Chapman & Hall/CRC Machine Learning & Pattern Recognition

TRANSFORMERS FOR MACHINE LEARNING

A Deep Dive

Uday Kamath
Kenneth L. Graham
Wael Emara

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Transformers for Machine Learning

Chapman & Hall/CRC Machine Learning & Pattern Recognition

A First Course in Machine Learning
Simon Rogers, Mark Girolami

Statistical Reinforcement Learning: Modern Machine Learning Approaches
Masashi Sugiyama

Sparse Modeling: Theory, Algorithms, and Applications
Irina Rish, Genady Grabarnik

Computational Trust Models and Machine Learning
Xin Liu, Anwitaman Datta, Ee-Peng Lim

Regularization, Optimization, Kernels, and Support Vector Machines
Johan A.K. Suykens, Marco Signoretto, Andreas Argyriou

Machine Learning: An Algorithmic Perspective, Second Edition
Stephen Marsland

Bayesian Programming
Pierre Bessiere, Emmanuel Mazer, Juan Manuel Ahuactzin, Kamel Mekhnacha

Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data
Haiping Lu, Konstantinos N. Plataniotis, Anastasios Venetsanopoulos

Data Science and Machine Learning: Mathematical and Statistical Methods
Dirk P. Kroese, Zdravko Botev, Thomas Taimre, Radislav Vaisman

Deep Learning and Linguistic Representation
Shalom Lappin

Artificial Intelligence and Causal Inference
Momiao Xiong

Introduction to Machine Learning with Applications in Information Security, Second Edition
Mark Stamp

Entropy Randomization in Machine Learning
Yuri S. Popkov, Alexey Yu. Popkov, Yuri A. Dubno

Transformers for Machine Learning: A Deep Dive
Uday Kamath, Kenneth L. Graham, and Wael Emara

For more information on this series please visit: https://www.routledge.com/Chapman--Hall CRC-Machine-Learning--Pattern-Recognition/book-series/CRCMACLEAPAT

Transformers for Machine Learning

A Deep Dive

Uday Kamath
Kenneth L. Graham
Wael Emara

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2022
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data

Names: Kamath, Uday, author.
Title: Transformers for machine learning : a deep dive / Uday Kamath, Kenneth L. Graham, Wael Emara.
Description: First edition. | Boca Raton : CRC Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021059529 | ISBN 9780367771652 (hardback) | ISBN 9780367767341 (paperback) | ISBN 9781003170082 (ebook)
Subjects: LCSH: Neural networks (Computer science). | Computational intelligence. | Machine learning.
Classification: LCC QA76.87 .K354 2022 | DDC 006.3/2--dc23/eng/20220218
LC record available at https://lccn.loc.gov/2021059529

ISBN: 978-0-367-77165-2 (hbk)
ISBN: 978-0-367-76734-1 (pbk)
ISBN: 978-1-003-17008-2 (ebk)

DOI: 10.1201/9781003170082

Typeset in Latin Modern font
by KnowledgeWorks Global Ltd.

Publisher's note: This book has been prepared from camera-ready copy provided by the authors.

To all the researchers and frontline COVID workers for their extraordinary service.
– Uday Kamath, Kenneth L. Graham, and Wael Emara

To my parents Krishna and Bharathi, my wife Pratibha, the kids Aaroh and Brandy, my family and friends for their support.
–Uday Kamath

To my wife Alyson, to my mother, my in-laws, my family and friends, thank you for the support and your willingness to sacrifice your time with me.
–Kenneth L. Graham

To my wife Noha, my parents Ali and Zainab, my sister Wesam, my extended family and friends, thank you all for being there for me all the time.
–Wael Emara

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Foreword

Renowned AI pioneer and Nobel laureate Herbert Simon underscored "attention" as the most valuable resource of the information economy, as necessary to allocate attention efficiently among the overabundance of information resources. Having written the foundational paper on meaning-aware AI and recently having served as MIT-Princeton-USAF-AFRL AI Faculty-SME, I had the privilege of publishing by invitation in the same journal's special issue of ASQ, and of being the Malcolm Baldrige National Quality Award administrator, as well as being ranked along with Dr. Simon in the same global academic citation impact studies.

Given the above background, I am thrilled to share with you the most thorough and up-to-date compendium of research, practices, case studies, and applications available today that can provide the best ROI on the latest AI technological advances on transformers inspired by the paper, "Attention is All You Need." Since Google introduced transformer architecture in 2017, transformers have provided exponential improvements in context-focused realization toward meaning-aware AI as deep (neural network) learning models based upon attention mechanisms such as dot-product attention and multi-head attention. Resulting advances in enhanced parallel processing of sequential data have made efficient context sensitive and hence more "meaningful" for ever-larger datasets and much more feasible than earlier.

Covering the latest advances in neural network architectures related to transformers spanning applications such as Natural Language Processing (NLP), speech recognition, time series analysis, and computer vision and domain-specific models spanning science, medicine, and finance, the book aims to meet the theoretical, research, application, and practical needs across academia and industry for multiple audiences including postgraduate students and researchers, undergraduate students, industry practitioners, and professionals. The book rounds off its theory-driven applied and practical coverage with hands-on case studies with

focus on AI explainability, an increasingly important theme in practice imposed by greater focus on issues such as ethical AI and trustable AI.

– Dr. Yogesh Malhotra
Founding Chairman and CEO
U.S. Venture Capital and Private Equity Firm
Global Risk Management Network LLC
scientist
www.yogeshmalhotra.com

Preface

WHY THIS BOOK?

Since 2012 deep learning architectures have started to dominate the machine learning field. However, most of the breakthroughs were in computer vision applications. The main driver of that success was convolutional neural network (CNN) architecture. The efficiency and parallelization of CNN have allowed computer vision architectures to pre-train on enormous data which proved to be a key factor in their success. For years afterward natural language processing (NLP) applications did not see much impact from the new deep learning revolution. Traditional sequence modeling architectures, such as recurrent neural networks (RNNs) and long short-term memory (LSTM), have been used for NLP applications. The sequential nature of such architectures has limited the possibilities to train on the same scale of data that showed value for computer vision.

In 2017 Google introduced the transformer architecture to process sequential data with much more parallelization. Such architecture allowed efficient training on much larger datasets than was possible before. This allowed transformers to revolutionize the NLP field the same way CNN had to computer vision.

Transformers are now becoming the core part of many neural architectures employed in a wide range of applications such as NLP, speech recognition, time series, and computer vision. OpenAI uses transformers in their GPT2/GPT3, which has state-of-the-art performances levels in various NLP tasks. DeepMind's AlphaStar program, which defeated a top professional Starcraft player, also uses transformer architecture. Transformers have gone through many adaptations and alterations, resulting in newer techniques and methods. There is no single book that captures the basics and various changes to the transformers in one place.

This book acts as a unique resource for providing data scientists and researchers (academic and industry) with

- A comprehensive reference book for detailed explanations for every algorithm and technique related to transformers.
- Over 60 transformer architectures covered in a comprehensive manner.
- A book for understanding how to apply the transformer techniques in different NLP applications, speech, time series, and computer vision.
- Practical tips and tricks for each architecture and how to use it in the real world.
- Hands-on case studies providing practical insights to real-world scenarios in diverse topics such as machine translation, topic mining, zero-shot multilingual classification, sentiment analysis, automatic speech recognition, and text classification/categorization are covered in sufficient detail from the task, process, and analysis perspective, all ready to run in Google Colab.

WHO IS THIS BOOK WRITTEN FOR?

The theoretical explanations of the state-of-the-art transformer architectures will appeal to postgraduate students and researchers (academic and industry) as it will provide a single-entry point with deep discussions of a quickly moving field. The practical hands-on case studies and code will appeal to undergraduate students, practitioners, and professionals as it allows for quick experimentation and lowers the barrier to entry into the field.

Transformers are already a cornerstone for NLP deep learning architectures. They are also rapidly employed in other applications such as computer vision and audio. Any course on neural networks, deep learning, or artificial intelligence must delve into discussing transformers as a key state-of-the-art architecture. The book can act as a reference for readers, to brush up on specific pieces of their understanding, or as a way to explore the uses of the transformer for specific challenges. We aim for the book to be a resource to refer back to multiple times, to gain insight and use as readers are faced with different challenges or when lacking understanding.

WHAT THIS BOOK COVERS

This book takes an in-depth approach to presenting the fundamentals of transformers through mathematical theory and practical use cases.

A brief description of each chapter is given below.

1. Chapter 1 will introduce readers to transformers from the timeline, history, and its impact on the academic and industrial world. We will then lay out a complete roadmap based on the taxonomy and how each chapter renders from the theory, practice, and application perspective. The chapter then proceeds with a comprehensive discussion on practical aspects such as resources, tools, books, and courses that will be employed in other chapters.

2. Chapter 2 starts by introducing the sequence-to-sequence models and their limitations. The chapter then lays out various building blocks of transformers such as attention, multi-headed attention, positional encodings, residual connections, and encoder-decoder frameworks in a step-by-step manner. All these functional units get detailed treatment from a theoretical and practical perspectives for the readers to get a complete handle on the topic. Finally, a real-world case study using transformers for machine translation tasks showing the operative aspects concludes the chapter.

3. The advent of BERT has revolutionized the field of natural language processing (NLP) and helped to get close to human-level performance in many conventionally challenging tasks. Chapter 3 introduces the details of the BERT architecture and how it is pre-trained and fine-tuned for classical NLP tasks such as single/pair text classification, token tagging, and question answering. The chapter also discusses the field of BERTology, which is research related to the inner workings of BERT and how it processes and analyzes text and information. Finally, the chapter introduces some deep learning architectures that modify BERT for more efficiency (e.g., RoBERTa) and other types of NLP applications (e.g., NLP for tabular data—TaBERT). The chapter concludes with real-world case studies on using BERT for sentiment classification and topic modeling applications.

4. Multilingual transfer learning is an area where transformer architectures have significantly impacted the field of machine learning. Chapter 4 introduces an overview of transformer-based

multilingual architectures and how cross-lingual transfer learning is pre-trained and fine-tuned for NLP tasks. The chapter also provides an overview of the state-of-the-art benchmarks used for multilingual NLP. The chapter further provides some insights into the research and techniques identifying the factors affecting cross-lingual and zero-shot transfer learning in NLP. Finally, a real-world case study of using multilingual universal sentence encoders for zero-shot cross-lingual sentiment classification is presented.

5. In Chapter 5 we discuss various modifications made to the standard transformer architecture to tackle longer sequences with limited memory, to build transformer models that are faster and of higher quality, and that perform better on text generation and summarization. We also discuss the key differences between the model architectures and approaches that are centered around key ideas such as knowledge distillation and making computations more efficient by reducing attention mechanism complexity. This chapter includes a case study that uses a pre-trained transformer model for sentiment classification, including a look at the contents of the model's multi-head attention mechanisms.

6. Since BERT, many flavors of pre-trained models have been made available across different domains, providing models that can be fine-tuned to domain-specific data across science, medicine, and finance. In addition, language-specific pre-trained models offer increasingly competitive results on downstream language specific tasks. In Chapter 6, we discuss the pre-trained models that are available, showing their benefits and applications to specific domains such as computer vision, speech, time series, and text. This chapter includes a case study that compares the performance of three transformer-based automatic speech recognition models.

7. There is a need to understand the models from an explainability standpoint in many critical applications and given the black-box nature of transformers-based models. In Chapter 7, we will cover the traits of the models that address explainability, related areas that impact explainability, the taxonomy of explainable methods applied to the transformer-based and attention-based systems, and finally, a detailed case study in the electronic health record systems using transformers with different explainable techniques to become more practical.

Authors

Uday Kamath has spent more than two decades developing analytics products and combines this experience with learning in statistics, optimization, machine learning, bioinformatics, and evolutionary computing. He has contributed to many journals, conferences, and books, is the author of *XAI: An Introduction to Interpretable XAI, Deep Learning for NLP and Speech Recognition, Mastering Java Machine Learning, and Machine Learning: End-to-End Guide for Java Developers*. He held many senior roles: chief analytics officer for Digital Reasoning, advisor for Falkonry, and chief data scientist for BAE Systems Applied Intelligence. Dr. Kamath has many patents and has built commercial products using AI in domains such as compliance, cybersecurity, financial crime, and bioinformatics. He currently works as the chief analytics officer for Smarsh. He is responsible for data science, research of analytical products employing deep learning, transformers, explainable AI, and modern techniques in speech and text for the financial domain and healthcare.

Kenneth L. Graham has two decades experience solving quantitative problems in multiple domains, including Monte Carlo simulation, NLP, anomaly detection, cybersecurity, and behavioral profiling. For the past ten years, he has focused on building scalable solutions in NLP for government and industry, including entity coreference resolution, text classification, active learning, automatic speech recognition, and temporal normalization. He currently works at AppFolio as a senior machine learning engineer. Dr. Graham has five patents for his work in natural language processing, seven research publications, and a PhD in condensed matter physics.

Wael Emara has two decades of experience in academia and industry. He has a PhD in computer engineering and computer science with emphasis on machine learning and artificial intelligence. His technical background and research spans signal and image processing, computer vision, medical imaging, social media analytics, machine learning,

and natural language processing. Dr. Emara has contributed to many peer-reviewed publications in various machine learning topics and he is active in the technical community in the greater New York area. He currently works as a principal data scientist at Smarsh, Inc. for Digital Reasoning where he is doing research on state-of-the-art artificial intelligence NLP systems.

Contributors

Krishna Choppella
BAE Systems AI
Toronto, Canada

Mitch Naylor
Smarsh, Inc.
Nashville, Tennessee

Vedant Vajre
Stone Bridge High School
Ashburn, Virginia

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

CHAPTER 1

Deep Learning and Transformers: An Introduction

TRANSFORMERS are deep learning models that have achieved state-of-the-art performance in several fields such as natural language processing, computer vision, and speech recognition. Indeed, the massive surge of recently proposed transformer model variants has meant researchers and practitioners alike find it challenging to keep pace. In this chapter, we provide a brief history of diverse research directly or indirectly connected to the innovation of transformers. Next, we discuss a taxonomy based on changes in the architecture for efficiency in computation, memory, applications, etc., which can help navigate the complex innovation space. Finally, we provide resources in tools, libraries, books, and online courses that the readers can benefit from in their pursuit.

1.1 DEEP LEARNING: A HISTORIC PERSPECTIVE

In the early 1940s, S. McCulloch and W. Pitts, using a simple electrical circuit called a "threshold logic unit", simulated intelligent behavior by emulating how the brain works [179]. The simple model had the first neuron with inputs and outputs that would generate an output 0 when the "weighted sum" was below a threshold and 1 otherwise, which later became the basis of all the neural architectures. The weights were not learned but adjusted. In his book *The Organization of Behaviour* (1949), Donald Hebb laid the foundation of complex neural processing

DOI: 10.1201/9781003170082-1

by proposing how neural pathways can have multiple neurons firing and strengthening over time [108]. Frank Rosenblatt, in his seminal work, extended the McCulloch–Pitts neuron, referring to it as the "Mark I Perceptron"; given the inputs, it generated outputs using linear thresholding logic [212].

The weights in the perceptron were "learned" by repeatedly passing the inputs and reducing the difference between the predicted output and the desired output, thus giving birth to the basic neural learning algorithm. Marvin Minsky and Seymour Papert later published the book *Perceptrons* which revealed the limitations of perceptrons in learning the simple exclusive-or function (XOR) and thus prompting the so-called The First AI Winter [186].

John Hopfield introduced "Hopfield Networks", one of the first recurrent neural networks (RNNs) that serve as a content-addressable memory system [117].

In 1986, David Rumelhart, Geoff Hinton, and Ronald Williams published the seminal work "Learning representations by back-propagating errors" [217]. Their work confirms how a multi-layered neural network using many "hidden" layers can overcome the weakness of perceptrons in learning complex patterns with relatively simple training procedures. The building blocks for this work had been laid down by various research over the years by S. Linnainmaa, P. Werbos, K. Fukushima, D. Parker, and Y. LeCun [164, 267, 91, 196, 149].

LeCun et al., through their research and implementation, led to the first widespread application of neural networks to recognize the handwritten digits used by the U.S. Postal Service [150]. This work is a critical milestone in deep learning history, proving the utility of convolution operations and weight sharing in learning the features in computer vision.

Backpropagation, the key optimization technique, encountered a number of issues such as vanishing gradients, exploding gradients, and the inability to learn long-term information, to name a few [115]. Hochreiter and Schmidhuber, in their work,"Long short-term memory (LSTM)" architecture, demonstrated how issues with long-term dependencies could overcome shortcomings of backpropagation over time [116].

Hinton et al. published a breakthrough paper in 2006 titled "A fast learning algorithm for deep belief nets"; it was one of the reasons for the resurgence of deep learning [113]. The research highlighted the effectiveness of layer-by-layer training using unsupervised methods followed by supervised "fine-tuning" to achieve state-of-the-art results in character recognition. Bengio et al., in their seminal work following this, offered

deep insights into why deep learning networks with multiple layers can hierarchically learn features as compared to shallow neural networks [27]. In their research, Bengio and LeCun emphasized the advantages of deep learning through architectures such as convolutional neural networks (CNNs), restricted Boltzmann machines (RBMs), and deep belief networks (DBNs), and through techniques such as unsupervised pre-training with fine-tuning, thus inspiring the next wave of deep learning [28]. Fei-Fei Li, head of the artificial intelligence lab at Stanford University, along with other researchers, launched ImageNet, which resulted in the most extensive collection of images and, for the first time, highlighted the usefulness of data in learning essential tasks such as object recognition, classification, and clustering [70]. Improvements in computer hardware, primarily through GPUs, increasing the throughput by almost 10× every five years, and the existence of a large amount of data to learn from resulted in a paradigm shift in the field. Instead of hand-engineered features that were the primary focus for many sophisticated applications, by learning from a large volume of training data, where the necessary features emerge, the deep learning network became the foundation for many state-of-the-art techniques.

Mikolov et al. and Graves proposed language models using RNNs and long short-term memory, which later became the building blocks for many natural language processing (NLP) architectures [184, 97]. The research paper by Collobert and Weston was instrumental in demonstrating many concepts such as pre-trained word embeddings, CNNs for text, and sharing of the embedding matrix for multi-task learning [60]. Mikolov et al. further improved the efficiency of training the word embeddings proposed by Bengio et al. by eliminating the hidden layer and formulating an approximate objective for learning giving rise to "word2vec", an efficient large-scale implementation of word embeddings [185, 183]. Sutskever's research, which proposed a Hessian-free optimizer to train RNNs efficiently on long-term dependencies, was a breakthrough in reviving the usage of RNNs, especially in NLP [237]. Sutskever et al. introduced sequence-to-sequence learning as a generic neural framework comprised of an encoder neural network processing inputs as a sequence and a decoder neural network predicting the outputs based on the input sequence states and the current output states [238]. As a result, the sequence-to-sequence framework became the core architecture for a wide range of NLP tasks such as constituency parsing, named entity recognition (NER), machine translation, question-answering, and summarization, to name a few. Furthermore, even Google started replacing its

monolithic phrase-based machine translation models with sequence-to-sequence neural machine translation models [272]. To overcome the bottleneck issues with the sequence-to-sequence framework, seminal work by Bahdanau et al. proposed the attention mechanism, which plays a crucial role in transformers and their variants [17].

1.2 TRANSFORMERS AND TAXONOMY

The transformer architecture [254] was introduced in 2017, in the paper *Attention Is All You Need*, for sequence-to-sequence problems. It was an alternative to using recurrent or convolutional layers. Since its introduction, there's been a wide variety of research into various ways to improve upon the standard transformer. Two surveys [163, 243] have categorized transformer-related papers. Transformer research has focused on three things: architecture modification, pre-training methods, and applications. In this book, we'll spend time on a subset of architecture modifications, pre-training methods of large language models, like BERT [71], and a few applications.

1.2.1 Modified Transformer Architecture

Modified transformer architectures can be split into two broad categories [163]: changes to the internal arrangement of the transformer block and changes to the layers that a transformer block is made of. A summary of the types of transformer modifications are shown in Table 1.1

1.2.1.1 Transformer block changes

Thus far, modifications to the transformer block have fallen into five categories [163]:

- Decreasing memory footprint and compute
- Adding connections between transformer blocks
- Adaptive computation time (e.g., allow early stopping during training)
- Recurrence or hierarchical structure
- Changing the architecture more drastically (e.g., neural architecture search)

TABLE 1.1 Types of modifications to the transformer block

Modification	**Transformer**
Lightweight transformers	
	Lite Transformer [274] Funnel Transformer [66] DeLighT [180]
Cross-block connectivity	
	Realformer [107] Transparent Attention [19]
Adaptive computation time	
	Universal Transformer [69] Conditional Computation Transformer [18] DeeBERT [276]
Recurrent	
	Transformer-XL [67] Compressive Transformer [204] Memformer [287]
Hierarchical	
	HIBERT [296] Hi-Transformer [270]
Different architectures	
	Macaron Transformer [174] Sandwich Transformer [201] Differentiable Architecture Search [299]

In this book, we'll focus on several architecture modifications that allow a transformer to process longer sequences and/or lower the computational complexity of the attention mechanism. We show a partial list of modified transformers in Table 1.1.

1.2.1.2 Transformer sublayer changes

In Chapter 2, we'll take a detailed look at the structure of a transformer block, covering its four components so we can later discuss ways in which researchers have modified them. In general, there are four parts to a transformer block [254]: positional encodings, multi-head attention, residual connections with layer normalization [13], and a position-wise feedforward network. Changes to transformer sublayers have focused on

TABLE 1.2 Types of modifications to the multi-head attention module

Modification	Transformer
Low-rank	
	Performer [53] Nystromformer [277] Synthesizer [241]
Attention with prior	
	Gaussian Transformer [102] Realformer [107] Synthesizer [241] Longformer [25]
Improved multi-head attention	
	Talking-heads Attention [227] Multi-Scale Transformer [234]
Complexity reduction	
	Longformer [25] Reformer [142] Big Bird [292] Performer [53] Routing Transformer [214]
Prototype queries	
	Clustered Attention [256] Informer [302]
Clustered key-value memory	
	Set Transformer [151] Memory Compressed Transformer [167] Linformer [259]

these four types of components, most of which has focused on changing aspects of the multi-head attention [163, 243]. Table 1.2 shows a selected list of transformer variants that have modified multi-head attention mechanisms.

Multi-head attention Much effort has been directed at the multi-head attention mechanism; studying its quadratic computational complexity, ways to address said complexity, and how it might be changed for specific kinds of problems. Most of this work falls into two broad cat-

egories: reducing the computational complexity of the attention mechanism, or changing the attention mechanism so it can learn more things.

As discussed in refs [163, 243], there are many ways to address the complexity of the attention mechanism. There are low-rank approximations, like Linformer [259] and Performer [53]. There are several ways to sparsify the attention mechanism, some of which effectively reduce the complexity of the attention mechanism to be linear in the sequence length. For example Longformer [25] and BigBird [292] add sparsity by fixing the positions to which a given token can attend. Some other transformers, like Reformer [142], introduce a learnable sparsity by sorting or clustering the input tokens. There are others still which reduce the size of the attention matrix [163].

There is also a variety of work that has tried to improve the multi-head attention mechanism [163]. For instance, attention heads have been allowed to "communicate" with each other and/or share information [158, 227, 65, 226], learn the optimal span to attend over [235], and use different attention spans in different attention heads [103]. This list is not exhaustive. We discuss several such methods in Chapter 5.

Positional encodings Positional encodings [254] are a way of encoding sequence order into the transformer. They also comprise another avenue for modifying the components of a transformer block. Thus far, four kinds of positional encodings have been used [163]: absolute positional encodings (like those of the standard transformer), relative positional encodings (such as in Transformer-XL), hybrid encodings that have absolute and relative position information, and implicit encodings that provide information about sequence order in other ways. This is shown in Table 1.3. We discuss the absolute positional encodings used in the standard transformer in Chapter 2 and the relative encodings used in Transformer-XL in Chapter 5.

TABLE 1.3 Changes to positional encodings

Modification	Transformer
Absolute position	Original Transformer [254]
Relative position	Transformer-XL [67]
Absolute/relative hybrid	Roformer [232]
Other representations	R-Transformer [262]

Residual connections and position-wise feedforward networks Some work has included changes to the residual blocks that come after the multi-head attention mechanism and after the position-wise feedforward network, including the position of the layer normalization, swapping layer normalization with something else, removal of layer normalization entirely [163], or the introduction or reversible residual layers to conserve memory (used in Reformer) [142]. Reformer will be discussed in Chapter 5. Other work has studied ways to change the position-wise feedforward network, including changing the activation function, increasing its representational capacity, or removing the feedforward network.

1.2.2 Pre-training Methods and Applications

A large body of work has focused on how a transformer can be pre-trained. There are encoder-only models, such as BERT [71], decoder-only models like the famed generative pre-trained transformer models GPT-3 [32], and encoder-decoder models like T5 [205] and ByT5 [280]. BERT is discussed in detail in Chapter 3, T5 in Chapter 5, and ByT5 in Chapter 6.

There have been many application and domain-specific transformers made for specific data domains (e.g., financial or medical text) and specific kinds of data (e.g., images or video). We discuss several such applications in Chapter 6.

1.3 RESOURCES

In this section, we will discuss some resources that can be useful for researchers and practitioners.

1.3.1 Libraries and Implementations

Here are some useful libraries, tools, and implementations (Table 1.4):

TABLE 1.4 Libraries and Tools

Organization	Language and Framework	API	Pre-trained
AllenNLP	Python and PyTorch	Yes	Yes
HuggingFace	Jax, PyTorch, and TensorFlow	Yes	Yes
Google Brain	TensorFlow	Yes	
GluonNLP	MXNet	Yes	Yes

1.3.2 Books

Some of the books that we found useful are:

- Transfer Learning for Natural Language Processing by Paul Azunre [12]
- Transformers for Natural Language Processing by Denis Rothman [213]
- Deep Learning Algorithms: Transformers, gans, encoders, rnns, cnns, and more by Ricardo A. Calix [35]
- Python Transformers By Huggingface Hands On by Joshua K. Cage [33]
- Deep Learning for NLP and Speech Recognition by Uday Kamath, John Liu, and James Whitaker [248]

1.3.3 Courses, Tutorials, and Lectures

Some of the very relevant online courses and tutorials:

- The Annotated Transformer by Alexander Rush et al. http://nlp.seas.harvard.edu/2018/04/03/attention.html
- HuggingFace course on transformers https://huggingface.co/course
- DeepLearning.AI course on sequence models https://www.coursera.org/learn/nlp-sequence-models
- DeepLearning.AI course on transformers and BERT https://www.coursera.org/learn/attention-models-in-nlp
- Stanford CS224N: NLP with Deep Learning by Christopher Manning http://web.stanford.edu/class/cs224n/
- UC Berkeley's Applied Natural Language Processing by David Bamman https://people.ischool.berkeley.edu/~dbamman/info256.html
- Advanced NLP with spaCy by Ines Montani https://course.spacy.io/en/

- Deep Learning for Coders with fastai and PyTorch by Sylvain Gugger and Jeremy Howard https://course.fast.ai/
- Jay Alammar's visual explanation of transformers and related architectures https://jalammar.github.io/

1.3.4 Case Studies and Details

At the end of Chapters 2–7, we include a case study that allows the reader to see how one or more of the models and methods discussed in the chapter can be applied, or how they stack up against one another when applied to the same problem. The aim is for the case study to provide a small starting point in working with transformer models from which one can branch out further. Each case study has been chosen to run within approximately one hour on GPUs at least as powerful as the NVIDIA K80 (Google Colaboratory provides these for free). Case studies are also available in the Github repository which accompanies this book: https://github.com/CRCTransformers/deepdive-book.

References

S. Abnar and W. Zuidema , *Quantifying attention flow in transformers*, arXiv preprint arXiv:2005.00928, (2020).

G. Alain and Y. Bengio , *Understanding intermediate layers using linear classifier probes*, arXiv preprint arXiv:1610.01644, (2016).

D. Alvarez-Melis and T. S. Jaakkola , *Towards robust interpretability with self-explaining neural networks*, arXiv preprint arXiv:1806.07538, (2018).

S. Amershi , M. Chickering , S. M. Drucker , B. Lee , P. Simard , and J. Suh , *Modeltracker: Redesigning performance analysis tools for machine learning*, in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 337–346.

D. Amodei , C. Olah , J. Steinhardt , P. Christiano , J. Schulman , and D. Mané , *Concrete problems in ai safety*, arXiv preprint arXiv:1606.06565, (2016).

A. Andoni , P. Indyk , T. Laarhoven , I. P. Razenshteyn , and L. Schmidt , *Practical and optimal lsh for angular distance*, in NIPS, 2015.

D. Araci , *Finbert: Financial sentiment analysis with pre-trained language models*, ArXiv, abs/1908.10063 (2019).

E. Arkhangelskaia and S. Dutta , *Whatcha lookin'at? deeplifting bert's attention in question answering*, arXiv preprint arXiv:1910.06431, (2019).

M. Artetxe , S. Ruder , and D. Yogatama , *On the cross-lingual transferability of monolingual representations*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4623–4637.

M. Artetxe , S. Ruder , and D. Yogatama , *On the cross-lingual transferability of monolingual representations*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky , J. Chai , N. Schluter , and J. R. Tetreault , eds., Association for Computational Linguistics, 2020, pp. 4623–4637.

M. Artetxe and H. Schwenk , *Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond* , Trans. Assoc. Comput. Linguistics, 7 (2019), pp. 597–610.

P. Azunre , Transfer Learning for Natural Language Processing, Manning, 2021.

L. J. Ba , J. R. Kiros , and G. E. Hinton , *Layer normalization*, CoRR, abs/1607.06450 (2016).

S. Bach , A. Binder , G. Montavon , F. Klauschen , K.-R. Müller , and W. Samek , *On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation* , PloS one, 10 (2015), p. e0130140.

D. Baehrens , T. Schroeter , S. Harmeling , M. Kawanabe , K. Hansen , and K.-R. Müller , *How to explain individual classification decisions* , The Journal of Machine Learning Research, 11 (2010), pp. 1803–1831.

A. Baevski , H. Zhou , A. rahman Mohamed , and M. Auli , *wav2vec 2.0: A framework for self-supervised learning of speech representations*, ArXiv, abs/2006.11477 (2020).

D. Bahdanau , K. Cho , and Y. Bengio , *Neural machine translation by jointly learning to align and translate*, CoRR, abs/1409.0473 (2014).

A. Bapna , N. Arivazhagan , and O. Firat , *Controlling computation versus quality for neural sequence models*, ArXiv, abs/2002.07106 (2020).

A. Bapna , M. Chen , O. Firat , Y. Cao , and Y. Wu , *Training deeper neural machine translation models with transparent attention*, in EMNLP, 2018.

S. Barocas and D. Boyd , *Engaging the ethics of data science in practice* , Communications of the ACM, 60 (2017), pp. 23–25.

O. Bastani , C. Kim , and H. Bastani , *Interpreting blackbox models via model extraction*, arXiv preprint arXiv:1705.08504, (2017).

K. Baum , M. A. Köhl , and E. Schmidt , *Two challenges for ci trustworthiness and how to address them*, in Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017), 2017.

Y. Belinkov , *Probing classifiers: Promises, shortcomings, and alternatives*, arXiv preprint arXiv:2102.12452, (2021).

I. Beltagy , K. Lo , and A. Cohan , *Scibert: A pretrained language model for scientific text*, in EMNLP, 2019.

I. Beltagy , M. E. Peters , and A. Cohan , *Longformer: The long-document transformer*, ArXiv, abs/2004.05150 (2020).

E. M. Bender , D. Hovy , and A. Schofield , *Integrating ethics into the nlp curriculum*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, 2020, pp. 6–9.

Y. Bengio , P. Lamblin , D. Popovici , and H. Larochelle , *Greedy layer-wise training of deep networks*, in Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06, MIT Press, 2006, pp. 153–160.

Y. Bengio and Y. LeCun , Scaling learning algorithms towards AI, MIT Press, 2007.

G. Bertasius , H. Wang , and L. Torresani , *Is space-time attention all you need for video understanding?*, in ICML, 2021.

A. Beutel , J. Chen , Z. Zhao , and E. H. Chi , *Data decisions and theoretical implications when adversarially learning fair representations*, arXiv preprint arXiv:1707.00075, (2017).

S. R. Bowman , L. Vilnis , O. Vinyals , A. Dai , R. Jozefowicz , and S. Bengio , *Generating sentences from a continuous space*, in Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, August 2016, Association for Computational Linguistics, pp. 10–21.

T. B. Brown , B. Mann , N. Ryder , M. Subbiah , J. Kaplan , P. Dhariwal , A. Neelakantan , P. Shyam , G. Sastry , A. Askell , S. Agarwal , A. Herbert-Voss , G. Krueger , T. Henighan , R. Child , A. Ramesh , D. M. Ziegler , J. Wu , C. Winter , C. Hesse , M. Chen , E. Sigler , M. Litwin , S. Gray , B. Chess , J. Clark , C. Berner , S. McCandlish , A. Radford , I. Sutskever , and D. Amodei , *Language models are few-shot learners*, ArXiv, abs/2005.14165 (2020).

J. Cage , *Python Transformers By Huggingface Hands On: 101 practical implementation hands-on of ALBERT/ViT/BigBird and other latest models with huggingface transformers*, 2021.

T. Calders , F. Kamiran , and M. Pechenizkiy , *Building classifiers with independency constraints*, in 2009 IEEE International Conference on Data Mining Workshops, IEEE, 2009, pp. 13–18.

R. A. Calixn , *Deep Learning Algorithms: Transformers, gans, encoders, rnns, cnns, and more*, 2020.

F. P. Calmon , D. Wei , B. Vinzamuri , K. N. Ramamurthy , and K. R. Varshney , *Optimized pre-processing for discrimination prevention*, in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 3995–4004.

O.-M. Camburu , T. Rocktäschel , T. Lukasiewicz , and P. Blunsom , *e-snli: Natural language inference with natural language explanations* , arXiv preprint arXiv:1812.01193, (2018).

D. M. Cer , M. T. Diab , E. Agirre , I. Lopez-Gazpio , and L. Specia , *Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation*, in SemEval@ACL, 2017.

M. Charikar , *Similarity estimation techniques from rounding algorithms*, in STOC ‘02, 2002.

C. Chen , O. Li , C. Tao , A. J. Barnett , J. Su , and C. Rudin , *This looks like that: deep learning for interpretable image recognition*, arXiv preprint arXiv:1806.10574, (2018).
L. Chen , K. Lu , A. Rajeswaran , K. Lee , A. Grover , M. Laskin , P. Abbeel , A. Srinivas , and I. Mordatch , *Decision transformer: Reinforcement learning via sequence modeling*, ArXiv, abs/2106.01345 (2021).
S. F. Chen , D. Beeferman , and R. Rosenfeld , *Evaluation metrics for language models*, (1998).
Z. Chen , H. Zhang , X. Zhang , and L. Zhao , *Quora question pairs*, 2017.
E. A. Chi , J. Hewitt , and C. D. Manning , *Finding universal grammatical relations in multilingual BERT*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, D. Jurafsky , J. Chai , N. Schluter , and J. R. Tetreault , eds., Association for Computational Linguistics, 2020, pp. 5564–5577.
Z. Chi , L. Dong , F. Wei , X. Mao , and H. Huang , *Can monolingual pretrained models help cross-lingual classification?*, in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4–7, 2020, K. Wong , K. Knight , and H. Wu , eds., Association for Computational Linguistics, 2020, pp. 12–17.
Z. Chi , L. Dong , F. Wei , W. Wang , X. Mao , and H. Huang , *Cross-lingual natural language generation via pre-training* , in The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, AAAI Press, 2020, pp. 7570–7577.
Z. Chi , L. Dong , F. Wei , N. Yang , S. Singhal , W. Wang , X. Song , X. Mao , H. Huang , and M. Zhou , *Infoxlm: An information-theoretic framework for cross-lingual language model pre-training*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6–11, 2021, K. Toutanova , A. Rumshisky , L. Zettlemoyer , D. Hakkani - Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty , and Y. Zhou , eds., Association for Computational Linguistics, 2021, pp. 3576–3588.
Z. Chi , L. Dong , F. Wei , N. Yang , S. Singhal , W. Wang , X. Song , X.-L. Mao , H. Huang , and M. Zhou , *InfoXLM: An information-theoretic framework for cross-lingual language model pre-training*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, June 2021, Association for Computational Linguistics, pp. 3576–3588.
M. Chidambaram , Y. Yang , D. M. Cer , S. Yuan , Y.-H. Sung , B. Strope , and R. Kurzweil , *Learning cross-lingual sentence representations via a multi-task dual-encoder model*, in RepL4NLP@ACL, 2019.
R. Child , S. Gray , A. Radford , and I. Sutskever , *Generating long sequences with sparse transformers*, ArXiv, abs/1904.10509 (2019).
K. Cho , B. van Merriënboer , C. Gulcehre , D. Bahdanau , F. Bougares , H. Schwenk , and Y. Bengio , *Learning phrase representations using RNN encoder–decoder for statistical machine translation*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 2014, Association for Computational Linguistics, pp. 1724–1734.
R. Choenni and E. Shutova , *What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties*, arXiv preprint arXiv:2009.12862, (2020).
K. Choromanski , V. Likhosherstov , D. Dohan , X. Song , A. Gane , T. Sarlós , P. Hawkins , J. Davis , A. Mohiuddin , L. Kaiser , D. Belanger , L. J. Colwell , and A. Weller , *Rethinking attention with performers*, ArXiv, abs/2009.14794 (2021).
H. W. Chung , T. Fevry , H. Tsai , M. Johnson , and S. Ruder , *Rethinking embedding coupling in pre-trained language models* , in International Conference on Learning Representations, 2021.
H. W. Chung , D. Garrette , K. C. Tan , and J. Riesa , *Improving multilingual models with language-clustered vocabularies*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020, B. Webber , T. Cohn , Y. He , and Y. Liu , eds., Association for Computational Linguistics, 2020, pp. 4536–4546.
J. H. Clark , E. Choi , M. Collins , D. Garrette , T. Kwiatkowski , V. Nikolaev , and J. Palomaki , *Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages*, Transactions of the Association for Computational Linguistics, (2020).
K. Clark , U. Khandelwal , O. Levy , and C. D. Manning , *What does bert look at? an analysis of bert's attention*, ArXiv, abs/1906.04341 (2019).
K. Clark , M.-T. Luong , Q. V. Le , and C. D. Manning , *Electra: Pre-training text encoders as discriminators rather than generators*, ArXiv, abs/2003.10555 (2020).
M. Coeckelbergh , AI ethics, MIT Press, 2020.
R. Collobert and J. Weston , *A unified architecture for natural language processing: Deep neural networks with multitask learning*, in Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 160–167.
A. Conneau , K. Khandelwal , N. Goyal , V. Chaudhary , G. Wenzek , F. Guzmán , E. Grave , M. Ott , L. Zettlemoyer , and V. Stoyanov , *Unsupervised cross-lingual representation learning at scale*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, D. Jurafsky , J. Chai , N. Schluter , and J.R. Tetreault , eds., Association for Computational Linguistics, 2020, pp. 8440–8451.
A. Conneau , G. Kruszewski , G. Lample , L. Barrault , and M. Baroni , *What you can cram into a single vector: Probing sentence embeddings for linguistic properties*, arXiv preprint arXiv:1805.01070, (2018).
A. Conneau and G. Lample , *Cross-lingual language model pretraining*, in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8–14 December 2019, Vancouver, BC, Canada, H. M. Wallach , H. Larochelle , A. Beygelzimer , F. d'Alché-Buc , E. B. Fox , and R. Garnett , eds., 2019, pp. 7057–7067.
A. Conneau , R. Rinott , G. Lample , A. Williams , S. R. Bowman , H. Schwenk , and V. Stoyanov , *Xnli: Evaluating cross-lingual sentence representations*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018.
J.-B. Cordonnier , A. Loukas , and M. Jaggi , *Multi-head attention: Collaborate instead of concatenate*, ArXiv, abs/2006.16362 (2020).
Z. Dai , G. Lai , Y. Yang , and Q. V. Le , *Funnel-transformer: Filtering out sequential redundancy for efficient language processing*, ArXiv, abs/2006.03236 (2020).
Z. Dai , Z. Yang , Y. Yang , J. Carbonell , Q. V. Le , and R. Salakhutdinov , *Transformer-xl: Attentive language models beyond a fixed-length context*, in ACL, 2019.
W. de Vries , A. van Cranenburgh , and M. Nissim , *What's so special about bert's layers? A closer look at the NLP pipeline in monolingual and multilingual models*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16–20 November 2020, T. Cohn , Y. He , and Y. Liu , eds., vol. EMNLP 2020 of Findings of ACL, Association for Computational Linguistics, 2020, pp. 4339–4350.

M. Dehghani , S. Gouws , O. Vinyals , J. Uszkoreit , and L. Kaiser , *Universal transformers*, ArXiv, abs/1807.03819 (2019).
J. Deng , W. Dong , R. Socher , L.-J. Li , K. Li , and L. Fei-Fei , *ImageNet: A Large-Scale Hierarchical Image Database*, in CVPR09, 2009.
J. Devlin , M.-W. Chang , K. Lee , and K. Toutanova , *Bert: Pre-training of deep bidirectional transformers for language understanding*, in NAACL-HLT, 2019.
J. Devlin , M.-W. Chang , K. Lee , and K. Toutanova , *BERT: Pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 4171–4186.
Y. Ding , Y. Liu , H. Luan , and M. Sun , *Visualizing and understanding neural machine translation*, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1150–1159.
S. Doddapaneni , G. Ramesh , A. Kunchukuttan , P. Kumar , and M. M. Khapra , *A primer on pretrained multilingual language models*, CoRR, abs/2107.00676 (2021).
W. B. Dolan and C. Brockett , *Automatically constructing a corpus of sentential paraphrases*, in IJCNLP, 2005.
Y. Dong , F. Liao , T. Pang , H. Su , J. Zhu , X. Hu , and J. Li , *Boosting adversarial attacks with momentum*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9185–9193.
Y. Dong , H. Su , J. Zhu , and B. Zhang , *Improving interpretability of deep neural networks with semantic information*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4306–4314.
A. Dosovitskiy , L. Beyer , A. Kolesnikov , D. Weissenborn , X. Zhai , T. Unterthiner , M. Dehghani , M. Minderer , G. Heigold , S. Gelly , J. Uszkoreit , and N. Houlsby , *An image is worth 16x16 words: Transformers for image recognition at scale*, ArXiv, abs/2010.11929 (2021).
P. Dufter and H. Schütze , *Identifying elements essential for BERT's multilinguality*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, November 2020, Association for Computational Linguistics, pp. 4423–4437.
V. P. Dwivedi and X. Bresson , *A generalization of transformer networks to graphs*, ArXiv, abs/2012.09699 (2020).
V. P. Dwivedi , C. K. Joshi , T. Laurent , Y. Bengio , and X. Bresson , *Benchmarking graph neural networks*, ArXiv, abs/2003.00982 (2020).
C. Dwork , N. Immorlica , A. T. Kalai , and M. Leiserson , *Decoupled classifiers for group-fair and efficient machine learning* , in Conference on Fairness, Accountability and Transparency, PMLR, 2018, pp. 119–133.
C. Eckart and G. M. Young , *The approximation of one matrix by another of lower rank* , Psychometrika, 1 (1936), pp. 211–218.
P. Erdös , On random graphs, *i* , 1959.
D. Erhan , Y. Bengio , A. Courville , and P. Vincent , Visualizing higher-layer features of a deep network, University of Montreal, 1341 (2009), p. 1.
K. Eykholt , I. Evtimov , E. Fernandes , B. Li , A. Rahmati , C. Xiao , A. Prakash , T. Kohno , and D. Song , *Robust physical-world attacks on deep learning visual classification*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.
W. Fedus , B. Zoph , and N. Shazeer , *Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity*, arXiv preprint arXiv:2101.03961, (2021).
F. Feng , Y. Yang , D. Cer , N. Arivazhagan , and W. Wang , *Language-agnostic BERT sentence embedding*, CoRR, abs/2007.01852 (2020).
R. C. Fong and A. Vedaldi , *Interpretable explanations of black boxes by meaningful perturbation*, in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.
F. B. Fuchs , O. Groth , A. R. Kosiorek , A. Bewley , M. Wulfmeier , A. Vedaldi , and I. Posner , *Neural stethoscopes: Unifying analytic, auxiliary and adversarial network probing.*, (2018).
K. Fukushima , *Neural network model for a mechanism of pattern recognition unaffected by shift in position—Neocognitron* , Trans. IECE, J62-A(10) (1979), pp. 658–665.
P. Gage , *A new algorithm for data compression* , The C Users Journal archive, 12 (1994), pp. 23–38.
e. a. Garofolo , John S. , *Timit acoustic-phonetic continuous speech corpus*, 1993.
A. N. Gomez , M. Ren , R. Urtasun , and R. B. Grosse , *The reversible residual network: Backpropagation without storing activations*, in NIPS, 2017.
I. J. Goodfellow , J. Shlens , and C. Szegedy , *Explaining and harnessing adversarial examples*, arXiv preprint arXiv:1412.6572, (2014).
P. Gordaliza , E. Del Barrio , G. Fabrice , and J.-M. Loubes , *Obtaining fairness using optimal transport theory*, in International Conference on Machine Learning, PMLR, 2019, pp. 2357–2365.
A. Graves , *Generating sequences with recurrent neural networks.*, CoRR, abs/1308.0850 (2013).
A. Graves , G. Wayne , and I. Danihelka , *Neural turing machines*, CoRR, abs/1410.5401(2014).
R. M. Gray and D. L. Neuhoff , *Quantization* , IEEE Trans. Inf. Theory, 44 (1998), pp. 2325–2383.
S. Gray , A. Radford , and D. P. Kingma , *GPU kernels for block-sparse weights*, 2017.
M. Grootendorst , *Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics.*, 2020.
M. Guo , Y. Zhang , and T. Liu , *Gaussian transformer: A lightweight approach for natural language inference*, in AAAI, 2019.
Q. Guo , X. Qiu , P. Liu , X. Xue , and Z. Zhang , *Multi-scale self-attention for text classification*, in AAAI, 2020.
K. He , H. Fan , Y. Wu , S. Xie , and R. Girshick , *Momentum contrast for unsupervised visual representation learning*, 2020.
K. He , X. Zhang , S. Ren , and J. Sun , *Deep residual learning for image recognition* , 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), pp. 770–778.
R. He , W. S. Lee , H. T. Ng , and D. Dahlmeier , *Effective attention modeling for aspect-level sentiment classification*, in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1121–1131.
R. He , A. Ravula , B. Kanagal , and J. Ainslie , *Realformer: Transformer likes residual attention*, in FINDINGS, 2021.
D. O. Hebb , The organization of behavior: A neuropsychological theory, Wiley, 1949.
L. A. Hendricks , Z. Akata , M. Rohrbach , J. Donahue , B. Schiele , and T. Darrell , *Generating visual explanations* , in European conference on computer vision, Springer, 2016, pp. 3–19.
J. Hewitt and P. Liang , *Designing and interpreting probes with control tasks*, arXiv preprint arXiv:1909.03368, (2019).
M. Hind , D. Wei , M. Campbell , N. C. Codella , A. Dhurandhar , A. Mojsilović , K. Natesan Ramamurthy , and K. R. Varshney , *Ted: Teaching ai to explain its decisions*, in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 123–129.

G. Hinton , O. Vinyals , and J. Dean , *Distilling the knowledge in a neural network*, arXiv preprint arXiv:1503.02531, (2015).
G. E. Hinton , S. Osindero , and Y.-W. Teh , *A fast learning algorithm for deep belief nets* , Neural Comput., 18 (2006), pp. 1527–1554.
J. Ho , N. Kalchbrenner , D. Weissenborn , and T. Salimans , *Axial attention in multidimensional transformers*, ArXiv, abs/1912.12180 (2019).
S. Hochreiter , *The vanishing gradient problem during learning recurrent neural nets and problem solutions* , International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6 (1998), pp. 107–116.
S. Hochreiter and J. Schmidhuber , *Long short-term memory* , Neural Comput., 9 (1997), pp. 1735–1780.
J. J. Hopfield , *Neural networks and physical systems with emergent collective computational abilities* , Proceedings of the National Academy of Sciences of the United States of America, 79 (1982), pp. 2554–2558.
B.-J. Hou and Z.-H. Zhou , *Learning with interpretable structure from rnn*, arXiv preprint arXiv:1810.10708, (2018).
J. Howard and S. Ruder , *Universal language model fine-tuning for text classification*, in ACL, 2018.
W.-N. Hsu , B. Bolte , Y.-H. H. Tsai , K. Lakhotia , R. Salakhutdinov , and A. Mohamed , *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*, ArXiv, abs/2106.07447 (2021).
J. Hu , M. Johnson , O. Firat , A. Siddhant , and G. Neubig , *Explicit alignment objectives for multilingual bidirectional encoders*, 2021.
J. Hu , S. Ruder , A. Siddhant , G. Neubig , O. Firat , and M. Johnson , *XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation*, in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event, vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 4411–4421.
R. Hu and A. Singh , *Transformer is all you need: Multimodal multitask learning with a unified transformer*, ArXiv, abs/2102.10772 (2021).
C.-Z. A. Huang , A. Vaswani , J. Uszkoreit , N. M. Shazeer , C. Hawthorne , A. M. Dai , M. Hoffman , and D. Eck , *An improved relative self-attention mechanism for transformer with application to music generation*, ArXiv, abs/1809.04281 (2018).
H. Huang , Y. Liang , N. Duan , M. Gong , L. Shou , D. Jiang , and M. Zhou , *Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, K. Inui , J. Jiang , V. Ng , and X. Wan , eds., Association for Computational Linguistics, 2019, pp. 2485–2494.
Z. Huang , X. Wang , L. Huang , C. Huang , Y. Wei , H. Shi , and W. Liu , *Ccnet: Criss-cross attention for semantic segmentation* , 2019 IEEE/CVF International Conference on Computer Vision (ICCV), (2019), pp. 603–612.
D. A. Hudson and C. D. Manning , *Compositional attention networks for machine reasoning*, arXiv preprint arXiv:1803.03067, (2018).
D. Hupkes , S. Veldhoen , and W. Zuidema , *Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure* , Journal of Artificial Intelligence Research, 61 (2018), pp. 907–926.
P. Indyk and R. Motwani , *Approximate nearest neighbors: towards removing the curse of dimensionality*, in STOC '98, 1998.
S. Ioffe and C. Szegedy , *Batch normalization: Accelerating deep network training by reducing internal covariate shift.*, CoRR, abs/1502.03167 (2015).
R. Iyer , Y. Li , H. Li , M. Lewis , R. Sundar , and K. Sycara , *Transparency and explanation in deep reinforcement learning neural networks* , in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 144–150.
S. Jain and B. C. Wallace , *Attention is not explanation*, arXiv preprint arXiv:1902.10186, (2019).
W. James and F. H. Burkhardt , *The principles of psychology, the works of William James* , Transactions of the Charles S. Peirce Society, 19 (1983).
H. Jiang , B. Kim , M. Y. Guan , and M. R. Gupta , *To trust or not to trust a classifier* ., in NeurIPS, 2018, pp. 5546–5557.
K. K. Z. Wang , S. Mayhew , and D. Roth , *Cross-lingual ability of multilingual BERT: an empirical study* , in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020.
L. Kaiser and I. Sutskever , *Neural GPUs learn algorithms*, arXiv: Learning, (2016).
D. Kakwani , A. Kunchukuttan , S. Golla , G. N.C. A. Bhattacharyya , M. M. Khapra , and P. Kumar , *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages*, in Findings of EMNLP, 2020.
K. S. Kalyan , A. Rajasekharan , and S. Sangeetha , *AMMUS: A survey of transformer-based pretrained models in natural language processing*, CoRR, abs/2108.05542 (2021).
F. Kamiran and T. Calders , *Data preprocessing techniques for classification without discrimination* , Knowledge and Information Systems, 33 (2012), pp. 1–33.
D. Kang , D. Raghavan , P. Bailis , and M. Zaharia , *Model assertions for debugging machine learning*, in NeurIPS MLSys Workshop, 2018.
S. Khanuja , D. Bansal , S. Mehtani , S. Khosla , A. Dey , B. Gopalan , D. K. Margam , P. Aggarwal , R. T. Nagipogu , S. Dave , S. Gupta , S. C. B. Gali , V. Subramanian , and P. Talukdar , *Muril: Multilingual representations for Indian languages*, 2021.
N. Kitaev , L. Kaiser , and A. Levskaya , *Reformer: The efficient transformer*, ArXiv, abs/2001.04451 (2020).
G. Kobayashi , T. Kuribayashi , S. Yokoi , and K. Inui , *Attention is not only a weight: Analyzing transformers with vector norms*, in EMNLP, 2020.
T. Kudo and J. Richardson , *SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, November 2018, Association for Computational Linguistics, pp. 66–71.
G. Lai , Q. Xie , H. Liu , Y. Yang , and E. H. Hovy , *Race: Large-scale reading comprehension dataset from examinations*, in EMNLP, 2017.
G. Lample and A. Conneau , *Cross-lingual language model pretraining*, Advances in Neural Information Processing Systems (NeurIPS), (2019).
S. Lapuschkin , A. Binder , G. Montavon , K.-R. Muller , and W. Samek , *Analyzing classifiers: Fisher vectors and deep neural networks*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2912–2920.
A. Lauscher , V. Ravishankar , I. Vulic , and G. Glavas , *From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers*, CoRR, abs/2005.00633 (2020).
Y. LeCun , *Une procédure d'apprentissage pour réseau a seuil asymmetrique (a learning scheme for asymmetric threshold networks)*, in Proceedings of Cognitiva 85, 1985, pp. 599–604.
Y. LeCun , B. Boser , J. S. Denker , D. Henderson , R. E. Howard , W. Hubbard , and L. D. Jackel , *Backpropagation applied to handwritten zip code recognition* , Neural Computation, 1 (1989), pp. 541–551.

J. Lee , Y. Lee , J. Kim , A. R. Kosiorek , S. Choi , and Y. Teh , *Set transformer: A framework for attention-based permutation-invariant neural networks*, in ICML, 2019.
J. Lee , W. Yoon , S. Kim , D. Kim , S. Kim , C. H. So , and J. Kang , *Biobert: a pre-trained biomedical language representation model for biomedical text mining* , Bioinformatics, 36 (2020), pp. 1234–1240.
T. Lei , R. Barzilay , and T. Jaakkola , *Rationalizing neural predictions*, arXiv preprint arXiv:1606.04155, (2016).
G. Letarte , F. Paradis , P. Giguère , and F. Laviolette , *Importance of self-attention for sentiment analysis*, in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018, pp. 267–275.
M. Lewis , Y. Liu , N. Goyal , M. Ghazvininejad , A. Mohamed , O. Levy , V. Stoyanov , and L. Zettlemoyer , *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 7871–7880.
P. S. H. Lewis , B. Oguz , R. Rinott , S. Riedel , and H. Schwenk , *MLQA: evaluating cross-lingual extractive question answering*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, D. Jurafsky , J. Chai , N. Schluter , and J.R. Tetreault , eds., Association for Computational Linguistics, 2020, pp. 7315–7330.
J. Li , W. Monroe , and D. Jurafsky , *Understanding neural networks through representation erasure*, arXiv preprint arXiv:1612.08220, (2016).
J. Li , Z. Tu , B. Yang , M. R. Lyu , and T. Zhang , *Multi-head attention with disagreement regularization*, ArXiv, abs/1810.10183 (2018).
O. Li , H. Liu , C. Chen , and C. Rudin , *Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.
Y. Liang , N. Duan , Y. Gong , N. Wu , F. Guo , W. Qi , M. Gong , L. Shou , D. Jiang , G. Cao , X. Fan , R. Zhang , R. Agrawal , E. Cui , S. Wei , T. Bharti , Y. Qiao , J. Chen , W. Wu , S. Liu , F. Yang , D. Campos , R. Majumder , and M. Zhou , *XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020, B. Webber , T. Cohn , Y. He , and Y. Liu , eds., Association for Computational Linguistics, 2020, pp. 6008–6018.
J. Libovický , R. Rosa , and A. Fraser , *How language-neutral is multilingual bert?*, arXiv preprint arXiv:1911.03310, (2019).
T. Limisiewicz , D. Marecek , and R. Rosa , *Universal dependencies according to BERT: both more specific and more general*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16–20 November 2020, T. Cohn , Y. He , and Y. Liu , eds., vol. EMNLP 2020 of Findings of ACL, Association for Computational Linguistics, 2020, pp. 2710–2722.
T. Lin , Y. Wang , X. Liu , and X. Qiu , *A survey of transformers*, ArXiv, abs/2106.04554 (2021).
S. Linnainmaa , *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors*, Master's thesis, Univ. Helsinki, 1970.
C. Liu , T. Hsu , Y. Chuang , and H. Lee , *A study of cross-lingual ability and language-specific information in multilingual BERT*, CoRR, abs/2004.09205 (2020).
H. Liu , Q. Yin , and W. Y. Wang , *Towards explainable nlp: A generative explanation framework for text classification*, arXiv preprint arXiv:1811.00196, (2018).
P. J. Liu , M. Saleh , E. Pot , B. Goodrich , R. Sepassi , L. Kaiser , and N. M. Shazeer , *Generating Wikipedia by summarizing long sequences*, ArXiv, abs/1801.10198 (2018).
Q. Liu , D. McCarthy , I. Vulić , and A. Korhonen , *Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation*, in Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, November 2019, Association for Computational Linguistics, pp. 33–43.
Y. Liu , J. Gu , N. Goyal , X. Li , S. Edunov , M. Ghazvininejad , M. Lewis , and L. Zettlemoyer , *Multilingual denoising pre-training for neural machine translation* , Transactions of the Association for Computational Linguistics, 8 (2020), pp. 726–742.
Y. Liu , M. Ott , N. Goyal , J. Du , M. Joshi , D. Chen , O. Levy , M. Lewis , L. Zettlemoyer , and V. Stoyanov , *Roberta: A robustly optimized bert pretraining approach*, ArXiv, abs/1907.11692 (2019).
Z. Liu , G. I. Winata , A. Madotto , and P. Fung , *Exploring fine-tuning techniques for pre-trained cross-lingual models via continual learning* , CoRR, abs/2004.14218 (2020).
I. Loshchilov and F. Hutter , *Decoupled weight decay regularization*, arXiv preprint arXiv:1711.05101, (2017).
J. Lu , D. Batra , D. Parikh , and S. Lee , *ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*, arXiv preprint arXiv:1908.02265, (2019).
Y. Lu , Z. Li , D. He , Z. Sun , B. Dong , T. Qin , L. Wang , and T.-Y. Liu , *Understanding and improving transformer from a multi-particle dynamic system point of view*, ArXiv, abs/1906.02762 (2019).
F. Luo , W. Wang , J. Liu , Y. Liu , B. Bi , S. Huang , F. Huang , and L. Si , *{VECO}: Variable encoder-decoder pre-training for cross-lingual understanding and generation*, 2021.
M. Luong , H. Pham , and C. D. Manning , *Effective approaches to attention-based neural machine translation*, CoRR, abs/1508.04025 (2015).
X. Ma , X. Kong , S. Wang , C. Zhou , J. May , H. Ma , and L. Zettlemoyer , *Luna: Linear unified nested attention*, ArXiv, abs/2106.01540 (2021).
A. Madry , A. Makelov , L. Schmidt , D. Tsipras , and A. Vladu , *Towards deep learning models resistant to adversarial attacks*, arXiv preprint arXiv:1706.06083, (2017).
W. S. McCulloch and W. Pitts , *Neurocomputing: Foundations of research*, MIT Press, 1988, ch. A Logical Calculus of the Ideas Immanent in Nervous Activity, pp. 15–27.
S. Mehta , M. Ghazvininejad , S. Iyer , L. Zettlemoyer , and H. Hajishirzi , *Delight: Deep and light-weight transformer*, in ICLR, 2021.
S. Mehta , R. Koncel-Kedziorski , M. Rastegari , and H. Hajishirzi , *Pyramidal recurrent unit for language modeling*, in EMNLP, 2018.
S. Mehta , R. Koncel-Kedziorski , M. Rastegari , and H. Hajishirzi , *Define: Deep factorized input word embeddings for neural sequence modeling*, ArXiv, abs/1911.12385 (2020).
T. Mikolov , K. Chen , G. Corrado , and J. Dean , *Efficient estimation of word representations in vector space*, CoRR, abs/1301.3781 (2013).
T. Mikolov , M. Karafiát , L. Burget , J. Cernocký , and S. Khudanpur , *Recurrent neural network based language model.* , inINTERSPEECH, T. Kobayashi , K. Hirose , and S. Nakamura , eds., ISCA, 2010, pp. 1045–1048.

T. Mikolov , I. Sutskever , K. Chen , G. S. Corrado , and J. Dean , *Distributed representations of words and phrases and their compositionality* , in Advances in Neural Information Processing Systems 26, C. J. C. Burges , L. Bottou , M. Welling , Z. Ghahramani , and K. Q. Weinberger , eds., Curran Associates, Inc., 2013, pp. 3111–3119.

M. Minsky and S. A. Papert , Perceptrons: An introduction to computational geometry, MIT press, 2017.

G. Montavon , S. Lapuschkin , A. Binder , W. Samek , and K.-R. Müller , *Explaining nonlinear classification decisions with deep Taylor decomposition* , Pattern Recognition, 65 (2017), pp. 211–222.

G. Montavon , W. Samek , and K.-R. Müller , *Methods for interpreting and understanding deep neural networks* , Digital Signal Processing, 73 (2018), pp. 1–15.

S.-M. Moosavi-Dezfooli , A. Fawzi , O. Fawzi , and P. Frossard , *Universal adversarial perturbations*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1765–1773.

N. Mostafazadeh , M. Roth , A. Louis , N. Chambers , and J. F. Allen , *LSDSem 2017 shared task: The story cloze test*, in LSDSem@EACL, 2017.

A. Nguyen , J. Yosinski , and J. Clune , *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436.

J. Nivre , M. Abrams , Ž. Agić , L. Ahrenberg , L. Antonsen , K. Aplonova , M. J. Aranzabe , G. Arutie , M. Asahara , L. Ateyah , M. Attia , and et. al.,*Universal dependencies 2.3*, 2018. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

J. Nivre , R. Blokland , N. Partanen , and M. Rießler , *Universal dependencies 2.2*, November 2018.

X. Ouyang , S. Wang , C. Pang , Y. Sun , H. Tian , H. Wu , and H. Wang , *ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora*, 2021.

D. H. Park , L. A. Hendricks , Z. Akata , A. Rohrbach , B. Schiele , T. Darrell , and M. Rohrbach , *Multimodal explanations: Justifying decisions and pointing to the evidence*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8779–8788.

D. B. Parker , *Learning-logic*, Tech. Rep. TR-47, Center for Comp. Research in Economics and Management Sci., MIT, 1985.

M. Peters , M. Neumann , M. Iyyer , M. Gardner , C. Clark , K. Lee , and L. Zettlemoyer , *Deep contextualized word representations*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 2227–2237.

J. Phang , I. Calixto , P. M. Htut , Y. Pruksachatkun , H. Liu , C. Vania , K. Kann , and S. R. Bowman , *English intermediate-task training improves zero-shot cross-lingual transfer too*, in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4–7, 2020, K. Wong , K. Knight , and H. Wu , eds., Association for Computational Linguistics, 2020, pp. 557–575.

T. Pires , E. Schlinger , and D. Garrette , *How multilingual is multilingual bert?*, in Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers, A. Korhonen , D. R. Traum , and L. Màrquez , eds., Association for Computational Linguistics, 2019, pp. 4996–5001.

E. M. Ponti , G. Glavaš , O. Majewska , Q. Liu , I. Vulić , and A. Korhonen , *XCOPA: A multilingual dataset for causal commonsense reasoning*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

O. Press , N. A. Smith , and O. Levy , *Improving transformer models by reordering their sublayers*, in ACL, 2020.

A. Radford and K. Narasimhan , *Improving language understanding by generative pre-training*, 2018.

A. Radford , J. Wu , R. Child , D. Luan , D. Amodei , and I. Sutskever , *Language models are unsupervised multitask learners*, 2019.

J. W. Rae , A. Potapenko , S. M. Jayakumar , and T. Lillicrap , *Compressive transformers for long-range sequence modelling*, ArXiv, abs/1911.05507 (2020).

C. Raffel , N. M. Shazeer , A. Roberts , K. Lee , S. Narang , M. Matena , Y. Zhou , W. Li , and P. J. Liu , *Exploring the limits of transfer learning with a unified text-to-text transformer*, ArXiv, abs/1910.10683 (2020).

M. Raghu , J. Gilmer , J. Yosinski , and J. Sohl-Dickstein , *SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability*, arXiv preprint arXiv:1706.05806, (2017).

G. Ras , M. van Gerven , and P. Haselager , *Explanation methods in deep learning: Users, values, concerns and challenges* , in Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, 2018, pp. 19–36.

M. T. Ribeiro , S. Singh , and C. Guestrin , *"why should i trust you?" explaining the predictions of any classifier*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

M. T. Ribeiro , S. Singh , and C. Guestrin , *Anchors: High-precision model-agnostic explanations*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

M. Robnik-Šikonja and I. Kononenko , *Explaining classifications for individual instances* , IEEE Transactions on Knowledge and Data Engineering, 20 (2008), pp. 589–600.

T. Rocktäschel , E. Grefenstette , K. M. Hermann , T. Kočiskỳ , and P. Blunsom , *Reasoning about entailment with neural attention*, arXiv preprint arXiv:1509.06664, (2015).

F. Rosenblatt , *The perceptron: A probabilistic model for information storage and organization in the brain* , Psychological Review, (1958), pp. 65–386.

D. Rothman , Transformers for Natural Language Processing, Packt, 2021.

A. Roy , M. Saffar , A. Vaswani , and D. Grangier , *Efficient content-based sparse attention with routing transformers* , Transactions of the Association for Computational Linguistics, 9 (2021), pp. 53–68.

U. Roy , N. Constant , R. Al-Rfou , A. Barua , A. Phillips , and Y. Yang , *LAReQA: Language-agnostic answer retrieval from a multilingual pool*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, November 2020, Association for Computational Linguistics, pp. 5919–5930.

S. Ruder , N. Constant , J. Botha , A. Siddhant , O. Firat , J. Fu , P. Liu , J. Hu , G. Neubig , and M. Johnson , *XTREME-R: towards more challenging and nuanced multilingual evaluation*, CoRR, abs/2104.07412 (2021).

D. E. Rumelhart , G. E. Hinton , and R. J. Williams , *Neurocomputing: Foundations of research*, MIT Press, 1988, ch. Learning Representations by Back-propagating Errors, pp. 696–699.

P. Samangouei , M. Kabkab , and R. Chellappa , *Defense-GAN: Protecting classifiers against adversarial attacks using generative models* , arXiv preprint arXiv:1805.06605, (2018).

S. Schneider , A. Baevski , R. Collobert , and M. Auli , *wav2vec: Unsupervised pre-training for speech recognition*, in INTERSPEECH, 2019.
M. Schuster and K. Nakajima , *Japanese and Korean voice search* , 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2012), pp. 5149–5152.
H. Schwenk and X. Li , *A corpus for multilingual document classification in eight languages* , in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), N. C. C. chair), K. Choukri , C. Cieri , T. Declerck , S. Goggi , K. Hasida , H. Isahara , B. Maegaard , J. Mariani , H. Mazo , A. Moreno , J. Odijk , S. Piperidis , and T. Tokunaga , eds., Paris, France, may 2018, European Language Resources Association (ELRA).
R. Sennrich , B. Haddow , and A. Birch , *Improving neural machine translation models with monolingual data*, 2016.
R. Sennrich , B. Haddow , and A. Birch , *Neural machine translation of rare words with subword units*, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, August 2016, Association for Computational Linguistics, pp. 1715–1725.
P. Shaw , J. Uszkoreit , and A. Vaswani , *Self-attention with relative position representations*, in NAACL-HLT, 2018.
N. Shazeer , A. Mirhoseini , K. Maziarz , A. Davis , Q. Le , G. Hinton , and J. Dean , *Outrageously large neural networks: The sparsely-gated mixture-of-experts layer*, arXiv preprint arXiv:1701.06538, (2017).
N. M. Shazeer , *Fast transformer decoding: One write-head is all you need*, ArXiv, abs/1911.02150 (2019).
N. M. Shazeer , Z. Lan , Y. Cheng , N. Ding , and L. Hou , *Talking-heads attention*, ArXiv, abs/2003.02436 (2020).
A. Shrikumar , P. Greenside , and A. Kundaje , *Learning important features through propagating activation differences* , inInternational Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.
J. Singh , B. McCann , R. Socher , and C. Xiong , *Bert is not an interlingua and the bias of tokenization*, in Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), 2019, pp. 47–55.
R. Socher , A. Perelygin , J. Wu , J. Chuang , C. D. Manning , A. Ng , and C. Potts , *Recursive deep models for semantic compositionality over a sentiment treebank*, in EMNLP, 2013.
K. Song , X. Tan , T. Qin , J. Lu , and T.-Y. Liu , *Mass: Masked sequence to sequence pre-training for language generation*, in ICML, 2019.
J. Su , Y. Lu , S. Pan , B. Wen , and Y. Liu , *Roformer: Enhanced transformer with rotary position embedding*, ArXiv, abs/2104.09864 (2021).
J. Su , D. V. Vargas , and K. Sakurai , *One pixel attack for fooling deep neural networks* , IEEE Transactions on Evolutionary Computation, 23 (2019), pp. 828–841.
S. Subramanian , R. Collobert , M. Ranzato , and Y.-L. Boureau , *Multi-scale transformer language models*, ArXiv, abs/2005.00581 (2020).
S. Sukhbaatar , E. Grave , P. Bojanowski , and A. Joulin , *Adaptive attention span in transformers*, in ACL, 2019.
M. Sundararajan , A. Taly , and Q. Yan , *Axiomatic attribution for deep networks* , in International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.
I. Sutskever , *Training recurrent neural networks*, Ph.D. Thesis from University of Toronto, Toronto, Ont., Canada, (2013).
I. Sutskever , O. Vinyals , and Q. V. Le , *Sequence to sequence learning with neural networks* , in Advances in neural information processing systems, 2014, pp. 3104–3112.
R. S. Sutton and A. G. Barto , *Reinforcement learning: An introduction* , IEEE Transactions on Neural Networks, 16 (2005), pp. 285–286.
S. Tan , R. Caruana , G. Hooker , P. Koch , and A. Gordo , *Learning global additive explanations for neural nets using model distillation*, arXiv preprint arXiv:1801.08640, (2018).
Y. Tay , D. Bahri , D. Metzler , D.-C. Juan , Z. Zhao , and C. Zheng , *Synthesizer: Rethinking self-attention in transformer models*, ArXiv, abs/2005.00743 (2021).
Y. Tay , D. Bahri , L. Yang , D. Metzler , and D.-C. Juan , *Sparse Sinkhorn attention*, in ICML, 2020.
Y. Tay , M. Dehghani , D. Bahri , and D. Metzler , *Efficient transformers: A survey*, ArXiv, abs/2009.06732 (2020).
I. Tenney , D. Das , and E. Pavlick , *BERT rediscovers the classical NLP pipeline*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, July 2019, pp. 4593–4601.
I. Tenney , P. Xia , B. Chen , A. Wang , A. Poliak , R. T. McCoy , N. Kim , B. Van Durme , S. R. Bowman , D. Das , et al., *What do you learn from context? probing for sentence structure in contextualized word representations*, arXiv preprint arXiv:1905.06316, (2019).
E. F. Tjong Kim Sang , *Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition* , in COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002), 2002.
E. F. Tjong Kim Sang and F. De Meulder , *Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition*, in Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147.
J. L. Uday Kamath and J. Whitaker , *Deep Learning for NLP and Speech Recognition*, 2019.
V. Valkov , *Sentiment Analysis with BERT and Transformers by Hugging Face using PyTorch and Python*. https://bit.ly/32Mb2mw, 2020.
A. van den Oord , Y. Li , and O. Vinyals , *Representation learning with contrastive predictive coding*, ArXiv, abs/1807.03748 (2018).
A. van den Oord , Y. Li , and O. Vinyals , *Representation learning with contrastive predictive coding*, 2019.
K. R. Varshney and H. Alemzadeh , *On the safety of machine learning: Cyber-physical systems, decision sciences, and data products* , Big data, 5 (2017), pp. 246–255.
A. Vaswani , N. Shazeer , N. Parmar , J. Uszkoreit , L. Jones , A. N. Gomez , L. Kaiser , and I. Polosukhin , *Attention is all you need*, CoRR, abs/1706.03762 (2017).
A. Vaswani , N. M. Shazeer , N. Parmar , J. Uszkoreit , L. Jones , A. N. Gomez , L. Kaiser , and I. Polosukhin , *Attention is all you need*, ArXiv, abs/1706.03762 (2017).
J. Vig , *A multiscale visualization of attention in the transformer model*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 37–42.
A. Vyas , A. Katharopoulos , and F. Fleuret , *Fast transformers with clustered attention*, ArXiv, abs/2007.04825 (2020).
A. Wang , A. Singh , J. Michael , F. Hill , O. Levy , and S. Bowman , *GLUE: A multi-task benchmark and analysis platform for natural language understanding*, in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, November 2018, Association for Computational Linguistics, pp. 353–355.
C. Wang , A. Wu , J. M. Pino , A. Baevski , M. Auli , and A. Conneau , *Large-scale self- and semi-supervised learning for speech translation*, ArXiv, abs/2104.06678 (2021).

S. Wang , B. Z. Li , M. Khabsa , H. Fang , and H. Ma , *Linformer: Self-attention with linear complexity*, ArXiv, abs/2006.04768 (2020).
X. Wang , Y. Jiang , N. Bach , T. Wang , F. Huang , and K. Tu , *Structure-level knowledge distillation for multilingual sequence labeling*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, D. Jurafsky , J. Chai , N. Schluter , and J.R. Tetreault , eds., Association for Computational Linguistics, 2020, pp. 3317–3330.
Y. Wang , W. Che , J. Guo , Y. Liu , and T. Liu , *Cross-lingual BERT transformation for zero-shot dependency parsing*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, K. Inui , J. Jiang , V. Ng , and X. Wan , eds., Association for Computational Linguistics, 2019, pp. 5720–5726.
Z. Wang , Y. Ma , Z. Liu , and J. Tang , *R-transformer: Recurrent neural network enhanced transformer*, ArXiv, abs/1907.05572 (2019).
Z. Wang , J. Xie , R. Xu , Y. Yang , G. Neubig , and J. G. Carbonell , *Cross-lingual alignment vs joint training: A comparative study and A simple unified framework* , in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020.
A. Warstadt , A. Singh , and S. R. Bowman , *Neural network acceptability judgments*, arXiv preprint arXiv:1805.12471, (2018).
D. Watts and S. Strogatz , *Collective dynamics of 'small-world' networks* , Nature, 393 (1998), pp. 440–442.
X. Wei , R. Weng , Y. Hu , L. Xing , H. Yu , and W. Luo , *On learning universal representations across languages* , in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021, OpenReview.net, 2021.
P. J. Werbos , *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, PhD thesis, Harvard University, 1974.
J. Whittlestone , R. Nyrup , A. Alexandrova , and S. Cave , *The role and limits of principles in ai ethics: towards a focus on tensions*, in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 195–200.
S. Wiegreffe and Y. Pinter , *Attention is not not explanation*, arXiv preprint arXiv:1908.04626, (2019).
C. Wu , F. Wu , T. Qi , and Y. Huang , *Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling*, ArXiv, abs/2106.01040 (2021).
S. Wu and M. Dredze , *Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, K. Inui , J. Jiang , V. Ng , and X. Wan , eds., Association for Computational Linguistics, 2019, pp. 833–844.
Y. Wu , M. Schuster , Z. Chen , Q. V. Le , M. Norouzi , W. Macherey , M. Krikun , Y. Cao , Q. Gao , K. Macherey , et al., *Google's neural machine translation system: Bridging the gap between human and machine translation*, arXiv preprint arXiv:1609.08144, (2016).
Y. Wu , M. Schuster , Z. Chen , Q. V. Le , M. Norouzi , W. Macherey , M. Krikun , Y. Cao , Q. Gao , K. Macherey , J. Klingner , A. Shah , M. Johnson , X. Liu , L. Kaiser , S. Gouws , Y. Kato , T. Kudo , H. Kazawa , K. Stevens , G. Kurian , N. Patil , W. Wang , C. Young , J. Smith , J. Riesa , A. Rudnick , O. Vinyals , G. Corrado , M. Hughes , and J. Dean , *Google's neural machine translation system: Bridging the gap between human and machine translation*, CoRR, abs/1609.08144 (2016).
Z. Wu , Z. Liu , J. Lin , Y. Lin , and S. Han , *Lite transformer with long-short range attention*, ArXiv, abs/2004.11886 (2020).
N. Xie , G. Ras , M. van Gerven , and D. Doran , *Explainable deep learning: A field guide for the uninitiated*, arXiv preprint arXiv:2004.14545, (2020).
J. Xin , R. Tang , J. Lee , Y. Yu , and J. J. Lin , *DeeBERT: Dynamic early exiting for accelerating bert inference*, in ACL, 2020.
Y. Xiong , Z. Zeng , R. Chakraborty , M. Tan , G. M. Fung , Y. Li , and V. Singh , *Nyströmformer: A Nyström-based algorithm for approximating self-attention*, in AAAI, 2021.
K. Xu , J. Ba , R. Kiros , K. Cho , A. Courville , R. Salakhudinov , R. Zemel , and Y. Bengio , *Show, attend and tell: Neural image caption generation with visual attention* , in International conference on machine learning, PMLR, 2015, pp. 2048–2057.
M. Xu , L. yu Duan , J. Cai , L. Chia , C. Xu , and Q. Tian , *HMM-based Audio Keyword Generation*, in PCM, 2004.
L. Xue , A. Barua , N. Constant , R. Al-Rfou , S. Narang , M. Kale , A. Roberts , and C. Raffel , *ByT5: Towards a token-free future with pre-trained byte-to-byte models*, ArXiv, abs/2105.13626 (2021).
L. Xue , N. Constant , A. Roberts , M. Kale , R. Al-Rfou , A. Siddhant , A. Barua , and C. Raffel , *mT5: A massively multilingual pre-trained text-to-text transformer*, in NAACL, 2021.
J. Yang , S. Ma , D. Zhang , S. Wu , Z. Li , and M. Zhou , *Alternating language modeling for cross-lingual pre-training* , Proceedings of the AAAI Conference on Artificial Intelligence, 34 (2020), pp. 9386–9393.
Y. Yang , G. H. Ábrego , S. Yuan , M. Guo , Q. Shen , D. Cer , Y. Sung , B. Strope , and R. Kurzweil , *Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax*, CoRR, abs/1902.08564 (2019).
Y. Yang , D. M. Cer , A. Ahmad , M. Guo , J. Law , N. Constant , G. H. Ábrego , S. Yuan , C. Tar , Y.-H. Sung , B. Strope , and R. Kurzweil , *Multilingual universal sentence encoder for semantic retrieval*, in ACL, 2020.
Y. Yang , Y. Zhang , C. Tar , and J. Baldridge , *PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification*, in Proc. of EMNLP, 2019.
Z. Yang , D. Yang , C. Dyer , X. He , A. Smola , and E. Hovy , *Hierarchical attention networks for document classification*, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.
Q. yang Wu , Z. Lan , J. Gu , and Z. Yu , *Memformer: The memory-augmented transformer*, ArXiv, abs/2010.06891 (2020).
J. Yim , D. Joo , J. Bae , and J. Kim , *A gift from knowledge distillation: Fast optimization, network minimization and transfer learning*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4133–4141.
P. Yin , G. Neubig , W.-t. Yih , and S. Riedel , *TaBERT: Pretraining for joint understanding of textual and tabular data*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 8413–8426.
C. ying Lee and J. R. Glass , *A nonparametric Bayesian approach to acoustic model discovery*, in ACL, 2012.
X. Yuan , P. He , Q. Zhu , and X. Li , *Adversarial examples: Attacks and defenses for deep learning* , IEEE transactions on neural networks and learning systems, 30 (2019), pp. 2805–2824.
M. Zaheer , G. Guruganesh , K. A. Dubey , J. Ainslie , C. Alberti , S. Ontañón , P. Pham , A. Ravula , Q. Wang , L. Yang , and A. Ahmed , *Big bird: Transformers for longer sequences*, ArXiv, abs/2007.14062 (2020).
R. Zellers , Y. Bisk , A. Farhadi , and Y. Choi , *From recognition to cognition: Visual commonsense reasoning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6720–6731.

Q. Zhang , R. Cao , Y. N. Wu , and S.-C. Zhu , *Growing interpretable part graphs on ConvNets via multi-shot learning*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, 2017.
W. E. Zhang , Q. Z. Sheng , A. A. F. Alhazmi , and C. Li , *Generating textual adversarial examples for deep learning models: A survey*, CoRR, abs/1901.06796, (2019).
X. Zhang , F. Wei , and M. Zhou , *Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization*, in ACL, 2019.
X. Zhang , J. J. Zhao , and Y. A. LeCun , *Character-level convolutional networks for text classification*, ArXiv, abs/1509.01626 (2015).
W. Zhao , S. Eger , J. Bjerva , and I. Augenstein , *Inducing language-agnostic multilingual representations*, CoRR, abs/2008.09112 (2020).
Y. Zhao , L. Dong , Y. Shen , Z. Zhang , F. Wei , and W. Chen , *Memory-efficient differentiable transformer architecture search*, in FINDINGS, 2021.
W. Zheng , Z. Chen , J. Lu , and J. Zhou , *Hardness-aware deep metric learning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
B. Zhou , A. Khosla , A. Lapedriza , A. Oliva , and A. Torralba , *Object detectors emerge in deep scene cnns*, arXiv preprint arXiv:1412.6856, (2014).
H. Zhou , S. Zhang , J. Peng , S. Zhang , J. Li , H. Xiong , and W. Zhang , *Informer: Beyond efficient transformer for long sequence time-series forecasting*, in AAAI, 2021.
Y. Zhu , R. Kiros , R. S. Zemel , R. Salakhutdinov , R. Urtasun , A. Torralba , and S. Fidler , *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books* , 2015 IEEE International Conference on Computer Vision (ICCV), (2015), pp. 19–27.
L. M. Zintgraf , T. S. Cohen , T. Adel , and M. Welling , *Visualizing deep neural network decisions: Prediction difference analysis*, arXiv preprint arXiv:1702.04595, (2017).
P. Zweigenbaum , S. Sharoff , and R. Rapp , *Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora*, in Proceedings of the 10th Workshop on Building and Using Comparable Corpora, Vancouver, Canada, August 2017, Association for Computational Linguistics, pp. 60–67.