BlackboxNLP 2021

Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP

November 11, 2021

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 978-1-955917-06-3

Introduction

BlackboxNLP is the workshop on analyzing and interpreting neural networks for NLP.

In the last few years, neural networks have rapidly become a central component in NLP systems. The improvement in accuracy and performance brought by the introduction of neural networks has typically come at the cost of our understanding of the system: How do we assess what the representations and computations are that the network learns? The goal of this workshop is to bring together people who are attempting to peek inside the neural network black box, taking inspiration from machine learning, psychology, linguistics, and neuroscience.

In this fourth edition of the workshop, hosted by the 2021 conference on Empirical Methods in Natural Language Processing (EMNLP), we accepted 41 archival papers and eight extended abstracts. The workshop also provided a platform for authors of four Findings of ACL papers, and seven Findings of EMNLP papers, to present their work as a poster at the workshop. In addition, for the first time, the workshop worked in collaboration with the ACL Rolling Review system (ARR), accepting two more archival papers through that avenue.

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 timely and high-quality reviews on a short notice. We are also grateful to our invited speakers for contributing to our program.

Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter and Hassan Sajjad

Organizers:

Jasmijn Bastings, Google Amsterdam
Yonatan Belinkov, Technion – Israel Institute of Technology
Emmanuel Dupoux, Ecole Normale Supérieure and INRIA
Mario Giulianelli, University of Amsterdam
Dieuwke Hupkes, Facebook AI Research
Yuval Pinter, Ben-Gurion University of the Negev
Hassan Sajjad, Qatar Computing Research Institute

Program Committee:

Badr Abdullah – Saarland University
Heike Adel – Bosch Center for Artificial Intelligence
Sachin Agarwal – Apple
Željko Agić – Unity Technologies
Leila Arras – Fraunhofer Heinrich Hertz Institute
Kalyani Asthana – University of California Irvine
Pepa Atanasova – University of Copenhagen
Tania Avgustinova – Saarland University
Parsa Bagherzadeh – Concordia University
Niranjan Balasubramanian – Stony Brook University
Vinayshekhar Bannihatti Kumar – Amazon Web Services
Laurent Besacier – LIG
Nikolay Bogoychev – University of Edinburgh
Gemma Boleda – Universitat Pompeu Fabra / ICREA
Jonathan Brophy – University of Oregon
Lisa Bylinina – Bookarang
Stergios Chatzikyriakidis – University of Gothenburg
Hanjie Chen – University of Virginia
Ting-Rui Chiang – Carnegie Mellon University
Marta R. Costa-jussà – Universitat Politècnica de Catalunya
Benoit Crabbé – University of Paris
Verna Dankers – University of Edinburgh
Debajyoti Datta – University of Virginia
Roberto Dessì – Facebook AI Research Paris / UPF
Anne Dirkson – Leiden University
Radina Dobreva – University of Edinburgh
Yanai Elazar – Bar-Ilan University
Allyson Ettinger – University of Chicago
Jean-Philippe – Fauconnier Apple Inc.
Ghazi Felhi – Université Paris Nord
Antske Fokkens – VU Amsterdam
Robert Frank – Yale University
Richard Futrell – University of California, Irvine
Sebastian Gehrmann – Google Research
Michael Goodale – Ecole Normale Superieure, Department of Cognitive Studies
Loïc Grobol – École Normale Supérieure

Prachi Gupta – New York University Tandon School of Engineering
David Harwath – The University of Texas at Austin
Peter Hase – University of North Carolina at Chapel Hill
Simon Hegelich – TUM
John Hewitt – Stanford University
Dewei Hu – University of Copenhagen
Cassandra L. Jacobs – University of Wisconsin, Madison
Alon Jacovi – Bar Ilan University
Robin Jia – Facebook AI Research
Zhiying Jiang – University of Waterloo
Richard Johansson – University of Gothenburg
Jaap Jumelet – University of Amsterdam
Lis Kanashiro Pereira – Ochanomizu University
Eugene Kharitonov – Facebook AI
Germán Kruszewski – Naver Labs Europe
Kunal Kukreja – University of Washington
Jenny Kunz – Linköping University
Shalom Lappin – King's College London
Alessandro Lenci – University of Pisa
Minghan Li – University of Waterloo
Sheng Liang – Center for Information and Language Processing, LMU Munich
Jindřich Libovický – Ludwig Maximilian University of Munich
Tomasz Limisiewicz – Charles University in Prague
Nelson F. Liu – Stanford University
David Mareček – Charles University
R. Thomas McCoy – Johns Hopkins University
Houman Mehrafarin – Iran University of Science and Technology
Koji Mineshima – Keio University
Saliha Muradoglu – The Australian National University
Anmol Nayak – ARiSE Labs at Bosch
Joakim Nivre – Uppsala University
Bhargavi Paranjape – University of Washington
Siddharth Patwardhan – Apple
Paolo Pedinotti – University of Pisa
Frithjof Petrick – RWTH Aachen University
Karl Pichotta – Memorial Sloan Kettering Cancer Center
Aman Priyanshu – Manipal Institute of Technology
Sara Rajaee – Iran University of Science and Technology
Abhilasha Ravichander – Carnegie Mellon University
Rudolf Rosa – Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Jan Rosendahl – RWTH Aachen University
Swarnadeep Saha – University of North Carolina at Chapel Hill
Wojciech Samek – Fraunhofer Heinrich Hertz Institute
Hendrik Schuff – Bosch Center for Artificial Intelligence
Robert Schwarzenberg – German Research Center For Artificial Intelligence (DFKI)
Djamé Seddah – Inria
Rico Sennrich – University of Zurich
Mattia Setzu – University of Pisa
Tatiana Shavrina – Sberbank, Sberdevices, HSE University
Sanchit Sinha – University of Virginia

Pia Sommerauer – Vrije Universiteit Amsterdam
Vinitra Swamy – EPFL
Aarne Talman – University of Helsinki
Arka Talukdar – New York University
shimin tao – huawei
Jörg Tiedemann – University of Helsinki
Ivan Titov – University of Edinburgh / University of Amsterdam
Jithendra Vepa – Observe AI
Sarah Wiegreffe – Georgia Institute of Technology
Adina Williams – Facebook, Inc.
Hao Yang – Huawei Co. Ltd
Naoki Yoshinaga – Institute of Industrial Science, the University of Tokyo
Fabio Massimo Zanzotto – University of Rome Tor Vergata
Sheng Zhang – North Carolina State University
Yichu Zhou – University of Utah

Invited Speakers:

Sara Hooker, Google Brain
Ana Marasović, Allen AI
Willem Zuidema, University of Amsterdam

Table of Contents

Conference Program

Thursday, November 11, 2021

Virtual Programme

2:00–2:15 ***Opening Remarks***

2:15–3:00 *Invited Talk by Jelle Zuidema & Q&A*

3:15–4:00 *Oral Session 1*

4:30–6:00 *Poster Session 1*

6:15–7:00 *Oral Session 2*

7:30–8:00 *Invited Talk by Ana Marasović*

8:00–8:30 *Invited Talk by Sara Hooker*

8:30–8:45 ***Closing***

Thursday, November 11, 2021 (continued)

Hybrid Programme

9:00–9:15 ***Opening Remarks & Best Paper Award***

9:15–10:00 *Invited Talk by Jelle Zuidema & Q&A*

10:00–10:30 *Oral Session 3*

11:00–12:00 *Poster Session 2*

13:00–13:45 *Invited Talk by Sara Hooker & Q&A*

13:45–14:15 *Oral Session 4*

14:45–16:15 *Poster Session 3*

16:45–17:15 *Oral Session 5*

17:15–18:00 *Invited Talk by Ana Marasović & Q&A*

18:00–18:15 ***Closing***

Oral Sessions

To what extent do human explanations of model behavior align with actual model behavior?
Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela and Adina Williams

Test Harder than You Train: Probing with Extrapolation Splits
Jenny Kunz and Marco Kuhlmann

Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Ratings
Hendrik Schuff, Hsiu-Yu Yang, Heike Adel and Ngoc Thang Vu

The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty Through Generation
Laura Aina and Tal Linzen

On the Limits of Minimal Pairs in Contrastive Evaluation
Jannis Vamvas and Rico Sennrich

What Models Know About Their Attackers: Deriving Attacker Information From Latent Representations
Zhouhang Xie, Jonathan Brophy, Adam Noack, Wencong You, Kalyani Asthana, Carter Perkins, Sabrina Reis, Zayd Hammoudeh, Daniel Lowd and Sameer Singh

Poster Sessions

ALL Dolphins Are Intelligent and SOME Are Friendly: Probing BERT for Nouns' Semantic Properties and their Prototypicality
Marianna Apidianaki and Aina Garí Soler

ProSPer: Probing Human and Neural Network Language Model Understanding of Spatial Perspective
Tessa Masis and Carolyn Anderson

Can Transformers Jump Around Right in Natural Language? Assessing Performance Transfer from SCAN
Rahma Chaabouni, Roberto Dessì and Eugene Kharitonov

Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?
Tobias Norlund, Lovisa Hagström and Richard Johansson

Discrete representations in neural models of spoken language
Bertrand Higy, Lieke Gelderloos, Afra Alishahi and Grzegorz Chrupała

Word Equations: Inherently Interpretable Sparse Word Embeddings through Sparse Coding
Adly Templeton

A howling success or a working sea? Testing what BERT knows about metaphors
Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini and Alessandro Lenci

How Length Prediction Influence the Performance of Non-Autoregressive Translation?
Minghan Wang, GUO Jiaxin, Yuxia Wang, Yimeng Chen, Su Chang, Hengchao Shang, Min Zhang, Shimin Tao and Hao Yang

On the Language-specificity of Multilingual BERT and the Impact of Fine-tuning
Marc Tanti, Lonneke van der Plas, Claudia Borg and Albert Gatt

Relating Neural Text Degeneration to Exposure Bias
Ting-Rui Chiang and Yun-Nung Chen

Efficient Explanations from Empirical Explainers
Robert Schwarzenberg, Nils Feldhus and Sebastian Möller

Variation and generality in encoding of syntactic anomaly information in sentence embeddings
Qinxuan Wu and Allyson Ettinger

Enhancing Interpretable Clauses Semantically using Pretrained Word Representation
Rohan Kumar Yadav, Lei Jiao, Ole-Christoffer Granmo and Morten Goodwin

Analyzing BERT's Knowledge of Hypernymy via Prompting
Michael Hanna and David Mareček

An in-depth look at Euclidean disk embeddings for structure preserving parsing
Federico Fancellu, Lan Xiao, Allan Jepson and Afsaneh Fazly

Training Dynamic based data filtering may not work for NLP datasets
Arka Talukdar, Monika Dagar, Prachi Gupta and Varun Menon

Interacting Knowledge Sources, Inspection and Analysis: Case-studies on Biomedical text processing
Parsa Bagherzadeh and Sabine Bergler

Attacks against Ranking Algorithms with Text Embeddings: A Case Study on Recruitment Algorithms
Anahita Samadi, debapriya banerjee and Shirin Nilizadeh

Controlled tasks for model analysis: Retrieving discrete information from sequences
Ionut-Teodor Sorodoc, Gemma Boleda and Marco Baroni

The Acceptability Delta Criterion: Testing Knowledge of Language using the Gradience of Sentence Acceptability
Héctor Vázquez Martínez

How Does BERT Rerank Passages? An Attribution Analysis with Information Bottlenecks
Zhiying Jiang, Raphael Tang, Ji Xin and Jimmy Lin

Do Language Models Know the Way to Rome?
Bastien Liétard, Mostafa Abdou and Anders Søgaard

Exploratory Model Analysis Using Data-Driven Neuron Representations
Daisuke Oba, Naoki Yoshinaga and Masashi Toyoda

Fine-Tuned Transformers Show Clusters of Similar Representations Across Layers
Jason Phang, Haokun Liu and Samuel R. Bowman

BERT Has Uncommon Sense: Similarity Ranking for Word Sense BERTology
Luke Gessler and Nathan Schneider

Non-archival papers (posters)

Language Models Use Monotonicity to Assess NPI Licensing – Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes and Shane Steinert-Threlkeld

BPE affects Training Data Memorization by Transformer Language Models – Eugene Kharitonov, Marco Baroni and Dieuwke Hupkes

Do contextual language embeddings distinguish between intersective and strictly subsective adjectives? – Michael Goodale and Salvador Mascarenhas

Generalization in neural sequence models: a case study in symbolic mathematics – Sean Welleck, Peter West, Jize Cao and Yejin Choi

Human Evaluation Study for Explaining Knowledge Graph Completion – Timo Sztyler and Carolin Lawrence

Transformers Scan both Left and Right – When they Have a Cue – Jan H. Athmer and Denis Paperno

Explaining NLP Models via Minimal Contrastive Editing (MiCE) – Alexis Ross, Ana Marasović and Matthew Peters

Probing structures in the visual region embeddings from multimodal BERT – Victor Milewski, Miryam de Lhoneux and Marie-Francine Moens

Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords – Taelin Karidi, Yichu Zhou, Nathan Schneider, Omri Abend and Vivek Srikumar

On Neurons Invariant to Sentence Structural Changes in Neural Machine Translation – Gal Patel, Leshem Choshen and Omri Abend

Explaining Classes through Word Attributions – Samuel Rönnqvist, Amanda Myntti, Aki-Juhani Kyröläinen, Sampo Pyysalo, Veronika Laippala and Filip Ginter

Testing the linguistics of transformer generalizations – Saliha Muradoglu and Mans Hulden