Transformer Networks and Their Real-World Business Applications

GLIMS Journal of Management Review and Transformation
1–13

DOI: 10.1177/jmrt.221110955
mrt.spectrumjps.com

Divam Gupta[1], Param Gupta[2] and Medha Goyal[3]

Abstract

Recent advances in deep learning models have presented many opportunities for businesses. This article focuses on the possible game changing development of transformer networks which have enabled self-supervised learning. These advances provide encouraging opportunities for business applications. The article discusses the different types of learning paradigms and the similarities and dissimilarities between them. The article also discusses how transformer networks enable self-supervised learning. The article finally discusses real-life business applications with data from text, audio and images.

Keywords

Transformer networks, supervised learning, unsupervised learning, self-supervised learning, business applications

Introduction

Over the past few years, the predominant focus of artificial intelligence (AI) and machine learning (ML) research has been on technical and theoretical aspects. With the explosion in the volume of data, especially unstructured data, coupled with exponential increase in processing power, businesses are now aggressively looking for ways to gain competitive advantage by deploying these technologies

[1] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2] University of Florida, Gainesville, Florida, USA
[3] Monash University, Clayton, Victoria, Australia

Corresponding Author:
Medha Goyal, Monash University, Wellington Rd, Clayton VIC 3800, Australia.
E-mail: goyalmedha12@gmail.com

(Davenport & Ronanki, 2018) to develop new products, serve their customers, market research, process automation, etc. (Duan et al., 2019). Practitioners and researchers alike suggest that Artificial Intelligence and Machine Learning (AIML) will provide the next frontier for both productivity and competition (Dwivedi et al., 2021). Some have even gone to claim that this is a revolution that will fundamentally transform how the industry itself operates (Ågerfalk, 2020). Firms are now beginning to figure out how investment in AIML can generate business value. There is a surge in companies venturing in the AIML world (Ransbotham et al., 2018) and investing in integrating AIML into the various aspects of the business.

Large volumes of unstructured data in the form of text, audio and images are an important by-product of the digital age. Posts on Facebook and Twitter and blogs provide deep insights into consumer behaviour and presents unprecedented potential opportunities to companies as well as academic researchers.

Existing models for deep learning, such as Convolutional Neural Networks (CNN) and Recurring Neural Networks (RNN), demonstrate issues with modelling of long-term contexts as they learn within a specific context which leads to a locality bias (Qiu et al., 2020). Also, since RNNs process inputs sequentially, that is, one word after the next, it makes limited use of parallel processing. A deep learning model, referred to as transformers, proposed by Vaswani et al. (2017), was designed to overcome these challenges. The model utilises 'self-attention' in learning, leading to self-supervised learning by the models. In essence, the model architecture design in self-attention utilises more parallel learning as compared to RNNs. The self-attention model can also take into account the long-term contexts. This is because every permutation and combination of all meaningful words, referred to as tokens, from the input sequence are used for learning (Vaswani et al., 2017). The architecture of transformers allows for it to learn even from complex language information. At the same time, it is an expensive as well as time-consuming process for generating humongous volumes of labelled data in the natural language processing (NLP) domain from the easily available unlabelled data.

Recent students suggest that transformer-based pre-trained language models have the potential to overcome the learning challenges of existing models and can likely successfully deliver in a variety of NLP tasks. Generative pre-trained transformer (GPT) and bidierctional encoder representations from transformers (BERT) are frontrunners in the evolution of such self-supervised learning models which have utilised the transformer architecture. As a result, the transformer-based pre-trained language models can potentially utilise self-supervised learning to learn any universal representations from humongous volumes of data. Subsequently, such knowledge can be transferred to downstream tasks. Since the self-supervised learning models provide good background knowledge for downstream tasks, the need to train downstream models from scratch is eliminated. In fact, the downstream models can then be trained with minimal data which provides the right context for the model.

Deep Learning Models

The earlier NLP systems were mainly rule-based and were subsequently replaced by machine learning models. For machine learning models to deliver successfully,

domain expertise of the modeller is critical and is, thus, a time- and resource-consuming process. Prominent learning models in recent years are unsupervised (USL), supervised (SL) and self-supervised learning models (SSL).

Supervised learning enables models to learn from data labelled by humans. This supervised learning has been a critical part in the progress that AI has achieved in recent years. Such models work well on specific tasks as they are trained using data labelled for that specific task. A drawback of such supervised training is that to attain good model performance, a large number of labelled data is required. Collecting and labelling such data is an expensive and time-consuming task. Additionally, for many domains, such as medical and legal, it is difficult to access labelled data. Furthermore, since these models only use the supplied labelled data for training, they are likely to suffer from a bias as there may be many other relationships that exist in reality which have not been captured by the labelled data used for training, thus, leading to generalisation error, and the models may probably even accept spurious relationships.

In summary, the drawbacks of supervised learning are that it has significant dependence on data labelled by humans, which is usually time-intensive and cost-ineffective, and that it lacks the ability to generalise the models that emerge from such learning. This can, at times, also lead to spurious relationships. In many areas, such as legal or medical, availability of labelled data could be a challenge, thereby creating a limit on applications for such models in such domains. These models are unable to capitalise on learning for easily available unlabelled data.

As the computer hardware accelerators such as Graphic Processing Units (GPU) and Tensor Processing Units (TPU) along with word association algorithms such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) evolved, the use of deep learning models such as CNN (Kalchbrenner et al., 2014) and RNN (Zhou et al., 2016) for NLP applications increased significantly (Zhou et al., 2016). As stated earlier, the major challenges in training these models were that they had to be trained from scratch (with the exception of word embeddings) and required large volumes of the labelled data for training, and such labelled data is expensive to generate.

Self-Supervised Learning and Transformers: A Game Changer?

Self-Supervised Learning

Self-supervised learning (SSL), a newer learning paradigm, has attracted significant attention from the AI research community. Its biggest advantage is its ability to learn from unlabelled data for gaining universal knowledge about language, speech or images when fed into pre-trained models.

For SSL to work effectively, first the system is pre-trained with a humongous volume of data. Such pre-training can involve one or more pre-training tasks. Since the pretraining is achieved with extremely large volumes of data, the model learns the universal language relationships for both information: syntactic and semantic.

When performing downstream tasks with these generalised models, these representations enable improved performance as they now require only a few instances of label data. In simpler terms, with the pre-training with large volumes of unlabelled data, the model learns fundamental common sense and basic knowledge. Now for specific tasks, these models require minimal labelled data to train.

While supervised learning has been a revolutionary step in the development of deep learning models over the last couple of decades, its drawbacks have pushed the research agenda towards developing alternate paradigms for learning. SSL is one such new learning paradigm. The biggest advantage of SSL is that it requires any human labelled data, and rather it can learn from humongous volumes of unlabelled data without any human supervision. The architecture that enables such learning was proposed by Vaswani et al. (2017) and has gained significant traction since then. Vaswani et al. (2017) named this architecture 'Transformers'. Since SSL is both data efficient and has the ability to generalise, it finds a variety of applications in the deep learning fields of NLP (Liu et al., 2020; Qiu et al., 2020), robotics (Liu et al., 2020), computer vision (Han et al., 2020; Khan et al., 2021) and speech (Baevski et al., 2020; Sivaraman & Kim, 2020).

SSL can be viewed as a hybrid of both supervised and unsupervised learning and some points of similarity and dissimilarity with both. As in unsupervised learning, SSL does not require labelled data. The primary objective of USL is to identify hidden patterns in the data, in contrast, the value of SSL is to identify and model meaningful relationships. In comparison to SL, SSL is similar to SL as they both require supervision and are different as SSL is able to generate labels without human involvement and while the goal of SL is to model a specific objective, SSL's aim is to train the model with a very large set of general knowledge. Thus, SSL uses the universal knowledge as background because of which downstream learning is possible with minimal training. Since the SSL models are trained on humongous amounts of freely available unlabelled data, such learning is more generalizable as compared to USL and SL which are useful for the specific objectives they were built for. As a result of these features, models such as GPT-1, BERT (Devlin et al., 2018), XLNet (Yang et al., 2020), Roberta, T5, ELECTRA, BART, ALBERT and PEGAUSUS have demonstrated encouraging success with learning using NLP and transferring this knowledge (also referred to as transfer knowledge) to downstream learning and models. Transformer-based architecture, a more recent development for deep learning, has made such pre-training of models with these humongous amounts of data possible.

Transformers

For many business applications, we can utilise deep-learning models that are pre-trained. This eliminates the intensive task of developing and training complex deep learning models. Rather we can use a model that is proven to be optimised and work well. This also eliminates the task of collecting and labelling

large amounts of data. This can conserve resources that can be deferred to other core aspects of the business application.

Typically, when we want to utilise deep learning models that are pre-trained, we need to find a model that is trained for our specific task. Besides this, a pre-trained model can be fine-tuned for our specific task. This still requires us to collect and label considerable data. In few-shot learning, we only need to provide a few example cases as input along with our query to the model. With this, the same model can be used for a variety of tasks without requiring any extra engineering. Extremely large transformer models have shown great success in few-shot learning. Transformers are able to perform few-shot learning by having extremely large number of parameters and training data. Since transformers can utilise parallel computing or hardware acceleration, they can achieve a scale large enough for few-shot learning that previous sequential models like long short-term memory (LSTM) could never achieve.

GPT-3 was one of the first popular and mainstream success for its breakthrough application with few-shot learning. GPT-3 can achieve few-shot learning due to its 175 billion trainable parameters being trained on 45TB of unlabelled text data collected partly from a data set of eight years of web crawling.

GPT-3 can be used for many breakthrough applications and produce astonishing results. There have been many applications that were built using GPT-3. For example, one application was a Figma plugin. In this, by just giving the text description of your desired application, a prototype can be generated within a matter of seconds. If one gave the description of an Instagram-like application, then the plugin will generate an appropriate prototype that is comparable to Instagram. Another application was a ReactJS application generator. In this, by giving the text description of a web app, the application would generate ReactJS code for it. When the text description of a to-do app was given to it, a code for the functioning of the to-do app was generated.

Transformers can enable businesses to use pre-trained models like GPT-3 to build many such breakthrough applications to complement their existing or create new products/services.

Transformer Network Architecture

Transformer networks contain a sequence encoder which takes the input sequence and encodes it to a sequence of vectors. The encoder of the transformer network uses a self-attention module which helps it decide which parts of the input are important. It is used to ignore the unrelated tokens for the given token in the sequence and focus on the related tokens. For example, in the sentence ‘The car is red’, the model might assign high attention to the token ‘red’ for the token ‘car’. For every token in the input sequence, an attention score is computed with respect to every other token in the sequence. This creates an attention matrix of size $N \times N$, where N is the sequence length. This attention matrix could be a memory bottleneck if the sequence length is large. The encoder also uses positional embeddings which helps the network (Figure 1).

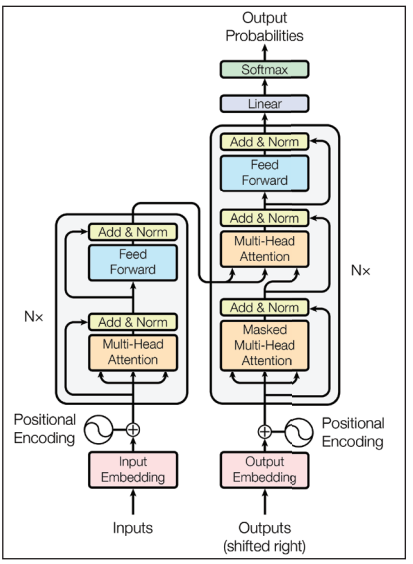

Figure 1. Architecture of the Transformer Network

Source: Vaswani et al. (2017).

Real-World Applications

Deep learning is widely used for several applications in various domains such as NLP, computer vision, video processing, audio processing, etc. There are several supervised and unsupervised deep learning techniques which are used in a wide variety of real-world applications. The supervised techniques require several labelled data points which are used to train a model. For example, to build an application to recognise objects present in an image, we would need several images with ground truth labels of the objects present. It is said that adding more training data usually gives us a higher prediction accuracy. That is not necessarily true, because the accuracy might start to saturate after a point. That is why it is beneficial to use models which are more scalable, such as the transformer network (Vaswani et al., 2017).

Transformer networks (Vaswani et al., 2017) were proposed for sequence-based tasks such as text processing. The authors proposed transformers for sequence-to-sequence applications and showed the efficacy of transformer networks on machine translation. The authors showed that transformer networks outperform other sequence processing techniques such as LSTM networks (Sundermeyer et al., 2012).

Transformer networks have been shown to be more scalable compared to LSTM networks, that is, they can generalise better on larger data sets. LSTM networks process each token of the sequence sequentially, whereas transformer networks process them in parallel; hence, they can better exploit hardware accelerators such as GPU or TPU as it allows parallel processing. Recent models like GPT-3 (Brown et al., 2020), which use transformer networks, are trained on several machines

connected through a network. For NLP applications, transformer networks can be first pre-trained for the task of next token prediction, where the network has to predict the next word in the sentence. This unsupervised pre-training is very useful because getting large amounts of labelled data is very expensive.

Unlike LSTM networks or CNNs, transformers do not make any prior assumptions about the input data, which helps them generalise better with large amounts of training data. Although models which do make assumptions about the structure of the data might perform better with smaller data sets.

Transformer Networks for Text-Based Applications

Transformer networks have been shown to be effective in several applications involving text data such as machine translation, question answering, named entity recognition, text summarisation, etc. In the application of machine translation, the model takes a sentence in language A and translates it to language B. This is usually done by training a sequence-to-sequence model on a parallel corpus having sentences of both language A and B. The application of the named entity recognition is to find named entities such as names of people, places, etc., in a sentence. This could be useful for analysing relationships between several entities from text data. Automatic question answering is very useful for chatbots, search engines and customer support. Here, the models take a query as input and select an appropriate response from a database of documents. The response is then re-worded to match the grammar according to the question. This uses a ranking model which ranks the responses by assigning similarity scores to all the responses from the database. Several models such as BERT (Devlin et al., 2018), ET (So et al., 2019) and transformer extra long (Transformer-XL) (Dai et al., 2019) have shown good performance for text-based applications.

Transformers can enable businesses to build highly accurate chatbots that are practically indiscernible from real human beings. They can also be used to summarise large texts accurately or make documents with legal jargon more comprehensible.

Transformer Networks for Computer Vision Applications

Transformer networks have shown good efficacy for image classification, object detection, semantic segmentation and image generation. Transformer networks can be used to replace the convolution operations in a convolutional neural network, or they can be used in conjunction with convolutional operations. The self-attention in transformer networks allows it to learn high-level information in the image. VIT (Dosovitskiy et al., 2020) splits the image into multiple patches and applies linear projection on each patch before passing it to the transformer module (Figure 2). The authors of VIT show that their method is superior to several CNN-based image classification models.

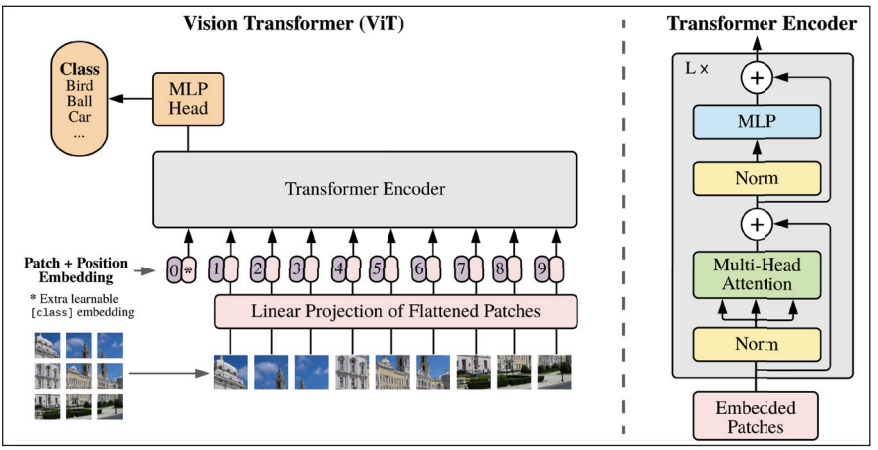

Figure 2. Architecture of the VIT Network

Source: Dosovitskiy et al. (2020).

DETR (Zhu et al., 2020) is a transformer-based end-to-end network to detect the classes and location of multiple objects in an image. VideoBERT (Sun et al., 2019) is a transformer-based model which learns a joint representation of text and videos, which can be used to generate captions from videos. The authors trained VideoBERT on data from YouTube. DALL-E (Ramesh et al., 2021) and CogView (Ding et al., 2021) are transformer-based models which are used to generate images by using a text description as input.

Using the idea of transfer learning, computer vision researchers trained large CNN models (He et al., 2016; Simonyan & Zisserman, 2014; Szegedy et al., 2016; Tan & Le, 2019) with the use of large data sets (e.g., ImageNet [Krizhevsky et al., 2012; Russakovsky et al., 2015]). Such models learn the common relationships from these images. Subsequently, these models pre-trained with large volumes of data are fine-tuned to downstream tasks with a very small data set specific to the task at hand (Kaur & Gandhi, 2019). Such models have demonstrated significant success in computer vision tasks (He et al., 2016; Ren et al., 2015).

Researchers from the area of NLP are optimistic with such pre-trained models and are increasingly combining the power of SSL and transformers in models such as GPT-3 (Brown et al., 2020), PANGU-(200B) (Zeng et al., 2021) and GShard (600B) (Lepikhin et al., 2020) which are trained using billions of parameters and use trillions of parameters for their switch-transformers (Fedus et al., 2021).

Transformer Networks for Audio Applications

Transformer networks have been used to encode audio signals for solving audio-based applications. Audio speech transformer (Dong et al., 2018) splits the audio

signals into chunks and applies a linear projection on each chunk to get a sequence of embeddings. The sequence is passed to a transformer network, which is then used to perform classification of audio clips. Streaming transformer (Moritz et al., 2020) is another transformer-based model which takes a chunk-wise stream of audio and performs real-time speech recognition. Music transformer (Huang et al., 2018) is a generative network which is used to generate music with long-term structure.

Audio chatbots can be built using transformers, which will be able to have normal human such as interactions and conversations. These bots can be used in place of human customer service agents and continue to provide comparable service.

Discussion and Conclusion

Supervised- and unsupervised-based learning models have found increasing use in the real-world applications. Such models have been found to be prone to and amplifying the biases that exist in the data sets used for training. Therefore, the decisions based on these models may end up being unintentionally biased. By using transformer-based pre-trained learning models, many of these biases are expected to reduce. A limitation of such transformer-based pre-trained models is that they are expensive to train. But with increasing processing power of computers, these costs are likely to come down.

Subsequent to the success with the transformer-based pre-trained learning models in the general domain of English, such architecture and learning is also being used in various other domains such as legal (Articles 32 and 33), programming (Ahmad et al., 2021; Feng et al., 2020; Guo et al., 2020; Lu et al., 2021; Phan et al., 2021), finance (Yang et al., 2020), news (Gururangan et al., 2020), networking (Louis, 2020), biomedical (Alsentzer et al., 2019; Gu et al., 2021; Lee et al., 2020; Peng et al., 2019, 2021), dialogue (Wu et al., 2020) and academics (Beltagy et al., 2019; Liu et al., 2021; Peng et al., 2021).

This article discusses several learning methods using transformer networks and provides their real-world business applications. Transformer networks show great potential due to their efficacy and scalability. While this article presents several transformer-based applications in the domains, such as text processing, image processing and audio processing, it has shown promising results in multiple domains such as legal, academics, finance and news. Further research in business applications of these models will help the business community adopt these for meaningful decision-making.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

References

Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems, 29*(1), 1–8.

Ahmad, W. U., Chakraborty, S., Ray, B., & Chang, K. W. (2021). Unified pre-training for program understanding and generation. arXiv preprint arXiv:2103.06333.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860.

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review, 96*(1), 108–116.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., & Tang, J. (2021). Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems, 34*, 19822–19835.

Dong, L., Xu, S., & Xu, B. (2018, April). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5884–5888). IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data: Evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63–71.

Dwivedi, Y. K., Ismagilova, E., Rana, N. P., & Raman, R. (2021). Social media adoption, usage and impact in business-to-business (B2B) context: A state-of-the-art literature review. *Information Systems Frontiers*, 1–23. https://doi.org/10.1007/s10796-021-10106-y

Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint arXiv:2101.03961.

Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., & Zhou, M. (2020). CodeBERT: A pre-trained model for programming and natural languages. arXiv preprint arXiv:2002.08155.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH), 3*(1), 1–23.

Guo, D., Ren, S., Lu, S., Feng, Z., Tang, D., Liu, S., Zhou, L., Duan, N., Svyatkovskiy, A., Fu, S., Tufano, M., Deng, S. K., Clement, C., Drain, D., Sundaresan, N., Yin, J., Jiang,

D., & Zhou, M. (2020). Graphcodebert: Pre-training code representations with data flow. arXiv preprint arXiv:2009.08366

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2020). A survey on visual transformer. arXiv preprint arXiv:2012.12556.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music transformer. arXiv preprint arXiv:1809.04281.

Kalchbrenner N., Grefenstette E., & Blunsom P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers, pp. 655–665). Association for Computational Linguistics.

Kaur, T., & Gandhi, T. K. (2019, December). Automated brain image classification based on VGG-16 and transfer learning. In *2019 International Conference on Information Technology (ICIT)* (pp. 94–98). IEEE.

Khan S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in vision: A survey. arXiv preprint arXiv:2101.01169.

Krizhevsky A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., & Chen, Z. (2020). Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668.

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A survey on contextual embeddings. arXiv preprint arXiv:2003.07278.

Liu X., Yin, D., Zhang, X., Su, K., Wu, K., Yang, H., & Tang, J. (2021). Oag-Bert: Pre-train heterogeneous entity-augmented academic language models. arXiv preprint arXiv:2103.02410.

Louis, A. (2020). NetBERT: A pre-trained language representation model for computer networking [Doctoral dissertation, Cisco Systems].

Lu, S., Guo, D., Ren, S., Huang, J., Svyatkovskiy, A., Blanco, A., Clement, C., Drain, D., Jiang, D., Tang, D., Li, G., Zhou, L., Shou, L., Zhou, L., Tufano, M., Gong, M., Zhou, M., Duan, N., Sundaresan, N., … & Liu, S. (2021). Codexglue: A machine learning benchmark dataset for code understanding and generation. arXiv preprint arXiv:2102.04664.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Moritz, N., Hori, T., & Le, J. (2020, May). Streaming automatic speech recognition with the transformer model. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6074–6078). IEEE.

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474.

Peng, S., Yuan, K., Gao, L., & Tang, Z. (2021). Mathbert: A pre-trained model for mathematical formula understanding. arXiv preprint arXiv:2105.00377.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics.

Phan, L., Tran, H., Le, D., Nguyen, H., Anibal, J., Peltekian, A., & Ye, Y. (2021). Cotext: Multi-task learning with code-text transformer. arXiv preprint arXiv:2105.08645.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences, 63*(10), 1872–1897.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, C., & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821–8831). PMLR.

Ransbotham, S., Gerbert, P., Reeves, M., Kiron, D., & Spira, M. (2018). Artificial intelligence in business gets real. *MIT Sloan Management Review.* Accessed, 24 April 2022, from artificial-intelligence-in-business-gets-real

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett (Eds), *Advances in neural information processing systems 28* (pp. 91–99). Curran Associates, Inc.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sivaraman, A., & Kim, M. (2020). Self-supervised learning from contrastive mixtures for personalized speech enhancement. arXiv preprint arXiv:2011.03426.

So, D., Le, Q., & Liang, C. (2019, May). The evolved transformer. In *International Conference on Machine Learning* (pp. 5877–5886). PMLR.

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7464–7473). IEEE/CVF.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association (ISCA)*. International Speech Communication Association (ISCA).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826). IEEE.

Tan, M., & Le, Q. (2019, May). EfficientNet: Rethinking model scaling for Convolutional Neural Networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems 30* (pp. 5998–6008).

Wu, C. S., Hoi, S., Socher, R., & Xiong, C. (2020). TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. arXiv preprint arXiv:2004.06871.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.

Zeng, W., Ren, X., Su, T., Wang, H., Liao, Y., Wang, Z., Jiang, X., Yang, Z., Wang, K., Zhang, X., Li, C., Gong, Z., Yao, Y., Huang, X., Wang, J., Yu, J., Guo, Q., Yu, Y., Zhang, Y., ... & Tian, Y. (2021). PanGu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. arXiv preprint arXiv:2104.12369.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. arXiv preprint arXiv:1611.06639.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.