ACL 2019

The BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP at ACL 2019

Proceedings of the Second Workshop

August 1, 2019
Florence, Italy

Sponsored by:

Google facebook Microsoft

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

978-1-950737-30-7

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 second edition of the workshop, hosted by the 2019 Annual Meeting of the Association of Computational Linguistics in Florence, Italy, we accepted 29 archival papers and 16 extended abstracts. We hope this workshop continues to bring together ideas and stimulating new ways of building methods and resources for the 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 timely and high-quality reviews on a short notice. We are also grateful to our invited speakers for contributing to our program. Finally, we are very thankful to our sponsors, Google, Facebook and Mircrosoft for supporting the workshop.

Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov and Dieuwke Hupkes

Organizers:

Tal Linzen, Johns Hopkins University
Grzegorz Chrupała, Tilburg University
Yonatan Belinkov, Harvard University and MIT
Dieuwke Hupkes, ILLC, University of Amsterdam

Program Committee:

Samira Abnar
Željko Agić
Afra Alishahi
Antonios Anastasopoulos
Niranjan Balasubramanian
Joost Bastings
Lisa Beinborn
Laurent Besacier
Or Biran
Samuel R. Bowman
Stergios Chatzikyriakidis
Miryam de Lhoneux
Ewan Dunbar
Jacob Eisenstein
Allyson Ettinger
Antske Fokkens
Robert Frank
Richard Futrell
Sharon Goldwater
Kristina Gulordava
David Harwath
Germán Kruszewski
Yair Lakretz
Shalom Lappin
Jindřich Libovický
Nelson F. Liu
Pranava Madhyastha
David Mareček
Paola Merlo
Raymond Mooney
Sebastian Padó
Yves Peirsman
Adam Poliak
Rudolf Rosa
Carolyn Rose
Hassan Sajjad
Wojciech Samek
Naomi Saphra
Rico Sennrich

Pia Sommerauer
György Szaszák
Francesca Toni
Adina Williams
Roberto Zamparelli
Fabio Massimo Zanzotto
Willem Zuidema

Table of Contents

Conference Program

August 1

9:00–9:10 **Opening remarks**

9:15–10:00 **Keynote speaker 1: Arianna Bisazza**

10:00–11:15 **Poster session 1**

Transcoding Compositionally: Using Attention to Find More Generalizable Solutions
Kris Korrel, Dieuwke Hupkes, Verna Dankers and Elia Bruni

Sentiment Analysis Is Not Solved! Assessing and Probing Sentiment Classification
Jeremy Barnes, Lilja Øvrelid and Erik Velldal

Second-order Co-occurrence Sensitivity of Skip-Gram with Negative Sampling
Dominik Schlechtweg, Cennet Oguz and Sabine Schulte im Walde

Can Neural Networks Understand Monotonicity Reasoning?
Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze and Johan Bos

Multi-Granular Text Encoding for Self-Explaining Categorization
Zhiguo Wang, Yue Zhang, Mo Yu, Wei Zhang, Lin Pan, Linfeng Song, Kun Xu and Yousef El-Kurdi

The Meaning of "Most" for Visual Question Answering Models
Alexander Kuhnle and Ann Copestake

Do Human Rationales Improve Machine Explanations?
Julia Strout, Ye Zhang and Raymond Mooney

Analyzing the Structure of Attention in a Transformer Language Model
Jesse Vig and Yonatan Belinkov

Detecting Political Bias in News Articles Using Headline Attention
Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi and Radhika Mamidi

Testing the Generalization Power of Neural Network Models across NLI Benchmarks
Aarne Talman and Stergios Chatzikyriakidis

(Un)natural Word-order Biases in Deep Agent Architectures
Rahma Chaabouni, Evgeny Kharitonov, Emmanuel Dupoux and Marco Baroni

Why Does a CNN Predict This Class? Interpreting Convolutional Neural Networks for Text Classification
Piyawat Lertvittayakumjorn and Francesca Toni

State-Regularized Recurrent Neural Networks
Cheng Wang and Mathias Niepert

Additional Evidences That BERT Learn Syntactic Structures
Ganesh Jawahar, Benoît Sagot and Djamé Seddah

Neural Networks as Explicit Word-Based Rules
Jindřich Libovický

Neural Pathways: A Method of Abstraction for Deep Neural Model Inspection and Comparison
James Fiacco, Samridhi Choudhary and Carolyn Rose

Can We Explain Natural Language Inference Decisions Taken with Neural Networks? Inference Rules in Distributed Representations
Fabio Massimo Zanzotto and Lorenzo Ferrone

Visualizing Deep Neural Networks for Speech Recognition with Learned Topographic Filter Maps
Andreas Krug and Sebastian Stober

10:30–11:00 ***Tea and coffee break***

August 1 (continued)

11:15–12:30 **Oral presentations 1 (5 x 15 minutes)**

Character Eyes: Seeing Language through Character-Level Taggers
Yuval Pinter, Marc Marone and Jacob Eisenstein

Faithful Multimodal Explanation for Visual Question Answering
Jialin Wu and Raymond Mooney

Evaluating Recurrent Neural Network Explanations
Leila Arras, Ahmed Osman, Klaus-Robert Müller and Wojciech Samek

On the Realization of Compositionality in Neural Networks
Joris Baan, Jana Leible, Mitja Nikolaus, David Rau, Dennis Ulmer, Tim Baumgärtner, Dieuwke Hupkes and Elia Bruni

Learning the Dyck Language with Attention-based Seq2Seq Models
Xiang Yu, Ngoc Thang Vu and Jonas Kuhn

12:30–14:00 ***Lunch***

14:00–14:50 **Keynote speaker 2: Michael F. Bonner**

14:50–16:00 **Poster session 2**

Modeling Paths for Explainable Knowledge Base Completion
Josua Stadelmaier and Sebastian Padó

Probing Word and Sentence Embeddings for Long-distance Dependencies Effects in French and English
Paola Merlo

Derivational Morphological Relations in Word Embeddings
Tomáš Musil, Jonáš Vidra and David Mareček

August 1 (continued)

August 1 (continued)

Learning to Make Accuracy Judgments from Intermediate and Final Outputs of a Neural Network for a Question Answering Task
Chad DeChant, Seungwook Han and Hod Lipson

Some Linguistic Correlates of Gradients and Attention Weights in BERT
Matthijs Westera

Towards Understanding Position Embeddings
Rishi Bommasani and Claire Cardie

On the Importance of Delexicalization for Fact Verification
Sandeep Suntwal, Mithun Paul, Rebecca Sharp and Mihai Surdeanu

15:30–16:00 ***Coffee, tea and snacks***

16:00–16:45 **Oral presentations 2 (3 x 15 minutes)**

GEval: Tool for Debugging NLP Datasets and Models
Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski and Tomasz Górecki

From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions
David Mareček and Rudolf Rosa

What Does BERT Look at? An Analysis of BERT's Attention
Kevin Clark, Urvashi Khandelwal, Omer Levy and Christopher D. Manning

August 1 (continued)

16:45–17:30 Panel discussion

17:30–17:40 Best paper award and closing remarks