Rethinking Data-driven Networking with Foundation Models: Challenges and Opportunities

Franck Le
IBM Research

Mudhakar Srivatsa
IBM Research

Raghu Ganti
IBM Research

Vyas Sekar
Carnegie Mellon University

ABSTRACT

Foundational models have caused a paradigm shift in the way artificial intelligence (AI) systems are built. They have had a major impact in natural language processing (NLP), and several other domains, not only reducing the amount of required labeled data or even eliminating the need for it, but also significantly improving performance on a wide range of tasks. We argue foundation models can have a similar profound impact on network traffic analysis, and management. More specifically, we show that network data shares several of the properties that are behind the success of foundational models in linguistics. For example, network data contains rich semantic content, and several of the networking tasks (e.g., traffic classification, generation of protocol implementations from specification text, anomaly detection) can find similar counterparts in NLP (e.g., sentiment analysis, translation from natural language to code, out-of-distribution). However, network settings also present unique characteristics and challenges that must be overcome. Our contribution is in highlighting the opportunities and challenges at the intersection of foundation models and networking.

CCS CONCEPTS

• **Networks** → **Network management**;

KEYWORDS

Machine learning, foundational models, network management and security.

HotNets '22, November 14–15, 2022, Austin, TX, USA

ACM ISBN 978-1-4503-9899-2/22/11…$15.00
https://doi.org/10.1145/3563766.3564109

ACM Reference Format:
Franck Le, Mudhakar Srivatsa, Raghu Ganti, and Vyas Sekar. 2022. Rethinking Data-driven Networking with Foundation Models: Challenges and Opportunities. In *The 21st ACM Workshop on Hot Topics in Networks (HotNets '22), November 14–15, 2022, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3563766.3564109

1 INTRODUCTION

Foundational models [3] describe the class of models that are behind the current paradigm shift for building artificial intelligence (AI) systems where neural networks that are trained on large corpora of unlabeled data, and then adapted to a wide range of downstream tasks with minimal fine-tuning. For example, the BERT (Bidirectional Encoder Representations from Transformers) model [14] was pre-trained on large unlabeled text corpora, before being fine-tuned and achieving state-of-the-art performance on eleven downstream tasks (e.g., classification, similarity and paraphrase, inference). Variants of BERT have since been developed, and outperformed state-of-the-art solutions [46, 67, 94].

In addition to the substantial performance increases, foundational models significantly reduce and even eliminate the need for data labeling, a process often considered tedious, error-prone, and expensive. The pre-training phase which is the most compute intensive phase is performed on unlabeled data in a self-supervised manner, and only the fine-tuning phase requires a small amount of labeled instances. Intuitively, in the pre-training phase, the models extract general useful features from raw text because of the large volume of unlabeled data that is usually available much more readily than labeled data (e.g., the common crawl corpus has billions of tokens and covers multiple languages). In this pre-training phase, the model learns numerical representations for words that capture semantic information, and relational knowledge behind them. For example, words belonging to a same lexical field, such as *Human, Face, Speech, Body*, have representations that tend to be closer [81]. In addition, foundational models learn to discern the different meanings a same word can have based on its context. For example, they will generate different numerical representations for the same word *die*

in the following two sentences: "Many more *die* from radiation sickness, starvation and cold", and "Players must always move a token according to the *die* value" [62]. Then, after having learnt general linguistic patterns, the models can be fine-tuned on specific downstream tasks, with considerably fewer labeled instances. GPT-3 [5], a more recent language model, even eliminates the need for fine-tuning; and reduces the amount of labeled examples further by one order or magnitude, and even eliminates the need for any labeled data for some tasks. In addition to NLP, foundational models are being applied in vision [9, 16, 59], tabular data [97], and programming languages [17, 27].

We observe that network traffic analysis, and management share a number of properties with NLP and other domains where foundational models have been very successful:

• *Similarities of downstream tasks:* First, similar to the status in NLP prior to the emergence of foundational models, specific solutions with their own preprocessing, features extraction, architectures, and datasets, are currently being developed for different tasks (e.g., congestion control [1, 24, 90], adaptive bitrate streaming [50], datacenter-scale automatic traffic optimization [7], job scheduling [31, 51, 51], resource management [65, 91, 99], network planning [104], packet classification [39], performance prediction [49], congestion prediction [56], performance estimation [100], malware detection [29, 93], mapping from a low-quality video to a high-quality version [96], or semi-automated generation of protocol implementations from specification text [95].) Next, we observe that most of the underlying adopted machine learning approach behind those solutions (e.g., classification, anomaly detection, generator, and reinforcement learning) are areas where foundational models have been successfully applied, or being explored (Section 3.1).

• *Abundance of unlabeled data:* Second, there is abundant unlabeled network data: For example, universities' networks have reported 1 to 2 TB of traffic on average per day [21, 30]; and data-center networks and content providers generate more than 10 TB on average daily [28, 38, 85] (Section 3.2).

• *Rich semantic content:* Third, network data has rich semantic information: For example, intuitively, for transport protocols, SCTP, TCP, and MPTCP may be considered more similar than UDP, since they handle congestion and packet loss recovery; for applications, HTTP and CoAP may be closer to each other than SMTP, IMAP, or POP3; for security, ciphersuites may form clusters (e.g., weak versus strong). In NLP, foundational models have been very effective in capturing semantic relationships behind, and between words [4] (Section 3.3).

We highlight early efforts that begun exploring foundational models in the networking domain, and which show promising results (Section 3.4). However, to fully realize the benefits of foundational models, we identify a number of challenges network data present, and that must be addressed:.

• *Extraction of general useful patterns from networking data:* Although network data has rich semantic content, how can we effectively extract it? First, we observe there are few dozens popular deployed network protocols (e.g., TCP, UDP, HTTP, SMTP, NTP, SIP, RTSP, DHCP, DNS, etc.) which have rich semantic relationships between and across them: For example, TCP and UDP are transport protocols, and HTTP runs on top of TCP. As such, can we define a single cross-protocol representation that capture those semantic relationships, and that can significantly improve the performance for a wide range of networking tasks (Section 4.1.1)? Next, *tokenization* (Section 4.1.2), *context* (Section 4.1.3), and *pre-training tasks* (Section 4.1.4) have played critical roles towards effectively extracting semantic relationships in NLP. What would their counterparts for networking data be, especially given its unique characteristics: e.g., packets from different connections may be interleaved, or fields (e.g., HTTP User-Agent, DNS answers) may contain different data structures (e.g., text, sets)?

• *Publicly available datasets:* The public availability of large text corpus (e.g., [60, 105], and labeled benchmarks (e.g., [61, 83, 98]) with a wide variety of downstream tasks, and groundtruth has been essential for research, and making advances in the emergence of foundational models and NLP. Although organizations can collect and analyze large amount of their own network data, can the networking community release training data available, and define networking benchmarks (Section 4.2)?

• *Dealing with rare and unseen events:* Reports raised uncertainty about the suitability of machine learning for network security, and more specifically anomaly detection [76], claiming that the task of detecting attacks may be fundamentally different from other applications where machine learning have been successfully deployed. Can foundational models therefore effectively help detect zero-day attacks as well as unusual behaviors (Section 4.3)?

• *Interpretability:* Understanding the reasons behind a model's prediction outcome allows users to validate models, and increases users' trust in the models [33, 103]. Researchers have argued for interpretation methods specifically developed for networking models [52] given the unique nature of networking inputs. Can we derive meaningful explanations for foundational models when applied to networking (Section 4.4)?

We elaborate on these challenges and identify opportunities to tackle them. Our paper is a "call to arms" for exploring the potential benefits of foundational models to the domain of networking. As such, our goal here is to ignite this discussion, and we acknowledge that we raise more questions than provide concrete answers.

2 BACKGROUND

The goal of this section is to given an overview on foundational models. More specifically, we describe how BERT models are trained, and how they learn numerical representations (i.e., high-dimensional vectors) – also called *embeddings* – for words. However, before going into the details of BERT models, we first briefly discuss why embeddings were introduced, and the solutions that were first developed.

Motivation: Neural networks cannot operate on text (e.g., string) directly. They require all inputs to be numerical. As such, text has to be converted into numerical values. One hot encoding, a popular method for categorical variables, assigns each value (e.g., "conference", "workshop", "journal") to a binary vector (e.g., 00...0001, 00...0010, 00...0100). As such, every pair of values is equidistant. Instead, for text, words with similar meanings should have similar vectors.

Word2Vec: To satisfy this requirement, Word2Vec was introduced in 2013 [53]. Relying on two neural network variants, it computes word embeddings based on the words' context: Continuous Bag-of-Words (CBOW) predicts the current word based on the context; and Skip-gram instead predicts the closely related context words to an input word. Word2Vec can as such learn high dimensional (e.g., 50, 100, 300, 600) vector representation for words, and was shown to learn very subtle semantic relationships between words, such as a currency and the name of the country that uses it, e.g. Angola is to kwanza as Iran is to rial [53], or that "King - Man + Woman" resulted in a vector very close to "Queen" [54].

BERT: In 2018, Devlin *et al.* presented BERT (Bidirectional Encoder Representations from Transformers), a new language representation model [14]. In contrast to Word2Vec which computes context-independent embeddings, BERT generates contextual embeddings. To illustrate the differences, we consider the two following sentences: "*Bark* is essential for a tree's survival", and "There can be many reasons behind a dog's *bark*". Word2Vec would generate the same vector representation for the word *bark* – independently of its position and meaning – in both sentences. In contrast, BERT would generate different vector representations for each occurrence of *bark* because their position, and surrounding words differ.

More generally, training BERT relies on two stages: pre-training and fine-tuning:

• *Pre-training*: The model is trained on unlabeled data over two pre-training tasks. In the first task, called, Masked Language Modeling, a fraction of the input tokens are randomly masked, and the goal is to predict those masked tokens. In the second task, Next Sentence Prediction, two sentences, `A` and `B`, are provided as inputs to the model, and the goal is to predict whether `B` is the actual sentence that follows `A` in the initial corpus, or a random sentence from the corpus.

• *Fine-tuning*: To fine-tune a BERT model to a downstream task, an additional layer is typically added to the pre-trained model, and then the entire model is trained over the labeled data for few epochs.

GPT-3: Generative Pre-trained Transformer 3 (GPT-3) is Transformer-based language model with 175 billion parameters, developed by OpenAI in 2020 [5]. It further reduces the amount of required labeled data during the fine-tuning stage, from thousands or tens of thousands of examples, to tens or hundreds of examples (few shot learning), one single example (one shot learning), or no example and only an instruction in natural language is given to the model (zero shot learning). Contrary to the fine tuning stage, GPT-3 does not perform any gradient update. Instead, the model applies "in-context learning", where the model is simply conditioned on a natural language instruction and/or a few examples of the task. More specifically, as input (also, called the prompt), the model is provided the instructions and/or labeled examples. In return, the model generates a text completion: As illustrated in the Open AI API documentation, given the prompt, "Write a tagline for an ice cream shop", the model returns the following completion: "We serve up smiles with every scoop!" In NLP tasks, GPT-3 demonstrated promising results in zero-shot and one-shot settings; and in few-shot setting, it was sometimes competitive with or even occasionally outperformed state-of-the-art models.

3 A CASE FOR NETWORK FOUNDATIONAL MODELS

We point out similarities between networking, and other domains (e.g., NLP) where foundational models had a large impact; and recently published early work that provide corroborating evidence of the potential of foundational models for the networking domain.

3.1 Range of downstream tasks

A wide range of network downstream tasks can benefit from foundational models. First, we observe that machine learning solutions have been developed for different network downstream tasks, including towards congestion control [1, 24, 90], adaptive bitrate streaming [50], datacenter-scale automatic traffic optimization [7], job scheduling [31, 51, 51], resource management [65, 91, 99], network planning [104], packet classification [39], performance prediction [49], congestion prediction [56], performance estimation [100], malware detection [29, 93] and , semi-automated generation of protocol implementations from specification text [95].

Next, we note that those solutions can be classified by the underlying adopted machine learning approach such as classification, anomaly detection, generator, and reinforcement learning; and most are areas where foundational models

have been very successfully applied, or where foundational models are currently being expanded to: For example, foundational models have set state-of-the-art performance for text classification [46], token classification [67], and text generation [5]; and foundational models have recently also been applied to reinforcement learning problems [8, 23], or translation from natural language to code [10]. Codex [10] is a GPT-3 language model fine-tuned on publicly available code from GitHub. Given a coding task in natural language as a prompt, the model returns blocks of code that satisfies it. This task can be considered to be similar to the network goal of generating protocol implementations from specification text [95].

3.2 Abundant unlabeled data

Unlabeled data is plentiful in network: For example, universities' networks carry 1 to 2 TB of traffic on average per day [21, 30]; and data-center networks and content providers generate more than 10 TB of traffic on average daily [28, 38, 85]. In comparison, BERT was trained on 16 GB of Books Corpus and English Wikipedia [14], RoBERTa uses 160 GB of text for pre-training [46], XLM-RoBERTa was trained from 2.5 TB of text [13], and the OpenAI GPT-3 model was trained on 45TB of text [5].

3.3 Rich semantic content

We envision foundational models to be applied to network data, and we argue that similar to text, network data has rich semantic content that the pre-training phase would be able to extract, and make it useful to a wide range of downstream tasks. More specifically, a packet trace can be viewed as a sequence of variables, some of which can be categorical, and others numerical. For example, the *total length* field in an IP header is a 16-bit numerical variable that indicates the entire size of the IP packet (header and data) in bytes. In contrast, the *protocol field* in the IP header is an 8-bit categorical variable that indicates the next protocol inside the IP packet. Possible values include TCP, UDP, ICMP, SCTP, EIGRP, DSR, IPv4, IPv6, and GRE; and one can observe that they can form semantic clusters, e.g., with TCP, UDP and SCTP being transport protocols, EIGRP and DSR being routing protocols, and IPv4, IPv6, and GRE indicating tunneling. Another example of categorical variable with rich semantic information is the DNS query field. Values may indicate mail servers (e.g., gmail.com, outlook.com), repository servers, time servers (e.g., time.nist.gov, ntp.org), news sites (e.g., npr.com, nytimes.com), or video streaming sites (e.g., netflix.com, primevideo.com).

3.4 Early successes

We summarize results from some early work exploring the potential of foundational models for the networking domain.

NetBERT: A recent study [47] provides evidence that network data includes rich semantic relationships. The authors trained a BERT model on a text corpus on computer networking. They do not apply foundational models directly to network data. However, the results reveals subtle relationships in the networking domain. For example, similar to "Man is to King as Woman is to Queen" (Section 2), the authors confirmed several similar analogies in the network domain including "BGP is to router as STP is to switch", "MAC is to switch as IP is to router", or "IP is to network as TCP is to transport".

NorBERT: A more recent study explored the adaptation of foundational models on network data, and preliminary results provide additional evidence of the semantic richness of network data, and performance improvement for networking downstream tasks [34].

• *Semantic relationships in networking data:* Adapting foundational models on networking data, the authors revealed interesting relationships between tokens' embeddings. For example, the closest neighbor to the token 80 (HTTP), was the token 443 (HTTPS); and the closest neighbor to the token 49199 (ciphersuite "ECDHE + RSA authentication AES-128 GCM SHA-256"), is token 49200 (ciphersuite "ECDHE + RSA authentication AES-256 GCM SHA-384"). These two ciphersuites differ only in the keys' lengths. The closeness of these tokens is according to intuition, and domain knowledge.

• *Performance improvement:* The authors also compared the performance of foundational models for downstream classification tasks. The authors pre-trained a foundational model (NorBERT) on DNS traffic, fine-tuned it on a labeled dataset, and evaluated its performance on an independent labeled dataset. The performance are compared with those of gated recurrent units (GRU) models, with both initialization to random values, and context-independent embeddings (GloVe) [55]. The performance of the GRU models drop considerably on the validation dataset (F-1 scores between 0.585 and 0.726). In contrast, the performance of NorBERT remains above 0.9, demonstrating significant performance improvements.

4 CHALLENGES AND OPPORTUNITIES

Despite the promising results of early efforts, we identify broader challenges that network data presents given their unique characteristics. More specifically, NetBERT essentially provided evidence that the networking domain contains rich semanctic relationships, but the study applied foundational models on text related to networking, and not networking data; and NorBERT demonstrated significant performance gains, but focused on two network downstream

classification tasks. For foundational models to be applicable, and useful to a broader range of networking downstream tasks, we identify challenges, and opportunities to tackle them.

4.1 Extraction of general useful patterns

4.1.1 Common Representation. Akin to languages, there are a hundred popularly used network protocols. A network protocol is a language of communication between two entities (e.g., client and server). An intrinsic property of a language is that an utterance between two entities at a certain point in space and time, bears a semblance to utterances between two other entities at an entirely different points in space and time. Consider a client and a web server communicating via the HTTP protocol; interactions between any pair of client and web server using the HTTP protocol (language) bears similarities between each other. For example, given a certain utterance from the client (e.g., HTTP GET), there are certain sets of valid utterances (responses) from the web server (e.g., STATUS 200). A wider context such as knowing the HTTP User Agent type or the size of the HTTP response, helps predict future utterances by both the client and server.

One key distinction between network protocols and natural languages is that network protocols are almost always multiparty (at least two) communication. The topology between the parties involved is an important aspect of the network protocol. Many of the foundation models in NLP are focused on monologues (e.g., document) or two party communication (e.g., chat bot). However, it is very common for an application to span multiple network protocols, and to involved multiple servers, to achieve a single task. Hence, a natural first step is for us to learn common representations within a single network protocol and then expand the foundation model to the multi-lingual domain that captures multi-party, multi-protocol applications (akin to the evolution in NLP domain from Roberta [46] to XLM-Roberta [13], a transformer-based multilingual masked language model pre-trained on text in 100 languages.)

4.1.2 Tokenizer. Tokenization is an important step in text pre-processing. It splits a piece of text (e.g., phrase, sentence, paragraph) into smaller units, called tokens. Tokens represent the smallest semantic unit of interest. A token can be a word, a subword, or even a character. Foundational models frequently use subword-based tokenization algorithms. More specifically, BERT [14] uses WordPiece [68], and RoBERTa [46] uses Byte-Pair Encoding [69]. Text is split into words based on delimiters, with common ones being space and punctuation; and to reduce the vocabulary size, to handle rare words, and to learn meaningful representations, words may be split into sub-words. For example WordPiece would split the word '`Hotnets`" into two tokens ['`hot`', '`##nets`']. Studies have revealed that subword tokenization plays an important role in the relationships that models learn from corpus [71].

However, with packet traces being often viewed as sequences of bytes, with no clear delimiters such as white spaces, and punctuation, how should network data therefore get tokenized? One approach could consist in applying character-based tokenizers [26, 35, 58]. Another approach may consist in recognizing the network protocol (language) and tokenizing it based on protocol format (e.g., 4 byte IP address, 2 byte port number, one byte TCP flag, HTTP fields, etc.). This would preserve the semantics of the tokens as per the underlying network protocol specifications.

4.1.3 Context. Word embeddings are based on the premise that words that occur in similar contexts tend to have similar meanings. For example, Word2Vec trains two neural networks to predict a word given its context and vice-versa [53]; and foundational models are generally trained using a masking approach: The key idea is that given a sequence of tokens (often called the context), a random percentage (typically 15%) of the tokens are masked and the model is trained to reconstruct the masked tokens. A byproduct of this training process is embeddings, which find application in various downstream tasks. This introduces the challenge of defining a context around tokens.

First, it may be possible to define a context based on packet boundaries (shorter context) or session boundaries (wider context). In addition, we observe that at a point of packet capture (e.g., border router), packets from different end points may be interleaved. Even when focusing on traffic from and to individual end points, their traffic may consist of packets belonging to concurrent connections. Focusing on individual connections may lose semantic relationships between connections, especially given that a transaction may consist of concurrent and sequential connections. Third, practical constraints may limit the size of context to about 512 tokens. As such, it may be vital to construct non-standard contexts over network protocols: e.g., use the first M tokens from each of the N successive IP packets sent or received from an endpoint as a context.

4.1.4 Pre-training tasks. What pre-training tasks would be most effective for network data? BERT defined two pre-training tasks: Masked language modeling, and next sentence prediction (Section 2). Since then, a number of studies explored alternative pre-training tasks [11, 12, 17, 20, 25, 42, 57, 79, 83, 84, 101, 102].

Given the network-specific downstream tasks (Section 3.1), and characteristics of network data (Sections 4.1.1, and 4.1.3), new network-specific training tasks may need to be defined. For example, query-answers are common transactions in computer networks, and new training tasks may be required

to capture the nature of the relationships between a query and its answers: In DNS, the answers may be viewed as the children of the query in a hierarchical tree. In addition, network fields may consist of different structures, such as sets where the order does not matter. An example is the DNS answer field, where multiple values may be returned for a single query. Such structures (e.g., list, set) may reflect stronger similarities between their members.

4.2 Publicly available data and benchmarks

The public availability of large text corpus (e.g., [60, 105], and labeled benchmarks (e.g., [61, 83, 98]) are one of the main reasons behind the active and prolific research in foundational models and NLP. For example, the General Language Understanding Evaluation (GLUE) benchmark [83] includes a collection of natural language understanding tasks including classification tasks (e.g., is a sentence grammatically acceptable? [86], or is a review positive? [75]), similarity and paraphrase tasks [6, 15] (e.g., are two questions semantically equivalent?), and inference tasks [37, 61] (e.g., does a premise sentence entail or contradict a hypothesis sentence? [89]).

Although network data is abundant (Section 3.2), and organizations can easily collect their own, concerns of leaking sensitive content have limited their public release. To address the lack of public networking data, researchers set up small private labs with a variety of devices, collect their traffic, and publicly release the packet captures (e.g., [72]). However, can the networking community offer larger amount of network data, and define benchmarks to facilitate research? Synthetic packet traces generators [64, 77, 82, 87, 92] may be one solution for mitigating the privacy concerns, and training foundational models on network data. Benchmarks could comprise a dozen of network downstream tasks including device classification, flow classification, performance prediction, congestion prediction, malware detection.

4.3 Rare and unseen events

Sommer and Paxson [76] reported that contrary to several other domains where machine learning was successfully commercially deployed (e.g., recommendation systems [2, 41], optical character recognition systems [74]), machine learning was rarely deployed in operational network settings for anomaly detection. The authors argue that that "*the strength of machine-learning tools is finding activity that is similar to something previously seen*", and not to find "*novel attacks*".

However, methods have recently been developed specifically to detect out-of-distribution instances, i.e., instances that differ from those seen during training (e.g., [18, 32, 36, 36, 40, 40, 43, 45]), and machine learning algorithms are now commercially deployed to identify defects (e.g., in car assembly line) where the costs of errors are also high. In other words, while machine learning may not have been suitable for network intrusion detection, recent advances in out-of-distribution may help effectively identify zero-day attacks, and unusual network behaviors.

4.4 Interpretability

Interpretability is critical towards validating models, and increasing users' trust in the models [33, 103]. As such, a large amount of research effort has been devoted towards understanding the reasons behind a model's prediction outcome (e.g., [22, 33, 44, 48, 63, 70, 73, 80, 88]). However, machine learning solutions for networking downstream tasks often work with domain specific inputs. For example, solutions may take a topology as input, and return routing paths [66]. Therefore, researchers have argued for interpretation methods specifically developed for networking models [52]. Similarly, we argue that interpretability methods specifically developed for foundational models applied to networking may be required. For example, if networking data is tokenized at the character level (Section 4.1.2), how can we derive meaningful explanations? To draw an analogy with computer vision, the notion of superpixels (i.e., set of adjacent pixels with similar color and underlying properties) [78] has allowed more meaningful features and explanations [19, 63].

4.5 Other Issues

The above challenges explore only a small sample of open issues. We conclude by identifying a few other open issues:

- *Energy footprint:* Large models training and inference often consume massive amount of energy, raising questions on the benefits and generability of those models.
- *Generalizability:* Although many downstream tasks can potentially benefit from common representations, we do not expect a single universal foundation model to be able to address most relevant tasks. Instead, distinct models may be required for different areas (e.g., security, resource management). As such, what is the minimum coverage set of foundation models that can cover the relevant desired tasks?
- *Learning complexity:* The dimensionality of networking data could be larger than that of text? If so, what would the required dimension of the embeddings be; and how much training data would be required?

ACKNOWLEDGMENTS

We thank the reviewers for their insightful feedback.

REFERENCES

[1] Soheil Abbasloo, Chen-Yu Yen, and H. Jonathan Chao. 2020. Classic Meets Modern: A Pragmatic Learning-Based Congestion Control for the Internet. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '20)*. Association for Computing Machinery, New York, NY, USA, 632–647. https://doi.org/10.1145/3387514.3405892

[2] J. Bennett and S. Lanning. 2007. The Netflix Prize. In *Proceedings of the KDD Cup Workshop 2007*. ACM, New York, 3–6. http://www.cs.uic.edu/~liub/KDD-cup-2007/NetflixPrize-description.pdf

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[4] Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing Relational Knowledge from BERT.. In *AAAI*. AAAI Press, 7456–7463. http://dblp.uni-trier.de/db/conf/aaai/aaai2020.html#BouraouiCS20

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. https://doi.org/10.18653/v1/S17-2001

[7] Li Chen, Justinas Lingys, Kai Chen, and Feng Liu. 2018. AuTO: Scaling Deep Reinforcement Learning for Datacenter-Scale Automatic Traffic Optimization. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '18)*. Association for Computing Machinery, New York, NY, USA, 191–205. https://doi.org/10.1145/3230543.3230551

[8] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative Pretraining From Pixels. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Hal Daumé III and Aarti Singh (Eds.), Vol. 119. PMLR, 1691–1703. https://proceedings.mlr.press/v119/chen20s.html

[10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[11] Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-Fine Entity Typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 87–96. https://doi.org/10.18653/v1/P18-1009

[12] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1xMH1BtvB

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. https://aclanthology.org/I05-5002

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[17] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1536–1547. https://doi.org/10.18653/v1/2020.findings-emnlp.139

[18] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, 1050–1059.

[19] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems* 32 (2019).

[20] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4803–4809. https://doi.org/10.18653/v1/D18-1514

[21] Seong-Cheol Hong, Jin Kim, Byungchul Park, Young J. Won, and James W. Hong. 2009. Traffic growth analysis over three years in enterprise networks. In *2009 15th Asia-Pacific Conference on Communications*. 896–899. https://doi.org/10.1109/APCC.2009.5375520

[22] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).

[23] Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems* 34 (2021), 1273–1286.

[24] Nathan Jay, Noga Rotman, Brighten Godfrey, Michael Schapira, and Aviv Tamar. 2019. A Deep Reinforcement Learning Perspective on Internet Congestion Control. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 3050–3059. https://proceedings.mlr.press/v97/jay19a.html

[25] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* 8 (2020), 64–77.

[26] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099* (2016).

[27] Aditya Kanade, Petros Maniatis, Gogul Balakrishnan, and Kensen Shi. 2020. Learning and Evaluating Contextual Embedding of Source Code. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Hal Daumé III and Aarti Singh (Eds.), Vol. 119. PMLR, 5110–5121. https://proceedings.mlr.press/v119/kanade20a.html

[28] Srikanth Kandula, Sudipta Sengupta, Albert Greenberg, Parveen Patel, and Ronnie Chaiken. 2009. The Nature of Data Center Traffic: Measurements & Analysis. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement (IMC '09)*. Association for Computing Machinery, New York, NY, USA, 202–208. https://doi.org/10.1145/1644893.1644918

[29] Amin Kharraz, , William Robertson, and Engin Kirda. 2018. Surveylance: Automatically Detecting Online Survey Scams. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 70–86.

[30] Stefan Kornexl, Vern Paxson, Holger Dreger, Anja Feldmann, and Robin Sommer. 2005. Building a Time Machine for Efficient Recording and Retrieval of High-Volume Network Traffic. In *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement (IMC '05)*. USENIX Association, USA, 23.

[31] Sanjay Krishnan, Zongheng Yang, Ken Goldberg, Joseph Hellerstein, and Ion Stoica. 2018. Learning to optimize join queries with deep reinforcement learning. *arXiv preprint arXiv:1808.03196* (2018).

[32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6405–6416.

[33] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications* 10, 1 (2019), 1–8.

[34] Franck Le, Davis Wertheimer, Seraphin Calo, and Erich Nahum. 2022. NorBERT: NetwOrk Representations through BERT for Network Analysis and Management. *arXiv preprint arXiv:2206.10472* (2022). https://doi.org/10.48550/ARXIV.2206.10472

[35] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics* 5 (2017), 365–378.

[36] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 7167–7177.

[37] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. AAAI Press, Rome, Italy, 552–561. https://cs.nyu.edu/faculty/davise/papers/WSKR2012.pdf

[38] Yuheng Li, Yiping Zhang, and Ruixi Yuan. 2011. Measurement and Analysis of a Large Scale Commercial Mobile Internet TV System. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC '11)*. Association for Computing Machinery, New York, NY, USA, 209–224. https://doi.org/10.1145/2068816.2068837

[39] Eric Liang, Hang Zhu, Xin Jin, and Ion Stoica. 2019. Neural Packet Classification. In *Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM '19)*. Association for Computing Machinery, New York, NY, USA, 256–269. https://doi.org/10.1145/3341302.3342221

[40] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690* (2017).

[41] G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* 7, 1 (2003), 76–80. https://doi.org/10.1109/MIC.2003.1167344

[42] Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design Challenges for Entity Linking. *Transactions of the Association for Computational Linguistics* 3 (2015), 315–328. https://doi.org/10.1162/tacl_a_00141

[43] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems* 33 (2020), 7498–7512.

[44] Shengzhong Liu, Franck Le, Supriyo Chakraborty, and Tarek Abdelzaher. 2021. On Exploring Attention-based Explanation for Transformer Models in Text Classification. In *2021 IEEE International Conference on Big Data (Big Data)*. 1193–1203. https://doi.org/10.1109/BigData52589.2021.9671639

[45] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-Based out-of-Distribution Detection. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1802, 12 pages.

[46] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[47] Antoine Louis. 2020. *NetBERT: A Pre-trained Language Representation Model for Computer Networking*. Master's thesis. University of Liège, Liège, Belgium. http://hdl.handle.net/2268.2/9060

[48] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[49] Antonis Manousis, Rahul Anand Sharma, Vyas Sekar, and Justine Sherry. 2020. Contention-Aware Performance Prediction For Virtualized Network Functions. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '20)*. Association for Computing Machinery, New York, NY, USA, 270–282. https://doi.org/10.1145/3387514.3405868

[50] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*. Association for Computing Machinery, New York, NY, USA, 197–210. https://doi.org/10.1145/3098822.3098843

[51] Hongzi Mao, Malte Schwarzkopf, Shaileshh Bojja Venkatakrishnan, Zili Meng, and Mohammad Alizadeh. 2019. Learning Scheduling Algorithms for Data Processing Clusters. In *Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM '19)*. Association for Computing Machinery, New York, NY, USA, 270–288. https://doi.org/10.1145/3341302.3342080

[52] Zili Meng, Minhu Wang, Jiasong Bai, Mingwei Xu, Hongzi Mao, and Hongxin Hu. 2020. Interpreting deep learning-based networking systems. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*. 154–171.

[53] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[54] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 746–751. https://aclanthology.org/N13-1090

[55] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[56] Konstantinos Poularakis, Qiaofeng Qin, Franck Le, Sastry Kompella, and Leandros Tassiulas. 2021. Generalizable and Interpretable Deep Learning for Network Congestion Prediction. In *29th IEEE International Conference on Network Protocols, ICNP 2021, Dallas, TX, USA, November 1-5, 2021*. IEEE, 1–10. https://doi.org/10.1109/ICNP52444.2021.9651937

[57] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*. Association for Computational Linguistics, Jeju Island, Korea, 1–40. https://aclanthology.org/W12-4501

[58] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444* (2017).

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[60] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[61] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[62] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and Measuring the Geometry of BERT. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[64] Chloé Rolland, Julien Ridoux, and Bruno Baynat. 2007. LiTGen, a Lightweight Traffic Generator: Application to P2P and Mail Wireless Traffic. In *Proceedings of the 8th International Conference on Passive and Active Network Measurement (PAM'07)*. Springer-Verlag, Berlin, Heidelberg, 52–62.

[65] Krzysztof Rusek, José Suárez-Varela, Albert Mestres, Pere Barlet-Ros, and Albert Cabellos-Aparicio. 2019. Unveiling the Potential of Graph Neural Networks for Network Modeling and Optimization in SDN. In *Proceedings of the 2019 ACM Symposium on SDN Research (SOSR '19)*. Association for Computing Machinery, New York, NY, USA, 140–151. https://doi.org/10.1145/3314148.3314357

[66] Krzysztof Rusek, José Suárez-Varela, Albert Mestres, Pere Barlet-Ros, and Albert Cabellos-Aparicio. 2019. Unveiling the potential of Graph Neural Networks for network modeling and optimization in SDN. In *Proceedings of the 2019 ACM Symposium on SDN Research*. 140–151.

[67] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[68] Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5149–5152. https://doi.org/10.1109/ICASSP.2012.6289079

[69] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*. Association for Computational Linguistics, Berlin, Germany, 1715–1725. https://doi.org/10.18653/v1/P16-1162

[70] Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731* (2019).

[71] Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. 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)*. Association for Computational Linguistics, Hong Kong, China, 47–55. https://doi.org/10.18653/v1/D19-6106

[72] Arunan Sivanathan, Hassan Habibi Gharakheili, Franco Loi, Adam Radford, Chamith Wijenayake, Arun Vishwanath, and Vijay Sivaraman. 2019. Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics. *IEEE Transactions on Mobile Computing* 18, 8 (2019), 1745–1759. https://doi.org/10.1109/TMC.2018.2866249

[73] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[74] R. Smith. 2007. An Overview of the Tesseract OCR Engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02 (ICDAR '07)*. IEEE Computer Society, USA, 629–633.

[75] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://aclanthology.org/D13-1170

[76] Robin Sommer and Vern Paxson. 2010. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *2010 IEEE Symposium on Security and Privacy*. 305–316. https://doi.org/10.1109/SP.2010.25

[77] Joel Sommers, Hyungsuk Kim, and Paul Barford. 2004. Harpoon: A Flow-Level Traffic Generator for Router and Network Tests. In *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '04/Performance '04)*. Association for Computing Machinery, New York, NY, USA, 392. https://doi.org/10.1145/1005686.1005733

[78] David Stutz, Alexander Hermans, and Bastian Leibe. 2018. Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding* 166 (2018), 1–27. https://doi.org/10.1016/j.cviu.2017.03.007

[79] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8968–8975.

[80] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[81] Jacob Turton, Robert Elliott Smith, and David Vinson. 2021. Deriving Contextualised Semantic Features from BERT (and Other Transformer Model) Embeddings. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Association for Computational Linguistics, Online, 248–262. https://doi.org/10.18653/v1/2021.repl4nlp-1.26

[82] Kashi Venkatesh Vishwanath and Amin Vahdat. 2006. Realistic and Responsive Network Traffic Generation. *SIGCOMM Comput. Commun. Rev.* 36, 4 (aug 2006), 111–122. https://doi.org/10.1145/1151659.1159928

[83] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

[84] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577* (2019).

[85] Zhaohua Wang, Zhenyu Li, Guangming Liu, Yunfei Chen, Qinghua Wu, and Gang Cheng. 2021. Examination of WAN Traffic Characteristics in a Large-Scale Data Center Network. In *Proceedings of the 21st ACM Internet Measurement Conference (IMC '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3487552.3487860

[86] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7 (2019), 625–641.

[87] Michele C. Weigle, Prashanth Adurthi, Félix Hernández-Campos, Kevin Jeffay, and F. Donelson Smith. 2006. Tmix: A Tool for Generating Realistic TCP Application Workloads in Ns-2. *SIGCOMM Comput. Commun. Rev.* 36, 3 (jul 2006), 65–76. https://doi.org/10.1145/1140086.1140094

[88] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626* (2019).

[89] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

[90] Keith Winstein and Hari Balakrishnan. 2013. TCP Ex Machina: Computer-Generated Congestion Control. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM (SIGCOMM '13)*. Association for Computing Machinery, New York, NY, USA, 123–134. https://doi.org/10.1145/2486001.2486020

[91] Yikai Xiao, Qixia Zhang, Fangming Liu, Jia Wang, Miao Zhao, Zhongxing Zhang, and Jiaxing Zhang. 2019. NFVdeep: Adaptive Online Service Function Chain Deployment with Deep Reinforcement Learning. In *2019 IEEE/ACM 27th International Symposium on Quality of Service (IWQoS)*. 1–10. https://doi.org/10.1145/3326285.3329056

[92] Shengzhe Xu, Manish Marwah, Martin Arlitt, and Naren Ramakrishnan. 2021. Stan: Synthetic network traffic generation with generative neural models. In *International Workshop on Deployable Machine Learning for Security Defense*. Springer, 3–29.

[93] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. 2021. Detecting ai trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 103–120.

[94] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[95] Jane Yen, Tamás Lévai, Qinyuan Ye, Xiang Ren, Ramesh Govindan, and Barath Raghavan. 2021. Semi-Automated Protocol Disambiguation and Code Generation. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference (SIGCOMM '21)*. Association for Computing Machinery, New York, NY, USA, 272–286. https://doi.org/10.1145/3452296.3472910

[96] Hyunho Yeo, Youngmok Jung, Jaehong Kim, Jinwoo Shin, and Dongsu Han. 2018. Neural Adaptive Content-Aware Internet Video Delivery. In *Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation (OSDI'18)*. USENIX Association, USA, 645–661.

[97] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314* (2020).

[98] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326* (2018).

[99] Menghao Zhang, Jiasong Bai, Guanyu Li, Zili Meng, Hongda Li, Hongxin Hu, and Mingwei Xu. 2019. When NFV Meets ANN: Rethinking Elastic Scaling for ANN-based NFs.. In *ICNP*. IEEE, 1–6. http://dblp.uni-trier.de/db/conf/icnp/icnp2019.html#ZhangBLMLHX19

[100] Qizhen Zhang, Kelvin K. W. Ng, Charles Kazer, Shen Yan, João Sedoc, and Vincent Liu. 2021. MimicNet: Fast Performance Estimates for Data Center Networks with Machine Learning. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference (SIGCOMM '21)*. Association for Computing Machinery, New York, NY, USA, 287–304. https://doi.org/10.1145/3452296.3472926

[101] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 35–45. https://doi.org/10.18653/v1/D17-1004

[102] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129* (2019).

[103] Ying Zheng, Ziyu Liu, Xinyu You, Yuedong Xu, and Junchen Jiang. 2018. Demystifying Deep Learning in Networking. In *Proceedings of the 2nd Asia-Pacific Workshop on Networking (APNet '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3232565.3232569

[104] Hang Zhu, Varun Gupta, Satyajeet Singh Ahuja, Yuandong Tian, Ying Zhang, and Xin Jin. 2021. Network Planning with Deep Reinforcement Learning. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference (SIGCOMM '21)*. Association for Computing Machinery, New York, NY, USA, 258–271. https://doi.org/10.1145/3452296.3472902

[105] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.