5 Background and Related work in NLP: from Symbolic methods to Foundation Models

Chapter's Summary

This chapter offers an overview of Natural Language Processing (NLP) methodologies, up to the development of recent large Foundation Models, and then transitions towards Few-shot learning, a strategy for learning from limited labeled data, before culminating in a discussion of FSL applied to NLP.
The initial section of this chapter outlines the progression of NLP research in understanding human language. This includes early rule-based or feature engineering methods, the utilization of word embeddings to create distributed, meaningful representations, and the development of various architectures for effective Language Models. In Section 5.7, we investigate the prevailing approach to addressing NLP tasks, which involves large pre-trained transformer-based Language Models and their subsequent evolution towards creating versatile central models capable of handling a diverse range of tasks, despite their distinct nature. Finally, we explore in Section 5.8 the realm of Few-Shot Learning, examining its principal techniques and intersection with current NLP paradigms, while shedding light on the latest progress and challenges in this research area.

5.1 Introduction

Natural Language Processing is a crucial subdomain of computer science and AI, focused on enabling computers to comprehend, interpret, and generate human languages. NLP methods have evolved over the years to handle the messiness of textual data. The primary challenges in NLP indeed stem from the inherent complexity of natural language, which is often ambiguous, context-dependent, and unstructured (Manning and Schütze 2001). To tackle these challenges, NLP encompasses a wide range of tasks:

- low-level tasks, such as tokenization (K. Church et al. 2021), filtering (Manning, Raghavan, et al. 2008), and stemming (Porter 1980), which prepare and process raw text,
- intermediate-level tasks, like part-of-speech tagging (Marcus et al. 1993) and named entity recognition (Bunescu et al. 2005), which analyze and label the data,

- and high-level tasks, including machine translation (Sutskever, Vinyals, et al. 2014), sentiment analysis (Pang et al. 2002), and question answering (Woods 1977), that draw from this analysis to perform complex language understanding and generation.

This section traces the history of NLP advances, beginning with early rule-based methods and feature engineering techniques. We then explore the development of word representation methods, focusing on word embeddings, which have become a crucial component in modern NLP systems. The next part delves into language models, from count-based approaches such as N-grams to neural language models based on RNN, encoder-decoder architectures, and attention mechanisms. Finally, we discuss the recent emergence of transformer-based models and large Foundation models, which bridge the gap between word embeddings and language models by leveraging contextual word representations.

5.2 Early NLP methods

Rule-based and feature engineering

Rule-based methods, which originated in the early days of NLP and AI in the 1950s, relied on manually crafted rules and expert knowledge to process and analyze text. These methods were based on a set of predefined linguistic rules or patterns that were applied to the text to extract or manipulate information (William John Hutchins 1986). Some popular rule-based NLP techniques included phrase structure grammars, and context-free grammars (Chomsky 1956). Techniques such as regular expressions and finite-state automata (Mohri 1997) were also used to identify patterns and perform basic text processing tasks, such as tokenization and stemming. Rule-based methods were widely used in early machine translation systems, such as the 1954 Georgetown-IBM experiment (W. John Hutchins 2004), and natural language interfaces (Androutsopoulos et al. 1995; Woods 1977). Rule-based methods have limitations, such as scalability and adaptability to new languages or domains, that respectively require the developments of new and complex rules. The manual creation of rules is time-consuming and requires significant domain knowledge, making these methods less efficient compared to more recent data-driven approaches.

Feature engineering is a process of extracting relevant features from raw data that can be used to build effective ML models. In the context of NLP, feature engineering often involved using expert knowledge to design features based on linguistic properties and domain-specific knowledge (Jurafsky 2000). Part-of-speech (POS) tagging was used as a preprocessing step in early NLP systems, identifying the grammatical role of each word in a sentence (Marcus et al. 1993). This information could then be used as input for other NLP tasks, such as parsing or information extraction. Named entity recognition (NER) is another example of feature engineering in early NLP systems, where the goal is to identify and classify proper nouns, such as people, organizations, and locations, within a text (Bunescu et al. 2005). Dependency parsing extracts the syntactic structure of a sentence by identifying the relationships between words (i.e., subject, object, modifiers). Like POS tagging and NER, dependency parsing was used as a feature in other NLP tasks (Y. Zhang et al. 2011). Similarly to rule-based methods, feature-engineering-approaches face several major limitations, such as the need for time-consuming expert knowledge for designing effective feature

extraction methods, that may not be generalizable to new tasks and do not scale efficiently to large datasets or long sequences. While feature engineering and expert knowledge played a significant role in early NLP tasks, another approach that emerged for handling unstructured textual data was the use of vector space models.

Vector space models: Bag-of-words and TF-IDF

In these models based on linear algebra, documents and words are represented as vectors (Salton et al. 1975) with the aim of leveraging some similarity between them. Bag of Words (BoW) (Harris 1954) is a simple and widely-used method for representing text data in NLP tasks. BoW converts text into a fixed-size vector by counting the frequency of words in a document and disregarding the order of words. BoW represents each document as a vector with the same length as the vocabulary size. Each element in the vector corresponds to a word in the vocabulary and contains the frequency of that word in the document. The main limitation of BoW is that it ignores word order and contextual information, making it less effective for capturing semantic relationships between words. Additionally, BoW can lead to high-dimensional and sparse representations, which can be computationally expensive for large vocabularies.
Term Frequency-Inverse Document Frequency (TF-IDF) is a technique that extends the BoW approach by incorporating the importance of words in a document relative to their importance in the entire corpus. TF-IDF is calculated as the product of the term frequency (TF) (Luhn 1957), which is the number of times a word appears in a document, and the inverse document frequency (IDF) (Sparck Jones 1972) , which is the logarithm of the ratio of the total number of documents in the corpus to the number of documents containing the word. Hence, for a word w and a document d from a corpus C:

$$\begin{aligned}\text{TF-IDF}(w, d, C) &= \text{TF}(w, d) \times \text{IDF}(w, C) \\ &= Card(\{x \in d | x = w\}) \times \log \frac{Card(C)}{Card(\{c \in C | w \in c\})}\end{aligned}$$

where $Card(C)$ denotes the cardinality of set C. The IDF weighting scheme assigns higher weights to words that are less frequent in the entire corpus, effectively reducing the impact of common words and emphasizing the importance of more informative words for a given document. Although TF-IDF provides a more sophisticated representation of text data compared to the BoW approach, it still has limitations. Similar to BoW, TF-IDF does not capture word order or contextual information.

While vector space models such as BoW and TF-IDF have proven effective in capturing document-level information and enabling the application of ML techniques to textual data without requiring engineering or expert knowledge, they do not inherently account for the sequential and structured nature of language. To address this shortcoming, researchers have turned to probabilistic frameworks that can model the dependencies and relationships between words in a sequence.

Probabilistic frameworks

Probabilistic frameworks, such as Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs), have been widely used in early NLP tasks to model sequences and dependencies between elements in a text. HMMs are generative probabilistic models that represent the joint probability distribution of observed and hidden variables (Rabiner 1989). CRFs, on their side, are discriminative probabilistic models that directly model the conditional probability of the hidden variables given the observed variables (Lafferty et al. 2001). These probabilistic models have been used in tasks like POS tagging, NER, and shallow parsing, among others (Finkel et al. 2005; Sha et al. 2003). While they have proven to be effective in capturing relationships and dependencies in sequential data, some limitations remain, such as their lack of scalability (when dealing with long sequences or datasets, CRF are computationally expensive, whereas HMM struggle in capturing long-range dependencies due to the Markov assumption) or the lack of semantic representation (these models operate at the level of individual words), preventing them to leverage the deep semantic structure of natural language.

Methods	Scalability to large datasets	Adaptability	Expert Knowledge	Robustness to Unknown Words	Dependencies Between Words	Long Sequences Scalability	Semantic Representation
Rule-based	-	-	+	-	-	-	-
Feature-Engineering Based	-	+/-	+	+/-	-	-	-
Vector Space Models (BoW, TF-IDF)	+/-	+	-	-	-	+/-	-
Probabilistic Frameworks (HMMs, CRFs)	-	-	+	+/-	+	-	-

Table 5.1: Summary of limitations of early NLP methods. "+" denotes significant presence/requirement of the criterion, "-" denotes significant lack/limitation, and "+/-" denotes moderate presence/requirement.

Takeaways

Despite the success of early NLP methods in addressing various language processing tasks, these early techniques struggle in capturing the rich semantic and syntactic information present in natural language. The BoW and TF-IDF models, for example, lack the ability to represent the semantic relationships between words and fail to account for word order, which is crucial for understanding the meaning of a text. Similarly, while probabilistic frameworks like HMMs and CRFs offer a way to model sequences and dependencies,

they still rely on hand-crafted features and do not scale well to large vocabularies or complex dependencies. The limitations of each of these methods are synthesized in Table 5.1. As the field of NLP evolved, researchers recognized the need for better word representations that could capture both the syntactic and semantic information in text. The development of word embeddings, which are continuous vector representations of words, emerged as a promising solution to address these limitations. In the next section, we delve into the world of word embeddings, exploring the various techniques that have been proposed to learn these representations, from count-based to prediction-based methods, and how they have significantly advanced the state-of-the-art in NLP.

5.3 Word embeddings

The limitations of early NLP methods led to the development of word embeddings as a way to better represent and capture semantic and syntactic information about words. Word embeddings are continuous and dense vector representations that map words from a large vocabulary into a lower-dimensional space. These embeddings are based on the distributional hypothesis, which states that words that occur in similar contexts tend to have similar meanings (Firth 1957; Harris 1954)[1]. They can be generated using various techniques, broadly categorized into count-based and prediction-based methods.

Count-based word embeddings

Count-based word embedding techniques take the idea to put information about contexts into word vectors literally, by manually designing a word-context matrix M in which columns represent potential contexts and rows represent words. In a second step, a dimension reduction technique is applied to the matrix to produce dense embeddings. As their name suggests, these approaches are based on global corpus statistics, and in that sense share some similarities with BoW and TF-IDF. However, those latter methods are not considered as count-based word embeddings because they represent documents rather than individual words and produce sparse vectors instead of dense embeddings.
From there, the different count-based word embeddings strategies differ in the way to consider what is context (hence defining what represent the matrix columns) and how to compute matrix elements. A simple co-occurrence-based approach is for instance to consider as contexts the surrounding words contained in a fixed-size sliding window, and to define M as a word-word matrix with M_{ij} being the number of times word w_i appears in context w_j (Lund et al. 1996). Based on the same definition of contexts, information theoretic measures such as Pointwise Mutual Information (PMI) (K. W. Church et al. 1990) and Positive Pointwise Mutual Information (PPMI) (Bullinaria et al. 2007) have been used to define word representations in matrix M. PMI of a words pair (w_i, w_j) is defined as the log ratio between joint probabilities and product of marginal probabilities: $PMI(w_i, w_j) = \log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}$. Intuitively, designing the matrix M such that

[1] Also found as "You shall know a word by the company it keeps"

$M_{ij} = PMI(w_i, w_j)$ will associate positive values to word pairs (w_i, w_j) that appear more frequently in a same context than if they were independent, and negative values to word pairs that appear less frequently than being independent. (Bullinaria et al. 2007) extend this idea by considering only positive values, that is defining $M_{ij} = PPMI(w_i, w_j) = \max(PMI(w_i, w_j), 0)$. Finally, a popular count-based word embedding technique is Latent Semantic Analysis (LSA) (Deerwester et al. 1990). Alternatively, LSA considers different documents from a corpus C as contexts, and hence designs matrix M as a word-document matrix, with $M_{ij} = \text{TF-IDF}(w_i, d_j, C)$. The second step is then to reduce the dimensionality of the term-document matrix through a singular value decomposition (SVD) to capture latent semantic relationships between words and documents. By doing so, LSA can identify and represent synonyms, polysemes, and other linguistic relationships in the reduced-dimensional space.
While count-based word embeddings capture dependencies between words and semantic relationships through their term-context matrix, constructing and factorizing such large matrices may undermine their scalability. Besides, count-based models generally struggle with out-of-vocabulary words since they are based on direct observation of the training corpus.

Prediction-based word embeddings

Prediction-based word embeddings are generated by training models to predict words or their contexts based on the local context information, which is generally a sliding window surrounding the target word. This approach aims to learn word representations that can effectively capture semantic and syntactic information while exploiting the co-occurrence patterns of words in their local contexts. Two popular prediction-based word embedding techniques are Word2Vec and FastText.
Word2Vec, developed by Mikolov et al., is a highly influential prediction-based word embedding method. Word2Vec consists of two main model architectures: Continuous Bag of Words (CBoW) (Tomás Mikolov, K. Chen, et al. 2013) and Skip-Gram (Tomas Mikolov et al. 2013). CBoW aims to predict the target word based on the surrounding context words, while Skip-Gram focuses on predicting context words given a target word (see Figure 5.1). For both architectures, word vectors are model parameters that are updated along the training through Maximum Likelihood Estimation (MLE) when moving the sliding window along the training corpus and predicting either target word or context words at each position. Word2Vec embeddings have been shown to produce state-of-the-art results on various NLP tasks when released.

FastText (Bojanowski et al. 2017) is an extension of the Word2Vec algorithm that focuses on learning representations for subword units. By representing words at the character scale, FastText can efficiently learn embeddings for rare and out-of-vocabulary words. FastText has been shown to improve performance on a range of NLP tasks, such as text classification (Dharma et al. 2022).
Finally, **GloVe**, a popular hybrid approach between count-based and prediction-based techniques has been developed in 2014 (Pennington et al. 2014). GloVe combines the benefits of matrix factorization techniques, like LSA, and local context window-based methods, such as Word2Vec. It constructs a word co-occurrence matrix from a large corpus and uses a weighted least squares objective function to learn word vectors that can effectively capture semantic and syntactic information. It hence captures both global and local context information, allowing for a more com-

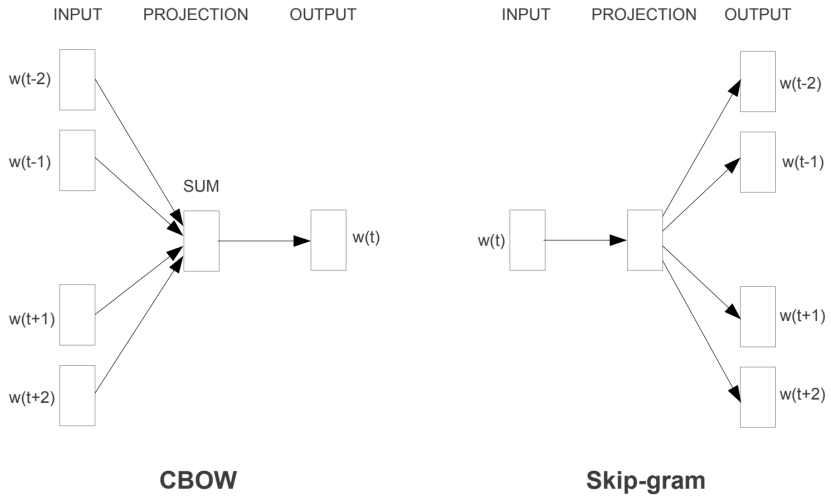

Figure 5.1: Comparison of CBoW and Skip-Gram approaches. CBoW projects context words to predict a central word (left), while Skip-Gram inversely projects a unique word to predicts its context (right). Figure from (Tomás Mikolov, Le, et al. 2013).

prehensive representation of word meaning. However, it requires explicit construction of the co-occurrence matrix, which can be computationally expensive for larger corpora, and it can be sensitive to the choice of hyperparameters, such as the window size and weighting scheme.

Interestingly, using similarity to build rich word representations is not reflected only in quantitative metrics of subsidiary tasks. (Tomas Mikolov et al. 2013) indeed qualitatively analyzed the learned vector space and pointed out geometrical patterns based on meanings similarity (see Figure 5.2). Thus, the difference between the representation vectors of many country/capital pairs seem to produce the same vector. Another example (Tomás Mikolov, Le, et al. 2013) shows the similar distribution of embedding vectors from a language to another one, suggesting a simple linear mapping for translation.

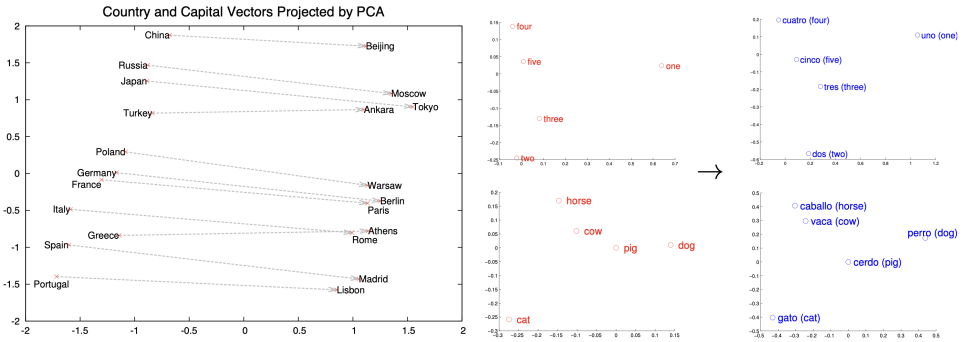

Figure 5.2: Qualitative results for Word2Vec embeddings. Subtracting capital vector to its related country vector produces similar vector among all country/capital pairs (left). Learned embeddings of number and animal words have very similar spatial distribution in English and Spanish (right).

Methods	Scalability to large datasets	Adaptability	Expert Knowledge	Robustness to Unknown Words	Dependencies Between Words	Long Sequences Scalability	Semantic Representation	Context-dependent representations
Rule-based	-	-	+	-	-	-	-	-
Feature-Engineering Based	-	+/-	+	+/-	-	-	-	-
Vector Space Models (BoW, TF-IDF)	+/-	+	-	-	-	+/-	-	-
Probabilistic Frameworks (HMMs, CRFs)	-	-	+	+/-	+	-	-	-
Count-Based Word Embeddings	+/-	+	-	-	+	+/-	+	-
Prediction-Based Word Embeddings	+	+	-	+	+	+/-	+	-

Table 5.2: Summary of limitations of early NLP methods and word embeddings techniques. "+" denotes significant presence/requirement of the criterion, "-" denotes significant lack/limitation, and "+/-" denotes moderate presence/requirement.

Takeaways

Word embeddings have become an essential tool in NLP, capturing semantic and syntactic relationships between words and providing a foundation for more advanced techniques. However, despite their ability to capture word relationships, word embeddings have limitations, particularly in representing context-dependent word meanings. Indeed, these representations are pre-computed in a static corpus, which may not be convenient when using a word in a different context afterwards (this is notably the case for polysemous words that have in this framework only one representation). Besides, long sequences can be handled well as the window size can be varied, but distant dependencies might be missed. The comparison of approaches is thus updated in Table 5.2.
We now delve into language models, which offer a comprehensive approach to capture the structure and context of language. Their development have led to powerful and versatile models capable of handling complex linguistic phenomena and significantly improving performance on a wide range of tasks, such as machine translation, speech recognition, and text generation.

5.4 Language models

Language models play a critical role in various NLP tasks by predicting the likelihood of a sequence of words, represented as a probability distribution over words. Given a sequence of words

$(w_1, w_2, ..., w_n)$, a language model assigns a probability $\mathbb{P}(w_1, w_2, ..., w_n)$ to this sequence. This can be used for numerous applications such as machine translation (Bahdanau et al. 2015; Koehn et al. 2003; Sutskever, Vinyals, et al. 2014), speech recognition (G. Hinton et al. 2012; Jelinek 1991), and text generation (Graves 2013). In this section, we explore the evolution of language modeling techniques, from early count-based approaches to more sophisticated neural models that have driven significant advances in the field of NLP.

Count-based language models

The early days of language modeling were dominated by count-based methods, with N-gram models being one of the most widely-used approaches (Jelinek 1991). N-grams are simply contiguous sequences of N words, where N is a fixed integer. An N-gram language model predicts the probability of a word given its preceding $N-1$ words by estimating the frequency of N-grams in a large corpus. Thus, an N-gram model makes a Markov assumption, which states that the probability of a word depends only on the previous $N-1$ words:

$$\mathbb{P}(w_n|w_1, \ldots, w_{n-1}) \approx \mathbb{P}(w_n|w_{n-N+1}, \ldots, w_{n-1})$$

N-gram probabilities $\mathbb{P}(w_n|w_{n-N+1}, \ldots, w_{n-1})$ can be estimated by counting in a corpus the occurrences of N-gram $(w_{n-N+1}, \ldots, w_{n-1}, w_n)$ and normalizing by the number of occurrences of $(w_{n-N+1}, \ldots, w_{n-1})$.
Despite their simplicity, N-gram models suffer from several limitations, such as data sparsity, which occurs when certain N-grams do not appear in the training corpus, leading to inaccurate probability estimates. To overcome this issue, various smoothing techniques have been proposed (S. F. Chen et al. 1996). Other drawbacks of N-gram models are their inability to capture long-range dependencies, as they only consider a fixed number of preceding words to predict the next word, or the curse of dimensionality they may face when considering large vocabulary (Bengio, Ducharme, et al. 2000).

While count-based language models have provided a foundation for early NLP research, their limitations have led to the development of more advanced techniques such as neural language models (Bengio, Ducharme, et al. 2000), that afterwards leveraged the power of deep learning to better understand and represent natural language.

Neural language models

Neural language models aim to provide a continuous representation of words and capture semantic and syntactic information in dense vector space. They have demonstrated their ability to overcome some of the limitations of count-based language models, such as the curse of dimensionality and the sparsity of N-grams. One of the first neural language models was a feedforward neural network (FFN) language model (Bengio, Ducharme, et al. 2000). This model aimed to predict the next word in a sequence by concatenating word embeddings of previous words and feeding them into the FFN. The output models the word probability given a context. The model's architecture is illustrated in Figure 5.3.

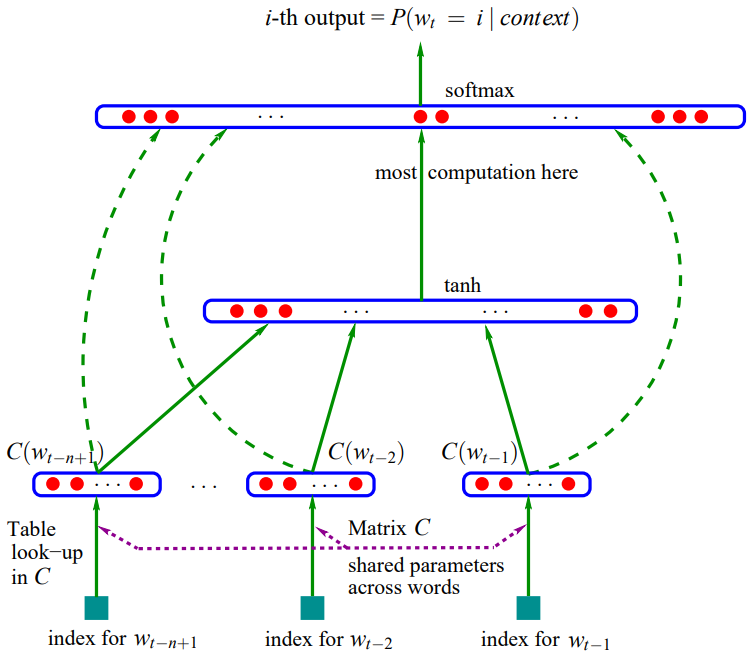

Figure 5.3: Neural Language Model architecture. The input sentence $(w_{t-n+1}, \ldots, w_{t-1})$ is converted to feature vectors stored in a matrix C, which are then fed to a neural network g represented by the green plain lines. The output of g estimates the probability of each word in the vocabulary, conditioned the input context. Figure from (Bengio, Ducharme, et al. 2000).

Recurrent Neural Networks were introduced as an extension to feedforward neural language models to better capture long-range dependencies in natural language data (Elman 1990). RNNs are designed to process sequences of variable length by maintaining a hidden state that can store information from previous time steps (Tomás Mikolov, Karafiát, et al. 2010). However, RNNs have some limitations, such as the vanishing gradient problem that makes learning long-range dependencies difficult (Hochreiter, Bengio, et al. 2001). To overcome the vanishing gradient problem in RNNs, **Long Short-Term Memory** (LSTM) networks were proposed (Hochreiter and Schmidhuber 1997). LSTMs introduce a gating mechanism that helps to maintain and propagate information over long sequences, making them more effective for learning long-range dependencies. LSTMs have thus been used as building blocks for Language Models (Sundermeyer et al. 2012). Finally, **Gated Recurrent Units** (GRU) are another variant of RNNs that simplify the LSTM architecture while retaining its ability to model long-range dependencies (Cho et al. 2014). GRUs use update and reset gates to control the flow of information in the hidden state, making them computationally more efficient than LSTMs, however they may not capture long-term dependencies as well as LSTM.

5.5 Encoder-decoder architecture

Many NLP tasks require not only an understanding of the input text but also the generation of a meaningful output sequence, such as in neural machine translation and text summarization. To tackle these challenges, a new class of models has emerged: encoder-decoder architectures, also known as **sequence-to-sequence models** (Sutskever, Vinyals, et al. 2014). The encoder-decoder architecture is composed of two main components: the encoder and the decoder. The encoder processes the input sequence and generates a fixed-length context vector that encapsulates the essential information of the input. The decoder, in turn, takes this context vector and generates an output sequence, conditioned on the input sequence. These architectures split the model into two parts, with one component (the encoder) focusing on processing the input sequence and the other (the decoder) generating the output sequence (Cho et al. 2014). In the early encoder-decoder models, both the encoder and decoder were typically implemented as RNNs, LSTMs, or GRUs. The encoder processes the input sequence one token at a time, updating its hidden state at each step. The final hidden state of the encoder is then used as the initial hidden state of the decoder, which generates the output sequence one token at a time. An illustration of this family of architectures is given in Figure 5.4 .

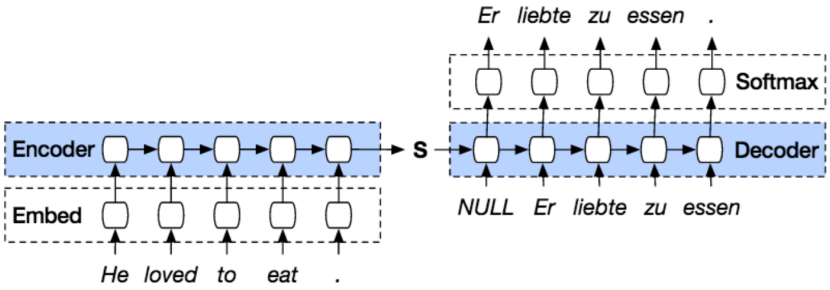

Figure 5.4: Sequence-to-Sequence architecture[2]. Every words of the input sentence are embedded and then sequentially fed to the encoder module, that stores the input information in a context S. Using this context and the previous generated token (starting with a special token), the decoder module sequentially generates the output.

While the encoder-decoder architecture was a significant improvement over the previous models, it still faced some limitations. One of the main challenges was that the encoder had to compress the entire input sequence into a single fixed-size context vector, which could result in loss of information, especially for long input sequences (Bahdanau et al. 2015). This limitation prompted researchers to explore more sophisticated ways to better capture and leverage the information in the input sequence, leading to the development of attention mechanism.

Attention mechanism

The key idea behind attention mechanism (Bahdanau et al. 2015) is that the decoder should be able to focus on different parts of the input sequence at different time steps, rather than relying

[2] Figure from https://www.guru99.com/seq2seq-model.html

solely on a single context vector. This allows the model to weight the importance of different input tokens and selectively retrieve information from the input sequence. In an attention-based encoder-decoder model, the encoder produces a sequence of hidden states, one for each input token. The decoder, at each time step, computes a weighted sum of these hidden states, where the weights are determined by the attention mechanism. These weights, also known as attention scores, indicate how much the decoder should "attend" to each input token when generating the output token at a given time step. The attention mechanism computes attention scores using a scoring function that takes as input the current hidden state of the decoder and the hidden states of the encoder. There are several variants of the scoring function, such as dot product, additive, and multiplicative attention (T. Luong et al. 2015). The introduction of attention mechanisms significantly improved the performance of encoder-decoder models on a wide range of NLP tasks, including neural machine translation (Bahdanau et al. 2015), text summarization (Rush et al. 2015), and speech recognition (Chorowski et al. 2015). The success of attention mechanisms in these tasks paved the way for further advancements in NLP, such as the development of transformers.

5.6 Transformers

Despite the success of attention mechanisms in improving the performance of encoder-decoder models, researchers continued to explore ways to further enhance the capabilities of NLP models. One significant drawback of the RNN-based models was their sequential nature, which makes it difficult to parallelize the computations and exploit the full potential of modern hardware, such as GPUs. In response to this challenge, (Vaswani et al. 2017) introduced the Transformer architecture, which replaces the recurrent layers in encoder-decoder models with self-attention mechanisms. This groundbreaking innovation has become the foundation for many state-of-the-art models in NLP, including BERT (Devlin et al. 2019), GPT(Radford, Narasimhan, et al. 2018b), and their variants, as well as in other domains (vision (Dosovitskiy et al. 2021), speech (Radford, Kim, T. Xu, et al. 2022), etc.).
The self-attention mechanism is at the core of the Transformer architecture. Unlike the attention mechanism used in encoder-decoder models, self-attention operates within a single sequence, allowing each token to attend to all other tokens in the sequence. This mechanism enables the model to capture long-range dependencies more effectively and allows for parallel computation across tokens. See Subsubsection 2.2.3 for a more detailed overview of the self-attention mechanism. The Transformer architecture is built upon a stack of self-attention layers and feed-forward layers, with residual connections and layer normalization applied throughout the model. The original Transformer model proposed in (Vaswani et al. 2017) consists of an encoder and a decoder, similar to the earlier encoder-decoder models. The encoder is composed of a stack of identical layers, each containing a multi-head self-attention mechanism followed by a position-wise feed-forward network. The decoder has a similar structure, with an additional layer of cross-attention that attends to the encoder's output. The global architecture is presented in Figure 5.5.

Transformers can also be designed as standalone encoders or decoders for various NLP tasks, depending on the nature of the problem and the desired model architecture. For instance, BERT (Devlin et al. 2019) is built upon a stack of Transformer encoder layers, while GPT (Radford, Narasimhan, et al. 2018b) uses a stack of Transformer decoder layers. Using only the encoder part of the Trans-

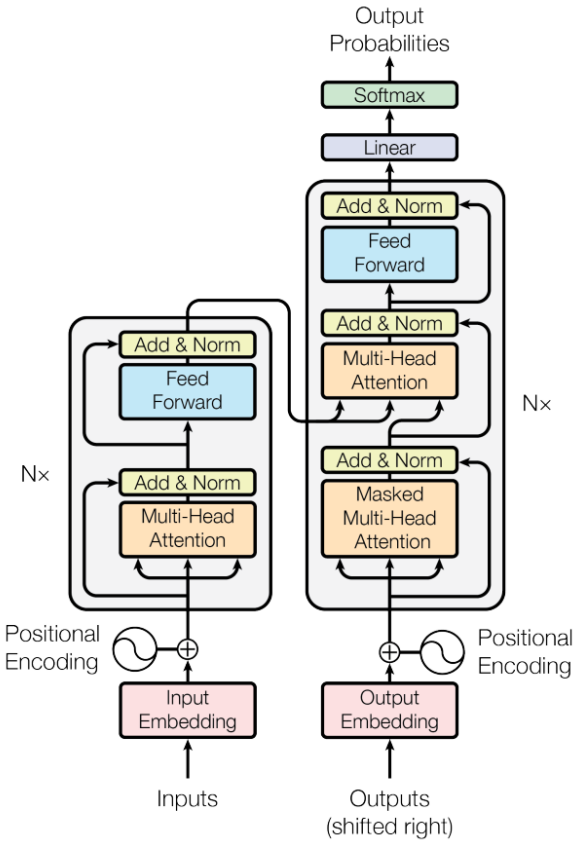

Figure 5.5: Original Transformer architecture. Similarly to encoder-decoder models, the embedded input is first encoded in a specific module before the decoder module generates the output autoregressively. The main difference is the use of Self-attention modules that make possible to model contextual dependencies between all parts of the sequences. The masking process in the decoder modules enables to parallelize the training. Figure from (Vaswani et al. 2017).

former architecture can be more suitable for tasks that require a fixed-length representation of the input sequence, such as sentence classification. The Transformer encoder processes the input sequence and produces a contextualized representation for each token, which can be aggregated or pooled to generate a fixed-length vector. On the other hand, using only the decoder part of the Transformer can be advantageous for tasks that involve generating text or predicting the next token in a sequence, such as language modeling, text generation, and summarization. The Transformer decoder is designed to handle autoregressive decoding, where the model generates one token at a time and feeds the generated tokens back as input for the subsequent steps. This architecture enables the model to leverage the self-attention mechanism for capturing dependencies between generated tokens, while still benefiting from the parallelizability and efficient handling of long-range dependencies offered by the Transformer architecture.

Methods	Scalability to large datasets	Adaptability	Expert Knowledge	Robustness to Unknown Words	Dependencies Between Words	Long Sequences Scalability	Semantic Representation	Context-dependent representations
Rule-based	-	-	+	-	-	-	-	-
Feature-Engineering Based	-	+/-	+	+/-	-	-	-	-
Vector Space Models (BoW, TF-IDF)	+/-	+	-	-	-	+/-	-	-
Probabilistic Frameworks (HMMs, CRFs)	-	-	+	+/-	+	-	-	-
Count-Based Word Embeddings	+/-	+	-	-	+	+/-	+	-
Prediction-Based Word Embeddings	+	+	-	+	+	+/-	+	-
Count-based Language Models	+/-	+/-	-	-	+/-	-	-	-
Recurrents Neural Networks (LSTM,GRU)	+	+	-	+	+	+/-	+	+
Transformers	+	+	+	+	+	+	++	++

Table 5.3: Summary of advantages and limitations of general NLP methods and word embeddings techniques. "+" denotes significant presence/requirement of the criterion, "-" denotes significant lack/limitation, and "+/-" denotes moderate presence/requirement.

Takeaways

Driven by the diverse requirements of NLP tasks and the inherent pursuit of comprehending and generating human language automatically, numerous frameworks and methodologies have been pursued and refined, successively diminishing the constraints of preceding methods (see Table 5.3). The advent of word embedding methods marked a significant milestone, providing dense, vector-based semantic representations that proved invaluable for a multitude of downstream tasks.

Recurrent Neural Networks, particularly LSTM, advanced this paradigm by capturing distributed, contextually-dependent representations via their hidden state. They led to the introduction of a new architectural framework: the Encoder-Decoder model. This approach is exceptionally suitable for tasks requiring contextual generation, such as machine translation.

The colossal breakthrough came with the advent of Transformer models, inspired by the Encoder-Decoder architecture and the introduction of the Attention Module. These models offer outstanding semantic and context-aware representations through their self-attention module, directly capturing all types of dependencies across sequence elements, rather than compressing pertinent information within a hidden state as is the case with

LSTM. Furthermore, the ability of Transformer models to parallelize efficiently permits impressive scaling, aligning seamlessly with the capabilities of modern hardware. This has resulted in Transformers becoming the cornerstone for the vast majority of today's architectural designs in NLP and other applications of Deep Learning.

5.7 Foundation models

Transformers have significantly impacted the field of NLP, and their introduction came with a change of paradigm in the field. Rather than using an end-to-end supervised framework composed of task-specific neural networks, most works in the recent years follow the pre-training and fine-tuning paradigm to achieve state-of-the-art performance across a wide range of NLP tasks. This has today led to the Foundation models era, that aim to unify all kind of NLP tasks within a single architecture.

Remark. Following the Center for Research on Foundation Models of Standford University[3], we refer to Foundation models (Bommasani et al. 2021) as the following: "In recent years, a new successful paradigm for building AI systems has emerged: Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model.". These models are based on Pre-trained Language Models (PLMs) architectures (see thereafter), and as they become larger and larger, are often referred to as Large Language Models. The interchange of these terms is hence frequent in the literature.

Pre-training and fine-tuning paradigm

The pre-training and fine-tuning paradigm has emerged as a successful approach for building Pre-trained Language Models in NLP. The idea is to first train a large neural network (mainly transformer-based one) on a massive amount of unsupervised text data (such as the C4 dataset (Raffel, Shazeer, et al. 2020)), and then fine-tune the pre-trained model on a specific supervised task (Howard et al. 2018; Peters et al. 2018). This approach leverages the ability of DL models to learn rich and meaningful representations from large-scale data, which can then be adapted to specific tasks with relatively small amounts of labeled data (see Figure 5.6).

Transfer learning is a key concept underlying the pre-training and fine-tuning paradigm. It refers to the process of transferring knowledge learned in one task or domain to another, usually related, task or domain (S. J. Pan et al. 2010). In NLP, transfer learning has been shown to be highly effective, as the knowledge learned from large-scale unsupervised text data can be generalized to a wide range of tasks (Ruder et al. 2019). The benefits of transfer learning in NLP are numerous. Firstly, it allows for more efficient learning and better generalization, as the pre-trained model has already learned meaningful language representations (Bengio, Courville, et al. 2013). Secondly, it reduces the need for labeled data in the target task, as the pre-trained model can be fine-tuned with relatively small amounts of labeled data (Peters et al. 2018). Finally, it leads to faster convergence

[3] https://crfm.stanford.edu/

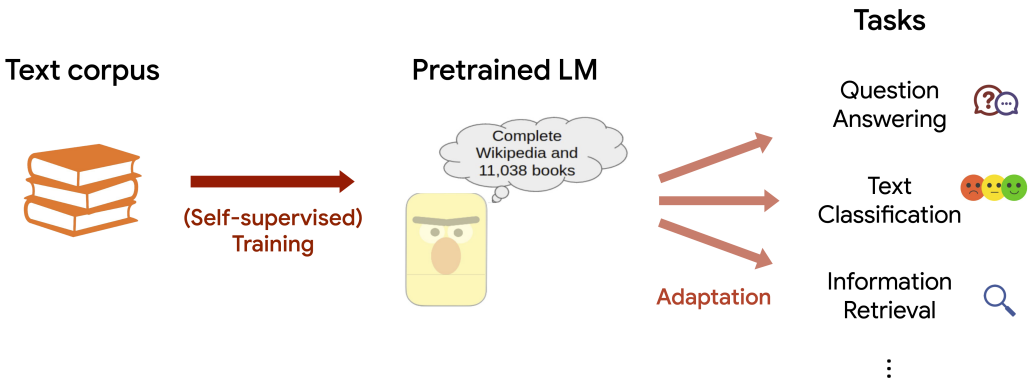

Figure 5.6: Pre-training and fine-tuning paradigm.[4] Large Language Models are first trained in an unsupervised fashion on massive textual corpora, and then fine-tuned on a specific supervised dataset for a related task.

and improved performance, as the model can leverage the knowledge learned during pre-training (Howard et al. 2018; Ruder et al. 2019).

Pioneering works: pre-trained language models to produce contextual word representations

As we discussed, Foundation models aim to acquire a vast amount of knowledge by pre-training on massive unsupervised corpora. The choice of pre-training tasks and associated losses is therefore crucial in enabling these models to gain the general linguistic knowledge necessary for effective downstream task performance. By carefully designing the pre-training objective, we can encourage the model to learn valuable patterns, structures, and relationships within the data that can be effectively transferred to a wide range of downstream tasks. In this context, pre-training losses play a pivotal role in guiding the learning process of foundation models and shaping their ability to generalize and adapt to various NLP challenges.

In the initial stages, **ELMo** (Peters et al. 2018) was developed to obtain context-sensitive word representations by first pre-training a bidirectional LSTM (biLSTM) network (rather than acquiring fixed word representations). Subsequently, the biLSTM network was fine-tuned to cater to particular downstream tasks.

BERT (Devlin et al. 2019) is a powerful model based on the Transformer encoder architecture. BERT is pre-trained on a large corpus of text using a **Masked Language Modeling** (MLM) objective, which enables it to learn bidirectional contextual representations. In this objective, a certain percentage of the input tokens are randomly masked (literally replaces by a MASK token), and the model is trained to predict the original token based on the context provided by the surrounding unmasked tokens. The MLM loss is calculated by comparing the predicted probabilities for the masked tokens with the true tokens using cross-entropy. This objective allows BERT to learn

[4] Figure from https://ai.stanford.edu/blog/linkbert/

deep bidirectional representations, capturing both the left and the right context of each token. BERT is also pre-trained using a Next Sentence Prediction (NSP) loss, in which the model shall predict if a sequence is subsequent to another one (but the NSP loss appeared to have low impact on performance). (Yamaguchi et al. 2021) explored other cheaper pre-training objectives, similar to MLM, and showed comparable performance (see Figure 5.7). Context-aware word representations of BERT and its variants (such as RoBERTa (Yinhan Liu et al. 2019)) have demonstrated state-of-the-art performance on a wide range of NLP predictive tasks, such as sentiment analysis, named entity recognition, and question-answering. Fine-tuning BERT on task-specific datasets allows it to adapt its powerful pre-trained representations to the target task, often with minimal additional training.

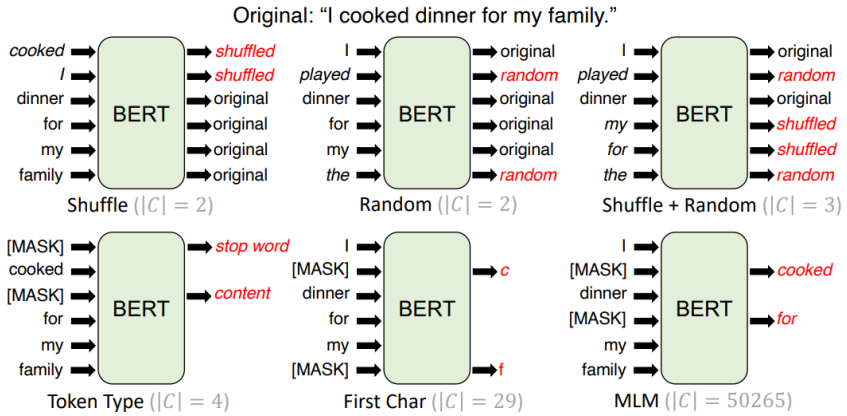

Figure 5.7: Masked Language Modeling and similar pre-training objectives. In each scenario, $|C|$ represents the number of classes of the pre-training objective, which considerably impacts computational efficiency. Figure from (Yamaguchi et al. 2021).

GPT (Radford, Narasimhan, et al. 2018b) is another significant milestone in contextual word representations. GPT models are based on the Transformer decoder architecture and are pre-trained using a unidirectional autoregressive **Language Modeling** (LM) objective. The primary goal of GPT is to predict the next token in a sequence given its preceding context. The LM loss is computed by comparing the predicted probabilities for the next token in the sequence with the true next token using cross-entropy. The unidirectional nature of GPT allows it to learn powerful contextual representations, capturing the left context of each token. However, due to their autoregressive loss, these models are especially suitable for generative tasks such as dialogues and document summarization. There have been several iterations of the GPT model, with GPT-2 (Radford, J. Wu, et al. 2019) and GPT-3 (Brown et al. 2020), especially differing by their sizes, both in number of parameters and training corpora. More recently, GPT-4 (OpenAI 2023) was released, once again crushing its previous version size with now 1 trillion (10^{12}) parameters, and now being multimodal, as it can process both text prompts and images as input. Like BERT, GPT models can be fine-tuned on task-specific datasets to adapt their pre-trained representations to the target tasks.

Conditional Language Modeling (CLM) objective is another type of pre-training loss used in some foundation models. Unlike standard LM loss used in GPT, which focuses on predicting the next word in a sequence given the previous words, or the MLM loss used in BERT that concentrates on predicting randomly masked words within a sentence, the CLM loss aims at reconstructing the input sequence after a specific kind of perturbations. A prominent encoder-decoder architecture that employs CLM objective is T5 (Raffel, Shazeer, et al. 2020), that adopts a text-to-text transfer learning approach, where both input and output sequences are represented as text strings. It is pre-trained on a denoising autoencoder task, which involves reconstructing the original text from a corrupted version. During pre-training, T5 introduces noise to the input text by applying transformations such as token masking or deletion. The model then learns to recover the original input sequence from the perturbed version. By learning to reconstruct the original sequence, T5 captures bidirectional context and adapts well to various NLP tasks. Another notable architecture that uses CLM loss is BART (Lewis et al. 2020). BART also adopts a denoising autoencoder setup, applying transformations such as token masking, token deletion, or text shuffling. The combination of bidirectional context and autoregressive nature allows both T5 and BART to excel in a wide range of tasks, taking advantage of both LM and MLM frameworks.

The different pre-training objectives are listed in Table 5.4. For each objective, the considered network aims to model the conditional probability p. It can be trained with maximum likelihood estimation.

Objective	**Loss**
MLM	$\mathcal{L}_{MLM} = -\sum_{\tilde{w} \in m(\mathbf{w})} \log p(\tilde{w} \mid \mathbf{w}_{\setminus m(\mathbf{w})})$
LM	$\mathcal{L}_{LM} = -\sum_{t=1}^{T} \log p(w_t \mid \mathbf{w}_{<t})$
CLM	$\mathcal{L}_{CLM} = -\sum_{t=1}^{T} \log p(w_t \mid \tilde{\mathbf{w}}, \mathbf{w}_{<t})$

Table 5.4: Pre-training objectives and their respective loss functions for a sentence $\mathbf{w} = (w_1, \ldots, w_T)$. $\mathbf{w}_{<t} := (w_1, \ldots, w_{t-1})$, while $m(\mathbf{w})$ designs masked words of $\mathbf{w}$, $\mathbf{w}_{\setminus m(\mathbf{w})}$ designs the unmasked elements of $\mathbf{w}$ and $\tilde{\mathbf{w}}$ designed corrupted sentence.

In summary, the introduction of PLM have revolutionized the field of NLP, providing general-purpose contextual word representations that have significantly improved performance across various tasks. Building on this success, following works developed larger architectures to still improve performances on downstream tasks.

Large language models

Several studies (Hoffmann et al. 2022; Kaplan et al. 2020; Rosenfeld et al. 2020) have demonstrated the advantages of scaling up language models in terms of model size, dataset size, and computational resources, by introducing scaling laws in terms of loss reduction. This led to the emergence of **Large Language Models** (LLMs). LLMs, typically composed of Transformer-based architec-

Model	Architecture	Pre-training Loss	Corpus
ELMo	LSTM	biLM	WikiText-103
GPT	Transformer Decoder	LM	BookCorpus
BERT	Transformer Encoder	MLM & NSP	WikiEn+BookCorpus
RoBERTa	Transformer Encoder	MLM	BCOS
BART	Transformer	CLM	BCOS
T5	Transformer	CLM	C4

Table 5.5: Overview of different Transformer-based models. BCOS stands for BookCorpus+CCNews+OpenWebText+STORIES. biLM is a bidirectional LM loss.

tures with hundreds of billions or more parameters, are trained on extensive text datasets. These scaled-up models, despite adopting similar Transformer architectures and pre-training objectives as smaller PLMs, benefit significantly from increased model size, data size, and computational power. Over the last years, several tech resource-rich organizations launched their own LLM, with for instance Google's PaLM (Chowdhery et al. 2022) and LaMDA (Thoppilan et al. 2022), OpenAI's GPT-4 (OpenAI 2023), DeepMind's Chinchilla (Hoffmann et al. 2022), or Meta's LLaMA (Touvron et al. 2023). In parallel, a team of researchers released BLOOM (Scao et al. 2022), a 176B-parameter open-access language with the aim to make this kind of models publicly accessible. Figure 5.8 provides an overview of the main LLM released over the last years.

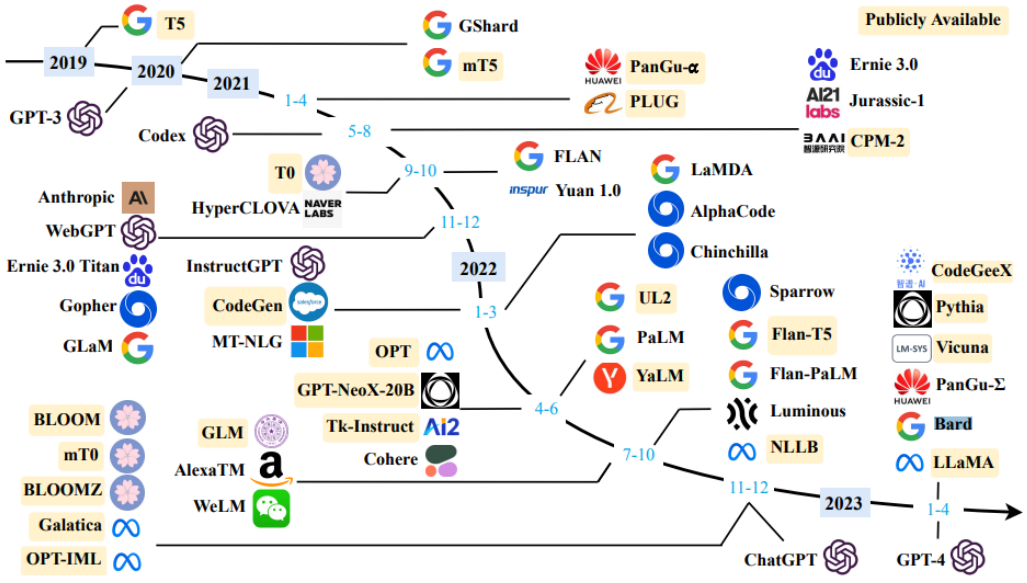

Figure 5.8: Timeline (left-to-right) of the released LLMs (bigger than 10B parameters) over the last years. The models marked in yellow are the ones made available for public use. The figures along the timeline represent the month of release. Figure from (W. X. Zhao et al. 2023).

Interesting learning abilities

LLMs exhibit strong capacities to understand natural language, generate text, and display emergent abilities, that "are not present in small models but arise in large models" (Wei, Tay, et al. 2022). These abilities include In-context learning (ICL) and Instruction formatting.

Introduced by GPT-3 (Brown et al. 2020), ICL allows language models to generate outputs at test time, given demonstrations of a task, without requiring additional fine-tuning or gradient updates. While the 175B GPT-3 model exhibits strong ICL abilities, the GPT-1 and GPT-2 models do not.

Besides, when fine-tuned on multi-task datasets using instructions (natural language descriptions), LLMs show considerable performance on unseen tasks that are also described by instructions (Ouyang et al. 2022; Sanh et al. 2022), without necessarily giving the model explicit examples, improving generalization abilities. Some studies (H. W. Chung et al. 2022; Wei, Bosma, et al. 2022) showed that this phenomenon induced by instruction-formatting essentially appears once a sufficient size has been reached. Some models such as Galactica (R. Taylor et al. 2022) even include Instruction formatting within the pre-training stage to achieve superior performance and better generalization capacity.

These emergent abilities are illustrated in Figure 5.9.

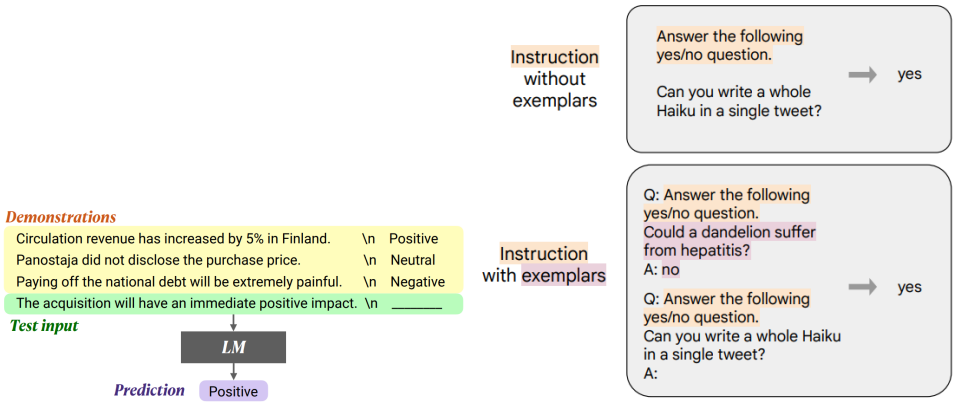

Figure 5.9: In-context learning (left): The model is given a prompt containing k input-label pairs (here $k = 3$) alongside with a test input (in the same prompt), and is asked to predict in response the test label. The model leverages the information contained in the demonstrations to effectively generate the label with no gradient update. Figure from (S. Min, X. Lyu, et al. 2022).
Instruction fine-tuning (right): The model is fine-tuned by providing Natural language descriptions of the task in preamble. It can also contain labeled examples in the prompt (bottom). Figure from (H. W. Chung et al. 2022).

Some limitations

Whereas LLMs have demonstrated impressive performance across a broad spectrum of NLP tasks, they sometimes produce unexpected outputs, or hallucinations, that may cause harm or mislead the user. To prevent this behavior, the concept of human alignment has been introduced to ensure LLMs outputs align with human expectations (Glaese et al. 2022; Ouyang et al. 2022). Reinforcement learning from human feedback (RLHF) (Christiano et al. 2017; Ziegler et al. 2019) for instance uses a policy-gradient RL algorithm to adjust LLMs based on human feedback. The integration

of human preferences via instructions, combined with training on both code and natural text segments, resulted in the development of the GPT-3.5 series. After undergoing a conversation-like training process, the widely-adopted chatbot ChatGPT was introduced, significantly influencing future AI research and underscoring the potential of human-like AI systems. Google similarly then released their chatbot BARD, aligned on human preferences with their own instruction fine-tuning method FLAN (Wei, Bosma, et al. 2022). Anthropic's Claude chatbot has on its side been aligned with human moral behavior using a technique called Constitutional AI (Bai et al. 2022), providing a principle-based approach to produce harmless outputs.

Takeaways

Large Language Models have made remarkable strides in the field of NLP by employing the **pre-training and fine-tuning paradigm**. This approach has enabled these models to achieve impressive results on a wide range of NLP tasks, even though the tasks themselves are quite diverse. While these models are yet subject to **hallucinations**, human **alignment** appeared as first step to ensure more control on their output. However, the fine-tuning process needs sizable labeled datasets for adapting the model to a new task, given the significant number of parameters involved. The challenge of gathering annotated data is amplified by the expenses involved and the scarcity of such data across different languages and domains. Consequently, there is a pressing need to develop effective methods for learning with limited annotated data. In parallel, LLMs show **emergent abilities**, such as **In-Context Learning**, that may be suitable for addressing this challenge. This leads us to the next section, which focuses on Few-Shot Learning (FSL) techniques for NLP.

5.8 Few-shot learning in NLP

Few-shot learning paradigm

Few-Shot Learning (FSL) refers to the ability to learn tasks with limited annotated examples. This ability of humans, that are able to use their previous experience to adapt fastly to new context, has been largely studied recently in the context of machine learning algorithms (Lake et al. 2015). As illustrated in Figure 5.10, it can concerns many tasks : classification , generation, etc.

Historically, **Meta-learning** -or learning to learn (Thrun et al. 1998)- approaches have for quite long stood as the *de-facto* paradigm for FSL (K. Lee et al. 2019; A. Raghu et al. 2020; A. Rusu et al. 2019; Snell et al. 2017; Q. Sun et al. 2019; Sung et al. 2018). Meta-learning refers to the process of improving a learning algorithm with multiple learning episodes **(episodic training)**. These learning episodes are a distribution of tasks and not data samples. This improved learning ability has then been applied to the FSL realm. For instance, MAML (Antoniou et al. 2019; Finn et al. 2017), arguably the most popular meta-learning method, tries to train a model such that it can be fine-tuned end-to-end using only a few supervised samples while retaining high generalization ability.
Meta-learning approaches are mainly divided into **optimization-based**, **model-based**, or **metric-based**. **Optimization-based meta-learning** methods focus on finding an optimal initialization

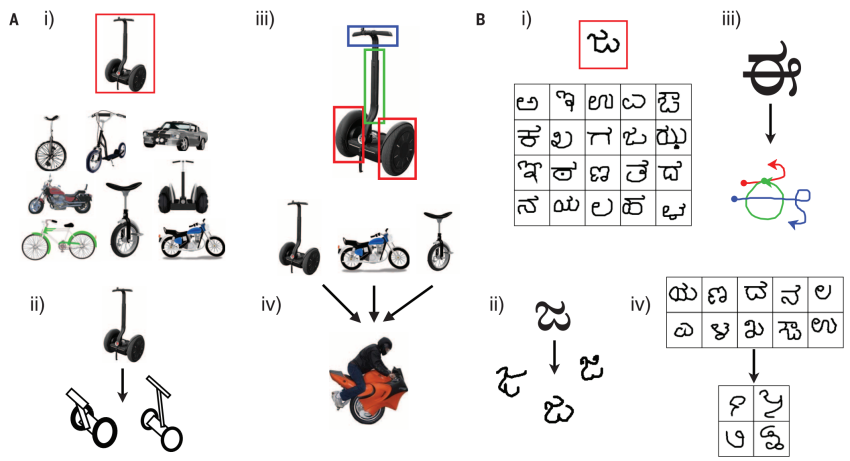

Figure 5.10: Few-shot learning paradigm. The objective is to leverage information from one or few annotated examples in order to perform many downstream tasks such as classification (i), generation of new examples (ii), segmentation and parsing (iii), new concepts generation. Figure from (Lake et al. 2015).

of model parameters, such that they can be fine-tuned efficiently with minimal supervision data (Finn et al. 2017; Ravi et al. 2017). **Model-based approaches** involve learning a model that can generate or adapt parameters for new tasks with the help of limited examples, often by using memory-augmented networks or modular architectures (Graves et al. 2014; N. Mishra et al. 2017). Lastly, **metric-based methods** rely on learning a similarity metric between instances, such that classification can be performed by comparing the relationships between few-shot examples and new instances in a latent space (Snell et al. 2017; Vinyals, Blundell, et al. 2016). Semi-supervised learning methods with few annotations also contribute to the FSL landscape, combining a small amount of labeled data with a larger pool of unlabeled data to improve performance on specific tasks (Oliver et al. 2018; Rasmus et al. 2015).

The majority of these methodologies have primarily been developed and tested within the realm of computer vision. Nonetheless, certain articles have shown that straightforward techniques rooted in transfer learning can competently compete with meta-learning approaches (Jiaxin Chen et al. 2020; Y. Tian, Yue Wang, et al. 2020). As a result, a significant number of modern investigations are centered around the **pre-training and efficient fine-tuning paradigm** as a means of developing effective methods for FSL (Jiaxin Chen et al. 2020). Similarly, in state-of-the-art NLP, FSL is predominantly executed through strategies that harness the power of Pre-trained Language Models.

Few-shot learning for NLP tasks using large language models

A significant body of research has addressed the challenge of FSL in NLP by leveraging Pre-trained Language Models (PLMs) (Devlin et al. 2019; Yinhan Liu et al. 2019; Radford, J. Wu, et al. 2019; Zhilin Yang et al. 2019). These approaches can be broadly categorized into three primary groups: **parameter-efficient tuning**, **prompt-based learning**, and **in-context learning**. Parameter-efficient tuning aligns with methods in the field of computer vision, introduced at the end of

previous paragraph, drawing heavily on the principles of transfer learning. On the other hand, the approaches of prompt-based learning and in-context learning are specific to the domain of NLP. They innovatively restructure tasks into natural language "prompts" and take advantage of Pre-trained Language Models (PLMs) to fill in these prompts.

Parameter-efficient tuning: These methods, such as adapters (Houlsby et al. 2019) have emerged as a promising solution for transfer learning and FSL in NLP tasks. These approaches involve adding lightweight, task-specific adapter layers to pre-trained transformer models, which allow for fine-tuning on limited labeled data while keeping the majority of the pre-trained model's parameters fixed (see Figure 5.11). Examples of such methods include AdapterHub (Pfeiffer et al. 2020), a framework for adapting transformers, and (D. Guo et al. 2021), referred to as "Diff-Pruning", accomplishing a similar objective by incorporating a sparse, task-specific difference vector to the original parameters. Moreover, in some cases, fine-tuning just a small fraction of the pre-trained model has proven to be effective. For instance, BitFit (Ben Zaken et al. 2022) only fine-tunes the bias parameters, which account for less than 1% of the total model parameters, yet it achieves competitive results on downstream tasks. More recently, T-FEW (Haokun Liu et al. 2022) proposed an approach consisting in adding learned vectors that rescale the network's internal activations.

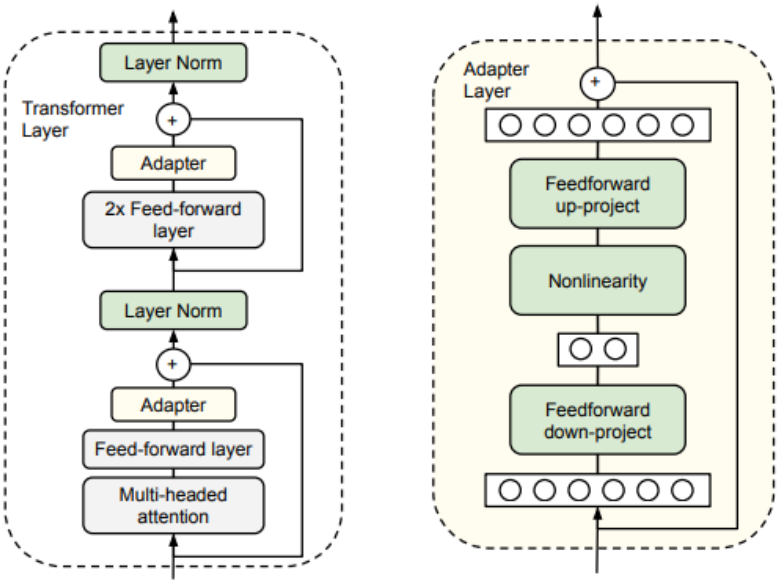

Figure 5.11: Adapter architecture (right) and its integration in Transformer (left). The Adapter consists in few-parameter modules that are inserted after Transformer FFN. When fine-tuning the modified architecture on a downstream tasks, only green modules (within Adapter and Layer Normalization) are updated. Figure from (Houlsby et al. 2019).

Prompt-Based Few-Shot Learning: In recent years, Pre-trained Language Models (PLMs) have been used to solve FSL tasks in NLP, notably using a prompting strategy. The idea is to frame the task as a language modeling problem by designing a template that guides the model towards generating a desired output. The seminal work (Schick et al. 2020) formalizes the prompt setting

by defining the template as pattern-verbalizer pairs, in which the pattern is a function mapping a set of input sentences to a cloze question. Verbalizers, on the other hand, are injective functions that map discrete labels into natural language phrases or tokens. This association leverages the generation capability of PLMs to perform classification tasks using a template, allowing the classification task to be formatted in a way that is intelligible to the PLM (Ding et al. 2022; P. Liu et al. 2023). This framework is illustrated in Figure 5.12. By varying the patterns and verbalizers, it is then possible to annotate a larger unlabeled dataset with soft labels, on which a classic classifier will be fine-tuned.

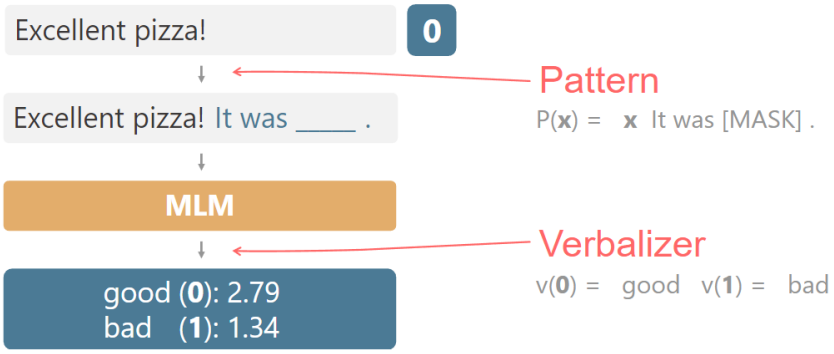

Figure 5.12: Prompt-based few-shot learning. The considered objective is to classify the input sentence "Excellent pizza!" as good or bad. The pattern P is first transforming the input as a cloze question $P(x)$. $P(x)$ is then fed to a PLM that outputs prediction scores for the masked word. Eventually, the verbalizer v converts the token prediction scores as classification logits.[5]

In-Context Learning: GPT3 (Brown et al. 2020), GPT4 (OpenAI 2023) and related chatbot ChatGPT based on InstructGPT model (Ouyang et al. 2022) showed that PLMs were also efficient for in-context FSL tasks. In this setting, the prompt is composed of the task description, but also some support input examples with their corresponding outputs and a query input with the objective to predict the query output (Wei, Xuezhi Wang, et al. 2022). ICL hence requires no parameter update, produces a new prediction model for each new prompting, and therefore quickly adapts to a new task (see Figure 5.9).

Inductive vs transductive few-shot learning

Learning an inductive classifier on embeddings generated by a pre-trained model, as proposed by (Snell et al. 2017), is a common baseline for performing FSL. This approach is prevalent in NLP, where a parametric model is trained on data to infer general rules that are applied to label new, unseen data (known as inductive learning (V. N. Vapnik 1999)). However, in FSL scenarios with limited labeled data, this approach can be highly ambiguous and lead to poor generalization. Transduction offers an attractive alternative to inductive learning (Sain 1996). Unlike inductive learning, which infers general rules from training data, transduction involves finding rules that work specifically for the unlabeled test data. By utilizing more data, such as unlabeled test instances, and

[5] Figure from http://timoschick.com/explanatory%20notes/2020/10/23/pattern-exploiting-training.html

aiming for a more localized rule rather than a general one (see Figure 5.13), transductive learning has shown promise and practical benefits in FSL for computer vision (Dhillon et al. 2020; Y. Guo et al. 2020; R. Hou et al. 2019; S. X. Hu et al. 2020; Y. Hu et al. 2021; J. Liu et al. 2020; Yanbin Liu et al. 2019; Yaoyao Liu et al. 2020; Qiao et al. 2019; Veilleux et al. 2021; Yikai Wang et al. 2020; Ling Yang et al. 2020; Ziko et al. 2020).

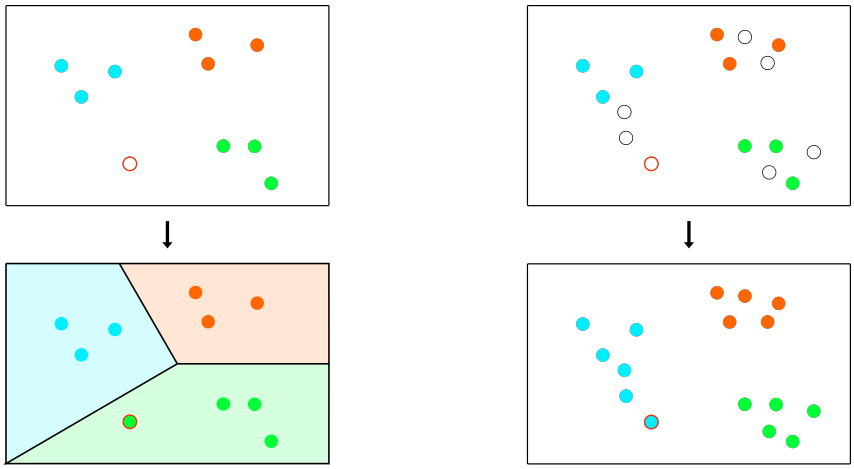

Figure 5.13: Inductive vs transductive settings. In the inductive setting (left), the model aims to learn general rules from labeled data, that will then serve to classify all unlabeled test samples, one by one. In the transductive setting (right), the model leverages information from both labeled data and all available unlabeled samples to adapt its classification to these samples. In this example, the same datapoint represented by a red circle is not classified the same way by the two approaches.

Transductive methods yield substantially better performance than their inductive counterparts by leveraging the statistics of the unlabeled data (such as batch normalization statistics (Nichol et al. 2018)). While (R. Hou et al. 2019; Yanbin Liu et al. 2019) use graphs or cross-attention modules to perform label propagation from support to query samples, other main strategies consist in minimizing the entropy of query samples predictions (Dhillon et al. 2020), using prototype rectification (J. Liu et al. 2020), Laplacian regularization (Ziko et al. 2020), optimal transport (Y. Hu et al. 2021), or maximizing Mutual Information measures (Boudiaf et al. 2020; Y. Guo et al. 2020; Veilleux et al. 2021). However, despite their success experienced in the vision community, this framework has not yet been explored in the context of textual data.

Conclusion

In conclusion, this chapter provided a comprehensive overview of the evolution and current state of NLP, delving into the various methodologies and techniques that have shaped the field. We began with early NLP approaches, including rule-based methods, vector space models, and probabilistic frameworks, before moving on to the groundbreaking de-

velopment of word embeddings that significantly advanced the state-of-the-art. The chapter then explored the emergence of language models and the attention mechanism, which have led to the transformative introduction of transformer architectures.
Large PLMs have revolutionized NLP by providing general-purpose contextual word representations that have greatly improved performance across a wide range of tasks. The pre-training and fine-tuning paradigm has proven highly successful, and has further pushed the boundaries of what is possible in NLP. However, these advancements based on the scaling paradigm require huge computational resources and available annotated data for fine-tuning. To handle this challenge, an interest in Few-shot Learning for NLP has grown. If universal efficient transfer-learning-based have been explored, new NLP-specific FSL paradigms have been developed, based on natural language prompts, and leveraging PLMs generation ability. Yet, they may not be suitable for realistic assumptions. A possible solution could be the use of transductive paradigm, that has not been explored in NLP. This is the main focus of Chapter 6.

Bibliography

Abed, Wathiq (2015). "A Robust Bearing Fault Detection and Diagnosis Technique for Brushless DC Motors Under Non-stationary Operating Conditions". *JCAES*.

Achille, Alessandro and Stefano Soatto (2018). "Emergence of invariance and disentanglement in deep representations". *The Journal of Machine Learning Research*.

Akhbardeh, Farhad, Travis Desell, and Marcos Zampieri (2020). "NLP Tools for Predictive Maintenance Records in MaintNet". 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: System Demonstrations, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*.

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. (2022). "Flamingo: a visual language model for few-shot learning". *Advances in Neural Information Processing Systems*.

Alayrac, Jean-Baptiste, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman (2020). "Self-supervised multimodal versatile networks". *Advances in Neural Information Processing Systems*.

Alemi, Alexander A., Ian Fischer, Joshua V. Dillon, and Kevin Murphy (2017). "Deep Variational Information Bottleneck". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Androutsopoulos, Ion, Graeme D Ritchie, and Peter Thanisch (1995). "Natural language interfaces to databases–an introduction". *Natural language engineering*.

Angelopoulos, Angelos, Emmanouel T Michailidis, Nikolaos Nomikos, Panagiotis Trakadas, Antonis Hatziefremidis, Stamatis Voliotis, and Theodore Zahariadis (2019). "Tackling faults in the industry 4.0 era—a survey of machine-learning solutions and key aspects". *Sensors*.

Antoniou, Antreas, Harrison Edwards, and Amos J. Storkey (2019). "How to train your MAML". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Arivazhagan, Naveen, Colin Cherry, Wolfgang Macherey, Chung Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel (2020). "Monotonic infinite lookback attention for simultaneous machine translation". *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

Atrey, Pradeep K, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli (2010). "Multimodal fusion for multimedia analysis: a survey". *Multimedia systems*.

Bachman, Philip, R Devon Hjelm, and William Buchwalter (2019). "Learning representations by maximizing mutual information across views". *Advances in neural information processing systems*.

Bagher Zadeh, AmirAli, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency (2018). "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. (2022). "Constitutional AI: Harmlessness from AI Feedback". *arXiv preprint arXiv:2212.08073*.

Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis Philippe Morency (2019). "Multimodal Machine Learning: A Survey and Taxonomy". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Barbieri, Francesco, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves (2020). "TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Proceedings of Findings of EMNLP*.

Ben Zaken, Elad, Yoav Goldberg, and Shauli Ravfogel (2022). "BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation learning: A review and new perspectives". *IEEE transactions on pattern analysis and machine intelligence*.

Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent (2000). "A neural probabilistic language model". *Advances in neural information processing systems*.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information". *Transactions of the Association for Computational Linguistics*.

Bommasani, Rishi, 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*.

Boudiaf, Malik, Ziko Imtiaz Masud, Jérôme Rony, Jose Dolz, Pablo Piantanida, and Ismail Ben Ayed (2020). "Transductive Information Maximization for Few-Shot Learning". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models Are Few-Shot Learners". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran (2012). "Distributional Semantics in Technicolor". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bullinaria, John A and Joseph P Levy (2007). "Extracting semantic representations from word co-occurrence statistics: A computational study". *Behavior research methods*.

Bunescu, Razvan and Raymond Mooney (2005). "A Shortest Path Dependency Kernel for Relation Extraction". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Cardoso, J-F (1997). "Infomax and maximum likelihood for blind source separation". *IEEE Signal processing letters*.

Casanueva, Iñigo, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić (2020). "Efficient intent detection with dual sentence encoders". *arXiv preprint arXiv:2003.04807*.

Castro, Santiago, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria (2019). "Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper)". 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*.

Chen, Baotong, Jiafu Wan, Lei Shu, Peng Li, Mithun Mukherjee, and Boxing Yin (2017). "Smart factory of industry 4.0: Key technologies, application case, and challenges". *Ieee Access*.

Chen, Jiaxin, Xiao-Ming Wu, Yanke Li, Qimai LI, Li-Ming Zhan, and Fu-lai Chung (2020). "A Closer Look at the Training Strategy for Modern Meta-Learning". In: *Advances in Neural Information Processing Systems*.

Chen, Junkun, Mingbo Ma, Renjie Zheng, and Liang Huang (2020). "Mam: Masked acoustic modeling for end-to-end speech-to-text translation". *arXiv preprint arXiv:2010.11445*.

Chen, Mark, 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*.

Chen, Stanley F. and Joshua Goodman (1996). "An Empirical Study of Smoothing Techniques for Language Modeling". In: Association for Computational Linguistics.

Chen, Yanda, Ruiqi Zhong, Sheng Zha, George Karypis, and He He (2022). "Meta-learning via Language Model In-context Tuning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*.

Cheng, Yiwei, Haiping Zhu, Jun Wu, and Xinyu Shao (2019). "Machine Health Monitoring Using Adaptive Kernel Spectral Clustering and Deep Long Short-Term Memory Recurrent Neural Networks". *IEEE Transactions on Industrial Informatics*.

Child, Rewon, Scott Gray, Alec Radford, and Ilya Sutskever (2019). "Generating long sequences with sparse transformers". *arXiv preprint arXiv:1904.10509*.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "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)*.

Chomsky, N. (1956). "Three models for the description of language". *IRE Transactions on Information Theory*.

Chorowski, Jan, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio (2015). "Attention-Based Models for Speech Recognition". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*.

Choudhary, Anurag, SL Shimi, and Aparna Akula (2018). "Bearing fault diagnosis of induction motor using thermal imaging". In: *2018 international conference on computing, power and communication technologies (GUCON)*.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. (2022). "Palm: Scaling language modeling with pathways". *arXiv preprint arXiv:2204.02311*.

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017). "Deep reinforcement learning from human preferences". *Advances in neural information processing systems*.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. (2022). "Scaling instruction-finetuned language models". *arXiv preprint arXiv:2210.11416*.

Chung, Joon Son and Andrew Zisserman (2016). "Out of Time: Automated Lip Sync in the Wild". In: *Computer Vision - ACCV 2016 Workshops - ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*. Ed. by Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle (2021). "Using statistics in lexical analysis". In: *Lexical acquisition: exploiting on-line resources to build a lexicon*.

Church, Kenneth Ward and Patrick Hanks (1990). "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics*.

Commission, European (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*.

– (2020). *Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act)*.

Conneau, Alexis, 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, ACL 2020, Online, July 5-10, 2020*.

Cover, Thomas M (1999). *Elements of information theory*. John Wiley & Sons.

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov (2019). "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context". 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*.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990). "Indexing by latent semantic analysis". *Journal of the American Society for Information Science*.

Degottex, Gilles, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer (2014). "COVAREP - A collaborative voice analysis repository for speech technologies". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*.

Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi (2020). "GoEmotions: A Dataset of Fine-Grained Emotions". In: *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*.

Dharma, EDDY MUNTINA, F Lumban Gaol, HLHS Warnars, and BENFANO Soewito (2022). "The accuracy comparison among Word2vec, Glove, and Fasttext towards convolution neural network (CNN) text classification". *Journal of Theoretical and Applied Information Technology*.

Dhillon, Guneet Singh, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto (2020). "A Baseline for Few-Shot Image Classification". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Diaz Rozo, Javier et al. (2017). "Machine Learning-based CPS for Clustering High throughput Machining Cycle Conditions". *Procedia Manufacturing*.

Ding, Ning, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun (2022). "OpenPrompt: An Open-source Framework for Prompt-learning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Elman, Jeffrey L. (1990). "Finding structure in time". *Cognitive Science*.

Esteva, Andre, Brett Kuprel, Roberto Novoa, Justin Ko, Susan Swetter, Helen Blau, and Sebastian Thrun (2017). "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature*.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. (2021). "Beyond english-centric multilingual machine translation". *The Journal of Machine Learning Research*.

Federici, Marco, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata (2020). "Learning Robust Representations via Multi-View Information Bottleneck". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Fei, Yu, Zhao Meng, Ping Nie, Roger Wattenhofer, and Mrinmaya Sachan (2022). "Beyond prompting: Making Pre-trained Language Models Better Zero-shot Learners by Clustering Representations". In: *Proceedings of the• 2022 Conference on Empirical Methods in Natural Language Processing*.

Fiérrez-Aguilar, Julian, Javier Ortega-Garcia, Daniel Garcia-Romero, and Joaquin Gonzalez-Rodriguez (2003). "A comparative evaluation of fusion strategies for multimodal biometric verification". In: *International Conference on Audio-and Video-Based Biometric Person Authentication*.

Finkel, Jenny Rose, Trond Grenager, and Christopher D Manning (2005). "Incorporating non-local information into information extraction systems by gibbs sampling". In: *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*.

Finn, Chelsea, Pieter Abbeel, and Sergey Levine (2017). "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International conference on machine learning*.

Firth, John (1957). "A synopsis of linguistic theory, 1930-1955". *Studies in linguistic analysis*.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky (2016). "Domain-Adversarial Training of Neural Networks". *J. Mach. Learn. Res.*

Gao, Tianyu, Adam Fisch, and Danqi Chen (2021). "Making Pre-trained Language Models Better Few-shot Learners". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*.

Gao, Tianyu, Xu Han, Zhiyuan Liu, and Maosong Sun (2019). "Hybrid attention-based prototypical networks for noisy few-shot relation classification". In: *Proceedings of the AAAI conference on artificial intelligence*.

Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra (2023). "ImageBind: One Embedding Space To Bind Them All". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Glaese, Amelia, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. (2022). "Improving alignment of dialogue agents via targeted human judgements". *arXiv preprint arXiv:2209.14375*.

Goldstein, Markus and Seiichi Uchida (2016). "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data". *PloS one*.

Gomes, Eduardo Dadalto Câmara, Florence Alberge, Pierre Duhamel, and Pablo Piantanida (2022). "Igeood: An Information Geometry Approach to Out-of-Distribution Detection". In: *The Tenth International Conference on Learning Representations*.

Goodfellow, Ian J., Yoshua Bengio, and Aaron C. Courville (2016). *Deep Learning*. Adaptive computation and machine learning. MIT Press.

Grandvalet, Yves and Yoshua Bengio (2004). "Semi-supervised learning by entropy minimization". *Advances in neural information processing systems*.

Graves, Alex (2013). "Generating sequences with recurrent neural networks". *arXiv preprint arXiv:1308.0850*.

Graves, Alex, Greg Wayne, and Ivo Danihelka (2014). "Neural turing machines". *arXiv preprint arXiv:1410.5401*.

Grigorescu, Sorin, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu (2020). "A survey of deep learning techniques for autonomous driving". *Journal of Field Robotics*.

Gulshan, Varun, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. (2016). "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". *Jama*.

Guo, Demi, Alexander Rush, and Yoon Kim (2021). "Parameter-Efficient Transfer Learning with Diff Pruning". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Guo, Liang et al. (2017). "A recurrent neural network based health indicator for remaining useful life prediction of bearings". *Neurocomputing*.

Guo, Wenzhong, Jianwen Wang, and Shiping Wang (2019). "Deep Multimodal Representation Learning: A Survey". *IEEE Access*.

Guo, Yiluan and Ngai-Man Cheung (2020). "Attentive weights generation for few shot learning via information maximization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Han, Wei, Hui Chen, and Soujanya Poria (2021). "Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Harris, Zellig S (1954). "Distributional structure". *Word*.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen (2021). "Deberta: decoding-Enhanced Bert with Disentangled Attention". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Henaff, Olivier (2020). "Data-efficient image recognition with contrastive predictive coding". In: *International conference on machine learning*.

Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". *IEEE Signal Processing Magazine*.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". *science*.

Hinton, Geoffrey E., Oriol Vinyals, and Jeffrey Dean (2015). "Distilling the Knowledge in a Neural Network". *ArXiv*.

Hjelm, R. Devon, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio (2019). "Learning deep representations by mutual information estimation and maximization". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Hochreiter, Sepp, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. (2001). *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". *Neural computation*.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. (2022). "Training compute-optimal large language models". *arXiv preprint arXiv:2203.15556*.

Hospedales, Timothy, Antreas Antoniou, Paul Micaelli, and Amos Storkey (2021). "Meta-learning in neural networks: A survey". *IEEE transactions on pattern analysis and machine intelligence*.

Hou, Ruibing, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen (2019). "Cross attention network for few-shot classification". *Advances in Neural Information Processing Systems*.

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019). "Parameter-efficient transfer learning for NLP". In: *International Conference on Machine Learning*.

Howard, Jeremy and Sebastian Ruder (2018). "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Hu, Shell Xu, Pablo Garcia Moreno, Yang Xiao, Xi Shen, Guillaume Obozinski, Neil D. Lawrence, and Andreas C. Damianou (2020). "Empirical Bayes Transductive Meta-Learning with Synthetic Gradients". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Hu, Weihua, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama (2017). "Learning Discrete Representations via Information Maximizing Self-Augmented Training". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*.

Hu, Yuqing, Vincent Gripon, and Stéphane Pateux (2021). "Leveraging the feature distribution in transfer-based few-shot learning". In: *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II 30*.

Huang, Yu, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang (2021). "What makes multi-modal learning better than single (provably)". *Advances in Neural Information Processing Systems*.

Huang, Zhiheng, Davis Liang, Peng Xu, and Bing Xiang (2020). "Improve Transformer Models with Better Relative Position Embeddings". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Hutchins, W. John (2004). "The Georgetown-IBM Experiment Demonstrated in January 1954". In: *Machine Translation: From Real Users to Research*.

Hutchins, William John (1986). *Machine translation: past, present, future*. Citeseer.

Isermann, Rolf (2005). *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media.

Jafar, Raed et al. (2010). "Application of Artificial Neural Networks (ANN) to model the failure of urban water mains". *Mathematical and Computer Modelling*.

Janssens, Olivier et al. (2015). "Thermal image based fault diagnosis for rotating machinery". *Infrared Physics and Technology*.

Jelinek, Frederick (1991). "Principles of lexical language modeling for speech recognition". *Advances in speech signal processing*.

Ji, Xu, Joao F Henriques, and Andrea Vedaldi (2019). "Invariant information clustering for unsupervised image classification and segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Jia, Feng et al. (2015). "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data". *Mechanical Systems and Signal Processing*.

Johnson, Alistair EW, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark (2016). "MIMIC-III, a freely accessible critical care database". *Scientific data*.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman,

Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis (2021). "Highly accurate protein structure prediction with AlphaFold". *Nature*.

Jurafsky, Dan (2000). *Speech & language processing*. Pearson Education India.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). "Scaling laws for neural language models". *arXiv preprint arXiv:2001.08361*.

Katharopoulos, Angelos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret (2020). "Transformers are rnns: Fast autoregressive transformers with linear attention". In: *International Conference on Machine Learning*.

Ke, Guolin, Di He, and Tie-Yan Liu (2021). "Rethinking Positional Encoding in Language Pretraining". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Keung, Phillip, Yichao Lu, György Szarvas, and Noah A. Smith (2020). "The Multilingual Amazon Reviews Corpus". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kirkpatrick, James, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. (2017). "Overcoming catastrophic forgetting in neural networks". *Proceedings of the national academy of sciences*.

Kitaev, Nikita, Lukasz Kaiser, and Anselm Levskaya (2020). "Reformer: The Efficient Transformer". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. (2023). "Chatgpt: Jack of all trades, master of none". *arXiv preprint arXiv:2302.10724*.

Koehn, Philipp, Franz J. Och, and Daniel Marcu (2003). "Statistical Phrase-Based Translation". In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Konar, Pratyay et al. (2009). "Bearing Fault Detection of Induction Motor using Wavelet and Neural Networks." In:

Lachs, Lorin (2017). "Multi-modal perception". *Noba textbook series: Psychology. Champaign: DEF Publishers*.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*.

Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum (2015). "Human-level concept learning through probabilistic program induction". *Science*.

Lan, Zhen-zhong, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann (2014). "Multimedia classification and event detection using double fusion". *Multimedia tools and applications*.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Larson, Stefan, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars (2019). "An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction". 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)*.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". *nature*.

Lee, Kwonjoon, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto (2019). "Meta-learning with differentiable convex optimization". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Lee, Mihee and Vladimir Pavlovic (2021). "Private-shared disentangled multimodal vae for learning of latent representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lehman, Eric, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer (2023). "Do We Still Need Clinical Language Models?" *arXiv preprint arXiv:2302.08091*.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (2020). "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, ACL 2020, Online, July 5-10, 2020*.

Lhoest, Quentin, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf (2021). "Datasets: A Community Library for Natural Language Processing". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*.

Li, Gen, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang (2020). "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training". In: *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang (2019). "Visualbert: A simple and performant baseline for vision and language". *arXiv preprint arXiv:1908.03557*.

Li, Ruilong, Shan Yang, David A Ross, and Angjoo Kanazawa (2021). "Ai choreographer: Music conditioned 3d dance generation with aist++". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li, Zhe (2018). "Deep learning driven approaches for predictive maintenance: A framework of intelligent fault diagnosis and prognosis in the industry 4.0 era".

Liang, Paul Pu, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency (2023). "Quantifying & Modeling Feature Interactions: An Information Decomposition Framework". *arXiv preprint arXiv:2302.12247*.

Liang, Paul Pu, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency (2021). "MultiBench: Multiscale Benchmarks for Multimodal Representation Learning". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Liang, Tianchen et al. (2018). "Bearing fault diagnosis based on improved ensemble learning and deep belief network". *Journal of Physics*.

Lichtenstein, Moshe, Prasanna Sattigeri, Rogerio Feris, Raja Giryes, and Leonid Karlinsky (2020). "Tafssl: Task-adaptive feature sub-space learning for few-shot classification". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII*.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick (2014). "Microsoft coco: Common objects in context". In: *European conference on computer vision*.

Linsker, Ralph (1988). "Self-Organization in a Perceptual Network". *Computer*.

Littlewort, Gwen, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett (2011). "The computer expression recognition toolbox (CERT)". In: *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*.

Liu, Han et al. (2018). "Unsupervised fault diagnosis of rolling bearings using a deep neural network based on generative adversarial networks". *Neurocomputing*.

Liu, Haokun, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel (2022). "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning". In: *NeurIPS*.

Liu, Jinlu, Liang Song, and Yongqiang Qin (2020). "Prototype rectification for few-shot learning". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig (2023). "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing". *ACM Computing Surveys*.

Liu, Yanbin, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang (2019). "Learning to Propagate Labels: Transductive Propagation Network for Few-Shot Learning". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Liu, Yaoyao, Bernt Schiele, and Qianru Sun (2020). "An ensemble of epoch-wise empirical bayes for few-shot learning". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*.

Liu, Yinhan, 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*.

Liu, Yueh-Cheng, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H Hsu (2021). "Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining". *arXiv preprint arXiv:2104.04687*.

Liu, Yukun et al. (2010). "Application to induction motor faults diagnosis of the amplitude recovery method combined with FFT". *Mechanical Systems and Signal Processing*.

Liu, Yunze, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi (2021). "Contrastive multimodal fusion with tupleinfonce". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Logan IV, Robert, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel (2022). "Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models". In: *Findings of the Association for Computational Linguistics: ACL 2022*.

Long, Xiang, Chuang Gan, Gerard De Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen (2018). "Multimodal keyless attention fusion for video classification". *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.

Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee (2019). "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*.

Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp (2022). "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Luhn, Hans Peter (1957). "A statistical approach to mechanized encoding and searching of literary information". *IBM Journal of research and development*.

Lund, Kevin and Curt Burgess (1996). "Producing high-dimensional semantic spaces from lexical co-occurrence". *Behavior research methods, instruments, & computers*.

Luo, Bo, Haoting Wang, Hongqi Liu, Bin Li, and Fangyu Peng (2018). "Early fault detection of machine tools based on deep learning and dynamic identification". *IEEE Transactions on Industrial Electronics*.

Luong, Thang, Hieu Pham, and Christopher D. Manning (2015). "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Ma, Xutai, Juan Miguel Pino, James Cross, Liezl Puzon, and Jiatao Gu (2020). "Monotonic Multihead Attention". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Mahabadi, Rabeeh Karimi, Luke Zettlemoyer, James Henderson, Marzieh Saeidi, Lambert Mathias, Veselin Stoyanov, and Majid Yazdani (2022). "PERFECT: Prompt-free and efficient few-shot learning with language models". *arXiv preprint arXiv:2204.01172*.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to information retrieval*. Cambridge University Press.

Manning, Christopher D. and Hinrich Schütze (2001). *Foundations of statistical natural language processing*. MIT Press.

Maragos, Petros, Alexandros Potamianos, and Patrick Gros (2008). *Multimodal Processing and Interaction, Audio, Video, Text*. Springer.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). "Building a Large Annotated Corpus of English: The Penn Treebank". *Computational Linguistics*.

McCarthy, John, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon (2006). "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955". *AI magazine*.

Mian, Tauheed, Anurag Choudhary, and Shahab Fatima (2022). "A sensor fusion based approach for bearing fault diagnosis of rotating machine". *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*.

Miech, Antoine, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman (2020). "End-to-End Learning of Visual Representations From Uncurated Instructional Videos". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, Tomás, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur (2010). "Recurrent neural network based language model". In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*.

Mikolov, Tomás, Quoc V. Le, and Ilya Sutskever (2013). "Exploiting Similarities among Languages for Machine Translation". *CoRR*.

Min, Sewon, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi (2022). "MetaICL: Learning to Learn In Context". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*.

Min, Sewon, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer (2022). "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?" In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*.

Mishra, Nikhil, Mostafa Rohaninejad, Xi Chen, and P. Abbeel (2017). "A Simple Neural Attentive Meta-Learner". In: *International Conference on Learning Representations*.

Mobley, R Keith (2002). *An introduction to predictive maintenance*. Elsevier.

Mohri, Mehryar (1997). "Finite-state transducers in language and speech processing". *Computational linguistics*.

Moritz, Niko, Takaaki Hori, and Jonathan Le Roux (2020). "Streaming automatic speech recognition with the transformer model". *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Muennighoff, Niklas, Nouamane Tazi, Loic Magne, and Nils Reimers (2023). "MTEB: Massive Text Embedding Benchmark". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*.

Neti, Chalapathy, Benoit Maison, Andrew W Senior, Giridharan Iyengar, P Decuetos, Sankar Basu, and Ashish Verma (2000). "Joint processing of audio and visual information for multimedia indexing and human-computer interaction." In: *RIAO*.

Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng (2011). "Multimodal Deep Learning". *Proceedings of the 28th International Con- ference on Machine Learning*.

Nichol, Alex, Joshua Achiam, and John Schulman (2018). "On first-order meta-learning algorithms". *arXiv preprint arXiv:1803.02999*.

Nor, Norazwan et al. (2019). "A review of data-driven fault detection and diagnosis methods: Applications in chemical process systems". *Reviews in Chemical Engineering*.

Oliver, Avital, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow (2018). "Realistic evaluation of deep semi-supervised learning algorithms". *Advances in neural information processing systems*.

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation learning with contrastive predictive coding". *arXiv preprint arXiv:1807.03748*.

OpenAI (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

Ouali, Yassine (2023). "Learning with Limited Labeled Data". PhD thesis. Université Paris-Saclay.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. (2022). "Training language models to follow instructions with human feedback". *Advances in Neural Information Processing Systems*.

Palade, Vasile and Cosmin Danut Bocaniala (2006). *Computational intelligence in fault diagnosis*. Springer Science & Business Media.

Pan, Jun et al. (2017). "LiftingNet: A Novel Deep Learning Network With Layerwise Feature Learning From Noisy Mechanical Data for Fault Classification". *IEEE TIE*.

Pan, Sinno Jialin and Qiang Yang (2010). "A survey on transfer learning". *IEEE Transactions on knowledge and data engineering*.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Pellegrain, Victor, Michel Batteux, William Lair, and Michel Kaczmarek (2022). "Démonstration de surveillance de défaillances sur un exemple applicatif". In: *Congrès Lambda Mu 23 «Innovations et maîtrise des risques pour un avenir durable»-23e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Institut pour la Maîtrise des Risques*.

Pellegrain, Victor, Myriam Tami, Michel Batteux, Céline Hudelot, and IRT SystemX (2022). "Apprentissage multimodal pour le diagnostic de fautes sur données séquentielles non alignées et arbitrairement longues". In: *Conférence Nationale d'Intelligence Artificielle Année 2022*.

Peng, Ying et al. (2010). "Current status of machine prognostics in condition-based maintenance: A review". *International Journal of Advanced Manufacturing Technology*.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*.

Perez, Ethan, Douwe Kiela, and Kyunghyun Cho (2021). "True Few-Shot Learning with Language Models". In: *Advances in Neural Information Processing Systems*.

Pérez-Rosas, Verónica, Rada Mihalcea, and Louis-Philippe Morency (2013). "Utterance-level multimodal sentiment analysis". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "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)*.

Pfeiffer, Jonas, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych (2020). "AdapterHub: A Framework for Adapting Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*.

Picot, Marine, Francisco Messina, Malik Boudiaf, Fabrice Labeau, Ismail Ben Ayed, and Pablo Piantanida (2023). "Adversarial Robustness Via Fisher-Rao Regularization". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Poole, Ben, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker (2019). "On variational bounds of mutual information". In: *International Conference on Machine Learning*.

Poria, Soujanya, Iti Chaturvedi, Erik Cambria, and Amir Hussain (2016). "Convolutional MKL based multimodal emotion recognition and sentiment analysis". In: *2016 IEEE 16th international conference on data mining (ICDM)*.

Porter, Martin F (1980). "An algorithm for suffix stripping". *Program*.

Qiao, Limeng, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian (2019). "Transductive episodic-wise adaptive metric for few-shot learning". In: *Proceedings of the IEEE/CVF international conference on computer vision*.

Rabiner, Lawrence R (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE*.

Radford, Alec, 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*.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2022). "Robust Speech Recognition via Large-Scale Weak Supervision". *CoRR* abs/2212.04356.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. (2018a). "Improving language understanding by generative pre-training".

– (2018b). "Improving language understanding by generative pre-training".

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language models are unsupervised multitask learners". *OpenAI blog*.

Raffel, Colin, Minh Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck (2017). "Online and linear-time attention by enforcing monotonic alignments". *34th International Conference on Machine Learning, ICML 2017*.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". *The Journal of Machine Learning Research*.

Raghu, Aniruddh, Maithra Raghu, Samy Bengio, and Oriol Vinyals (2020). "Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Raghu, Maithra and Eric Schmidt (2020). "A Survey of Deep Learning for Scientific Discovery". *CoRR*.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen (2022). "Hierarchical text-conditional image generation with clip latents". *arXiv preprint arXiv:2204.06125*.

Rasmus, Antti, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko (2015). "Semi-supervised learning with ladder networks". *Advances in neural information processing systems*.

Ravi, Sachin and Hugo Larochelle (2017). "Optimization as a model for few-shot learning". In: *International conference on learning representations*.

Reid, Alistair et al. (2013). "Fault Location and Diagnosis in a Medium Voltage EPR Power Cable". *IEEE TDEI*.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". 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)*.

Reis, Marco S. and Geert Gins (2017). "Industrial Process Monitoring in the Big Data/Industry 4.0 Era: from Detection, to Diagnosis, to Prognosis". *Processes*.

Rogers, A.P. et al. (2019). "A review of fault detection and diagnosis methods for residential air conditioning systems". *Building and Environment*.

Rosenfeld, Jonathan S., Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit (2020). "A Constructive Prediction of the Generalization Error Across Scales". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Ruder, Sebastian, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf (2019). "Transfer Learning in Natural Language Processing". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*.

Rush, Alexander M., Sumit Chopra, and Jason Weston (2015). "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Rusu, Andrei, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell (2019). "Meta-Learning with Latent Embedding Optimization". In: *International Conference on Learning Representations (ICLR)*.

Ryalat, Mutaz, Hisham ElMoaqet, and Marwa AlFaouri (2023). "Design of a smart factory based on cyber-physical systems and internet of things towards industry 4.0". *Applied Sciences*.

Sain, Stephan R (1996). *The nature of statistical learning theory*.

Salton, Gerard, Anita Wong, and Chung-Shu Yang (1975). "A vector space model for automatic indexing". *Communications of the ACM*.

Sanh, Victor, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Ur-

mish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *The Tenth International Conference on Learning Representations, 2022, Virtual Event, April 25-29, 2022*.

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. (2022). "Bloom: A 176b-parameter open-access multilingual language model". *arXiv preprint arXiv:2211.05100*.

Schick, Timo and Hinrich Schütze (2020). "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference". In: *Conference of the European Chapter of the Association for Computational Linguistics*.

– (2021). "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners". 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*.

– (2022). "True Few-Shot Learning with Prompts—A Real-World Perspective". *Transactions of the Association for Computational Linguistics*.

Schwab, Klaus (2017). *The fourth industrial revolution*. Currency.

Sha, Fei and Fernando Pereira (2003). "Shallow parsing with conditional random fields". In: *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*.

Shannon, Claude E (1948). "A mathematical theory of communication". *The Bell system technical journal*.

Shao, Haidong et al. (2018). "A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders". *MSSP*.

Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani (2018). "Self-Attention with Relative Position Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Shi, Yangyang, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer (2021). "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Silberer, Carina and Mirella Lapata (2014). "Learning grounded meaning representations with autoencoders". *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*.

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. (2016). "Mastering the game of Go with deep neural networks and tree search". *Nature*.

Sipos, Ruben, Dmitriy Fradkin, Fabian Moerchen, and Zhuang Wang (2014). "Log-based predictive maintenance". In: *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*.

Sipple, John (2020). "Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*.

Snell, Jake, Kevin Swersky, and Richard Zemel (2017). "Prototypical networks for few-shot learning". *Advances in neural information processing systems*.

Soatto, Stefano and Alessandro Chiuso (2016). "Modeling Visual Representations: Defining Properties and Deep Approximations". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Solaiman, Irene (2023). "The Gradient of Generative AI Release: Methods and Considerations". *arXiv preprint arXiv:2302.04844*.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu (2020). "Mpnet: Masked and permuted pre-training for language understanding". *Advances in Neural Information Processing Systems*.

Song, Yisheng, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo (2022). "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities". *ACM Computing Surveys*.

Souza, Vinicius M.A., Denis M. dos Reis, André G. Maletzke, and Gustavo E.A.P.A. Batista (2020). "Challenges in benchmarking stream learning algorithms with real-world data". *Data Mining and Knowledge Discovery*.

Sparck Jones, Karen (1972). "A statistical interpretation of term specificity and its application in retrieval". *Journal of documentation*.

Sridharan, Karthik and Sham M. Kakade (2008). "An Information Theoretic Framework for Multi-view Learning". In: *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*.

Srivastava, Nitish and Ruslan Salakhutdinov (2012). "Learning representations for multimodal data with deep belief nets". *International Conference on Machine Learning Workshop*.

Srivastava, Nitish and Russ R Salakhutdinov (2012). "Multimodal learning with deep boltzmann machines". *Advances in neural information processing systems*.

Stein, Barry E and M Alex Meredith (1993). *The merging of the senses.* The MIT press.

Stokes, Jonathan M, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. (2020). "A deep learning approach to antibiotic discovery". *Cell*.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP". 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*.

Sun, Chen, Fabien Baradel, Kevin Murphy, and Cordelia Schmid (2019). "Contrastive Bidirectional Transformer for Temporal Representation Learning". *CoRR* abs/1906.05743.

Sun, Chen, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid (2019). "Videobert: A joint model for video and language representation learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Sun, Jiedi et al. (2017). "Intelligent Bearing Fault Diagnosis Method Combining Compressed Data Acquisition and Deep Learning". *IEEE TIM*.

Sun, Qianru, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele (2019). "Meta-transfer learning for few-shot learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sun, Shengli, Qingfeng Sun, Kevin Zhou, and Tengchao Lv (2019). "Hierarchical attention prototypical networks for few-shot text classification". 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)*.

Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney (2012). "LSTM neural networks for language modeling". In: *Thirteenth annual conference of the international speech communication association*.

Sung, Flood, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales (2018). "Learning to compare: Relation network for few-shot learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Sutskever, Ilya, James Martens, George Dahl, and Geoffrey Hinton (2013). "On the importance of initialization and momentum in deep learning". In: *Proceedings of the 30th International Conference on Machine Learning*.

Sutskever, Ilya, Oriol Vinyals, and Quoc Le (2014). "Sequence to Sequence Learning with Neural Networks". *Advances in Neural Information Processing Systems*.

Taheri-Garavand, Amin et al. (2015). "An intelligent approach for cooling radiator fault diagnosis based on infrared thermal image processing technique". *Applied Thermal Engineering*.

Taj, SM, SM Rizwan, BM Alkali, DK Harrison, and GL Taneja (2017). "Reliability analysis of a single machine subsystem of a cable plant with six maintenance categories". *International Journal of Applied Engineering Research*.

Talmor, Alon, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant (2021). "MultiModalQA: complex question answering over text, tables and images". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Tam, Derek, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel (2021). "Improving and Simplifying Pattern Exploiting Training". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*.

Tay, Yi, Mostafa Dehghani, Dara Bahri, and Donald Metzler (2022). "Efficient transformers: A survey". *ACM Computing Surveys*.

Taylor, Ross, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic (2022). "Galactica: A large language model for science". *arXiv preprint arXiv:2211.09085*.

Taylor, Wilson L (1953). ""Cloze procedure": A new tool for measuring readability". *Journalism quarterly*.

Teske, JJ, JC Liljegren, and DL Sisterson (2001). *Long-term analysis of the corrective maintenance records of the ARM SGP CART.* Technical report. Argonne National Lab., IL (US).

Thoppilan, Romal, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. (2022). "Lamda: Language models for dialog applications". *arXiv preprint arXiv:2201.08239*.

Thrun, Sebastian and Lorien Y. Pratt, eds. (1998). *Learning to Learn*. Springer.

Tian, Yonglong, Dilip Krishnan, and Phillip Isola (2020). "Contrastive multiview coding". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*.

Tian, Yonglong, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola (2020). "What makes for good views for contrastive learning?" *Advances in neural information processing systems*.

Tian, Yonglong, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola (2020). "Rethinking few-shot image classification: a good embedding is all you need?" In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*.

Tian, Zhengkun, Jiangyan Yi, Ye Bai, Jianhua Tao, Shuai Zhang, and Zhengqi Wen (2020). "Synchronous Transformers for end-to-end Speech Recognition". *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Tishby, Naftali, Fernando C Pereira, and William Bialek (2000). "The information bottleneck method". *arXiv preprint physics/0004057*.

Tosh, Christopher, Akshay Krishnamurthy, and Daniel Hsu (2021). "Contrastive learning, multiview redundancy, and linear models". In: *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. (2023). "Llama: Open and efficient foundation language models". *arXiv preprint arXiv:2302.13971*.

Tripathi, Anshuman, Jaeyoung Kim, Qian Zhang, Han Lu, and Hasim Sak (2020). "Transformer transducer: One model unifying streaming and non-streaming speech recognition". *arXiv preprint arXiv:2010.03192*.

Tsai, Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov (2019). "Multimodal transformer for unaligned multimodal language sequences". In: *Proceedings of the conference. Association for Computational Linguistics. Meeting*.

Turing, Alan M (2009). *Computing machinery and intelligence*. Springer.

Vapnik, Vladimir (2000). *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer.

Vapnik, Vladimir N (1999). "An overview of statistical learning theory". *IEEE transactions on neural networks*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". *Advances in neural information processing systems*.

Veilleux, Olivier, Malik Boudiaf, Pablo Piantanida, and Ismail Ben Ayed (2021). "Realistic evaluation of transductive few-shot learning". *Advances in Neural Information Processing Systems*.

Venkatasubramanian, Venkat et al. (2003). "A review of process fault detection and diagnosis. Part I: Quantitative model-based methods 27(3), 293–311. Part II: Qualitative models and search strategies 27(3), 313–32. Part III: Process history based methods 27(3), 327–346". *Computers & Chemical Engineering*.

Vinyals, Oriol, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, et al. (2019). "Alphastar: Mastering the real-time strategy game starcraft ii". *DeepMind blog*.

Vinyals, Oriol, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu koray, and Daan Wierstra (2016). "Matching Networks for One Shot Learning". In: *Advances in Neural Information Processing Systems*.

Wan, Zhibin, Changqing Zhang, Pengfei Zhu, and Qinghua Hu (2021). "Multi-view information-bottleneck representation learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, Chengyi, Yu Wu, Liang Lu, Shujie Liu, Jinyu Li, Guoli Ye, and Ming Zhou (2020). "Low Latency End-to-End Streaming Speech Recognition with a Scout Network". In: *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*.

Wang, Feng et al. (2016). "Bilevel feature extraction-based text mining for fault diagnosis of railway systems". *IEEE TITS*.

Wang, Jinjiang et al. (2019). "Machine vision intelligence for product defect inspection based on deep learning and Hough transform". *Journal of Manufacturing Systems*.

Wang, Qi, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou (2019). "Deep multi-view information bottleneck". In: *Proceedings of the 2019 SIAM International Conference on Data Mining*.

Wang, Sen, Xiaoqin Liu, Tangfeng Yang, and Xing Wu (2018). "Panoramic crack detection for steel beam based on structured random forests". *IEEE Access*.

Wang, Sinong, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma (2020). "Linformer: Self-attention with linear complexity". *arXiv preprint arXiv:2006.04768*.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou (2020). "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers". *Advances in Neural Information Processing Systems*.

Wang, Yikai, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu (2020). "Instance credibility inference for few-shot learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2022). "Finetuned Language Models are Zero-Shot Learners". In: *The Tenth International Conference on Learning Representations, 2022, Virtual Event, April 25-29, 2022*.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus (2022). "Emergent Abilities of Large Language Models". *Trans. Mach. Learn. Res.*

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *NeurIPS*.

Wen, Long et al. (2017). "A New Convolutional Neural Network Based Data-Driven Fault Diagnosis Method". *IEEE TIE*.

Wen, Long et al. (2019). "A New Snapshot Ensemble Convolutional Neural Network for Fault Diagnosis". *IEEE Access*.

Woods, William A (1977). "Lunar rocks in natural English: Explorations in natural language question answering."

Wu, Bingjie et al. (2021). "Simultaneous-fault diagnosis considering time series with a deep learning transformer architecture for air handling units". *Energy and Buildings*.

Wu, Chunyang, Yongqiang Wang, Yangyang Shi, Ching-Feng Yeh, and Frank Zhang (2020). "Streaming Transformer-Based Acoustic Models Using Self-Attention with Augmented Memory". In: *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*.

Wu, Mike, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah D. Goodman (2020). "On Mutual Information in Contrastive Learning for Visual Representations". *CoRR* abs/2005.13149.

Xia, Min et al. (2017). "Fault Diagnosis for Rotating Machinery Using Multiple Sensors and Convolutional Neural Networks". *IEEE/ASME Transactions on Mechatronics*.

Xie, Junyuan, Ross Girshick, and Ali Farhadi (2016). "Unsupervised deep embedding for clustering analysis". In: *International conference on machine learning*.

Xie, Yuan and Tao Zhang (2018). "Imbalanced learning for fault diagnosis problem of rotating machinery based on generative adversarial networks". In: *2018 37th Chinese Control Conference (CCC)*.

Xu, Peng, Xiatian Zhu, and David A Clifton (2022). "Multimodal learning with transformers: a survey". *arXiv preprint arXiv:2206.06488*.

Yam, R. et al. (2001). "Intelligent Predictive Decision Support System for Condition-Based Maintenance". *IJAMT*.

Yamaguchi, Atsuki, George Chrysostomou, Katerina Margatina, and Nikolaos Aletras (2021). "Frustratingly Simple Pretraining Alternatives to Masked Language Modeling". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Yang, Ling, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, and Yu Liu (2020). "Dpgn: Distribution propagation graph network for few-shot learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Yang, Bo-Suk et al. (2008). "Random forests classifier for machine fault diagnosis". *JMST*.

Yang, Zhe et al. (2021). "A multi-branch deep neural network model for failure prognostics based on multimodal data". *Journal of Manufacturing Systems*.

Yang, Zhilin, 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*.

Yang, Zichao, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola (2016). "Stacked attention networks for image question answering". *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Ye, Qinyuan, Bill Yuchen Lin, and Xiang Ren (2021). "Crossfit: A few-shot learning challenge for cross-task generalization in nlp". *arXiv preprint arXiv:2104.08835*.

Yeh, Ching-Feng, Yongqiang Wang, Yangyang Shi, Chunyang Wu, Frank Zhang, Julian Chan, and Michael L Seltzer (2021). "Streaming attention-based models with augmented memory for end-to-end speech recognition". In: *2021 IEEE Spoken Language Technology Workshop (SLT)*.

Yu, Kun et al. (2019). "A bearing fault and severity diagnostic technique using adaptive deep belief networks and Dempster–Shafer theory". *Structural Health Monitoring*.

Yu, Wenmeng, Hua Xu, Ziqi Yuan, and Jiele Wu (2021). "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis". In: *Proceedings of the AAAI conference on artificial intelligence*.

Yuan, Jiahong and Mark Y. Liberman (2008). "Speaker identification on the SCOTUS corpus". *Journal of the Acoustical Society of America*.

Zadeh, Amir, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency (2017). "Tensor Fusion Network for Multimodal Sentiment Analysis". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*.

Zadeh, Amir, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis Philippe Morency (2018). "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph". *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*.

Zadeh, Amir, Soujanya Poria, Paul Pu Liang, Erik Cambria, Navonil Mazumder, and Louis Philippe Morency (2018). "Memory fusion network for multi-view sequential learning". *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.

Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. (2020). "Big bird: Transformers for longer sequences". *Advances in neural information processing systems*.

Zarei, Jafar et al. (2014). "Vibration analysis for bearing fault detection and classification using an intelligent filter". *Mechatronics*.

Zhang, Qian, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar (2020). "Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss". *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Zhang, Shen et al. (2020). "Deep Learning Algorithms for Bearing Fault Diagnostics—A Comprehensive Review". *IEEE Access*.

Zhang, Yue and Joakim Nivre (2011). "Transition-based Dependency Parsing with Rich Non-local Features". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Zhang, Zhenyou et al. (2013). "Fault diagnosis and prognosis using wavelet packet decomposition, Fourier transform and artificial neural network". *Journal of Intelligent Manufacturing*.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. (2023). "A survey of large language models". *arXiv preprint arXiv:2303.18223*.

Zhao, Zihao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh (2021). "Calibrate Before Use: Improving Few-shot Performance of Language Models". In: *Proceedings of the 38th International Conference on Machine Learning*.

Zhou, Funa et al. (2018). "A Multimodal Feature Fusion-Based Deep Learning Method for Online Fault Diagnosis of Rotating Machinery". *Sensors*.

Zhu, Xiaojin Jerry (2005). "Semi-supervised learning literature survey".

Ziegler, Daniel M, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving (2019). "Fine-tuning language models from human preferences". *arXiv preprint arXiv:1909.08593*.

Ziko, Imtiaz, Jose Dolz, Eric Granger, and Ismail Ben Ayed (2020). "Laplacian regularized few-shot learning". In: *International conference on machine learning*.