balkir, Kathleen C. Fraser and Svetlana Kiritchenko

Investigating the Characteristics of a Transformer in a Few-Shot Setup: Does Freezing Layers in RoBERTa Help?
Digvijay Anil Ingle, Rishabh Kumar Tripathi, Ayush Kumar, Kevin Patel and Jithendra Vepa

It Is Not Easy To Detect Paraphrases: Analysing Semantic Similarity With Antonyms and Negation Using the New SemAntoNeg Benchmark
Teemu Vahtola, Mathias Creutz and Jörg Tiedemann

Controlling for Stereotypes in Multimodal Language Model Evaluation
Manuj Malik and Richard Johansson

On the Compositional Generalization Gap of In-Context Learning
Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni and Aaron Courville

Explaining Translationese: why are Neural Classifiers Better and what do they Learn?
Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Van Genabith and Cristina España-Bonet

Probing GPT-3's Linguistic Knowledge on Semantic Tasks
Lining Zhang, Mengchen Wang, Liben Chen and Wenxin Zhang

Garden Path Traversal in GPT-2
William Jurayj, William Rudman and Carsten Eickhof

Testing Pre-trained Language Models' Understanding of Distributivity via Causal Mediation Analysis
Pangbo Ban, Yifan Jiang, Tianran Liu and Shane Steinert-Threlkeld

Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence
Jingcheng Niu, Wenjie Lu, Eric Corlett and Gerald Penn

DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models
Royi Rassin, Shauli Ravfogel and Yoav Goldberg

Practical Benefits of Feature Feedback Under Distribution Shift
Anurag Katakkar, Clay H. Yoo, Weiqin Wang, Zachary Chase Lipton and Divyansh Kaushik

Identifying the Source of Vulnerability in Explanation Discrepancy: A Case Study in Neural Text Classification
Ruixuan Tang, Hanjie Chen and Yangfeng Ji

Probing Pretrained Models of Source Codes
Sergey Troshin and Nadezhda Chirkova

Probing the representations of named entities in Transformer-based Language Models
Stefan Frederik Schouten, Peter Bloem and Piek Vossen

Decomposing Natural Logic Inferences for Neural NLI
Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino and Andre Freitas

Probing with Noise: Unpicking the Warp and Weft of Embeddings
Filip Klubicka and John D. Kelleher

Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering
Kazutoshi Shinoda, Saku Sugawara and Akiko Aizawa

A Continuum of Generation Tasks for Investigating Length Bias and Degenerate Repetition
Darcey Riley and David Chiang

Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation
Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova and Tatiana Shavrina

The Rediscovery Hypothesis: Language Models Need to Meet Linguistics
Vassilina Nikoulina, Maxat Tezekbayev, Nuradil Kozhakhmet, Madina Babazhanova, Matthias Gallé and Zhenisbek Assylbekov

The Solvability of Interpretability Evaluation Metrics
Yilun Zhou and Julie Shah

Discontinuous Constituency and BERT: A Case Study of Dutch
Konstantinos Kogkalidis and Gijs Wijnholds

The BERT Walked Down the Garden Path Assigned Semantic Roles
Tovah Irwin, Kyra Wilson and Alec Marantz

Analyzing Transformers in Embedding Space
Guy Dar, Mor Geva, Ankit Gupta and Jonathan Berant

FIDAM-Eval: A Framework for Evaluating Feature Interaction Detection and Attribution Methods
Jaap Jumelet and Willem H. Zuidema

Faithful, Interpretable Model Explanations via Causal Abstraction
Atticus Geiger, Zhengxuan Wu, Karel D'Oosterlinck, Elisa Kreiss, Noah Goodman, Thomas Icard and Christopher Potts

Behavioral Testing of Knowledge Graph Embedding Models for Link Prediction
Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert and Naoaki Okazaki

Do LSTMs See Gender? Probing the Ability of LSTMs to Learn Abstract Syntactic Rules
Priyanka Sukumaran, Conor Houghton and Nina Kazanina

FRAME: Evaluating Rationale-Label Consistency Metrics for Free-Text Rationales
Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi and Xiang Ren

Tracing and Manipulating Intermediate Results in Neural Math Problem Solvers
Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa and Kentaro Inui

A Critical Look at Fine-tuning Generalization: Are Patterns All We Need?
Marius Mosbach, Shauli Ravfogel, Dietrich Klakow and Yanai Elazar

A Human-Centric Assessment Framework for AI
Sascha Saralajew, Ammar Shaker, Zhao Xu, Kiril Gashteovski, Bhushan Kotnis, Wiem Ben Rim, Jürgen Quittek and Carolin Lawrence

Do Language Models Understand Measurements?
Edward Choi, Seungwoo Ryu and Sungjin Park

Outlier Dimensions that Disrupt Transformers are Driven by Frequency
Felice Dell'Orletta, Aleksandr Drozd, Anna Rogers and Giovanni Puccetti

On the Impact of Temporal Concept Drift on Model Explanations
Nikolaos Aletras, Kalina Bontcheva, George Chrysostomou and Zhixue Zhao

Lexical Generalization Improves with Larger Models and Longer Training
Yanai Elazar, Yoav Goldberg and Elron Bandel

What Has Been Enhanced in my Knowledge-Enhanced Language Model?
Mrinmaya Sachan, Guoji Fu and Yifan Hou

SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings
Markus Strohmaier, Marlene Lutz, Sandipan Sikdar and Jan Engler

Identifying Human Strategies for Generating Word-Level Adversarial Examples
Lewis Griffin, Bennett Kleinberg and Maximilian Mozes

Transformer Language Models without Positional Encodings Still Learn Positional Information
Omer Levy, Peter Izsak, Ofir Press, Ori Ram and Adi Haviv

Probing Relational Knowledge in Language Models via Word Analogies
Jose Camacho-Collados and Kiamehr Rezaee

Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup
Oana-Maria Camburu, Thomas Lukasiewicz, Vid Kocijan and Yordan Yordanov

The Curious Case of Absolute Position Embeddings
Adina Williams, Dieuwke Hupkes, Joelle Pineau, Siva Reddy, Amirhossein Kazemnejad and Koustuv Sinha

Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining
Siva Reddy, Vaibhav Adlakha, Nicholas Meade and Andreas Madsen

Are Large Pre-Trained Language Models Leaking Your Personal Information?
Kevin Chen-Chuan Chang, Hanyin Shao and Jie Huang

Exploring The Landscape of Distributional Robustness for Question Answering Models
Ludwig Schmidt, Hannaneh Hajishirzi, Ian Magnusson, Sewon Min, Gabriel Ilharco, Mitchell Wortsman and Anas Awadalla

Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning
Sameer Singh, Matt Gardner, Robert L Logan IV and Yasaman Razeghi

What do Large Language Models Learn beyond Language?
Shashank Srivastava and Avinash Madasu

How Much Does Attention Actually Attend? Questioning the Importance of Attention in Pretrained Transformers
Roy Schwartz, Noah A. Smith, Ivan Montero, Jungo Kasai, Daniel Rotem, Hao Peng and Michael Hassid

Probing for Constituency Structure in Neural Language Models
Hassan Sajjad, Laura Kallmeyer, Younes Samih and David Arps

Recursive Neural Networks with Bottlenecks Diagnose (Non-)Compositionality
Ivan Titov and Verna Dankers

CAT-probing: A Metric-based Approach to Interpret How Pre-trained Models for Programming Language Attend Code Structure
Ming Gao, Xuesong Lu, Xiang Li, Renyu Zhu, Qiushi Sun and Nuo Chen

ER-Test: Evaluating Explanation Regularization Methods for Language Models
Xiang Ren, Hamed Firooz, Maziar Sanjabi, Shaoliang Nie, Ziyi Liu, Aaron Chan and Brihi Joshi

Can Language Models Serve as Temporal Knowledge Bases?
Hai Jin, Sixiao Zhang, Guandong Xu, Feng Zhao and Ruilin Zhao

Baked-in State Probing
Kevin Gimpel, Karen Livescu, Sam Wiseman and Shubham Toshniwal

Towards Tracing Knowledge in Language Models Back to the Training Data
Kelvin Guu, Jacob Andreas, Ian Tenney, Binbin Xiong, Frederick Liu, Tolga Bolukbasi and Ekin Akyurek

Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes
Mark Riedl, Sarah Wiegreffe and Kaige Xie

CodeExp: Explanatory Code Document Generation
Nan Duan, Jianfeng Gao, Bo Wang, Todd Mytkowicz, Jeevana Priya Inala, Junjie Huang, Chenglong Wang and Haotian Cui

Influence Functions for Sequence Tagging Models
Ani Nenkova, Byron Wallace, Varun Manjunatha and Sarthak Jain

CORE: A Retrieve-then-Edit Framework for Counterfactual Data Generation
Luke Zettlemoyer, Hannaneh Hajishirzi, Bhargavi Paranjape and Tanay Dixit

17:30 - 17:45 *Mini Break*

17:45 - 18:45 *Invited Talk 3 - David Bau*

18:45 - 19:00 *Closing Remarks*