Symbiosis

www.symbiosisonlinepublishing.com
ISSN Online: 2474-9257

Review Article | Journal of Computer Science Applications and Information Technology | Open Access

Beyond Text: Exploring Multimodal BERT Models

Sarika Kondra[1*], Vijay kumar Adari[2] and Vijay Raghavan[1]

[1]*University of Louisiana, Lafayette, Louisiana, USA*

[2]*Cognizant Technologies Solutions, USA*

Received: January 05, 2025; Accepted: January 17, 2025; Published: February 03, 2025

***Corresponding author:** Sarika Kondra, University of Louisiana, Lafayette, Louisiana, USA, Tel: +1 337- 806-6587; E-mail: venkatasarika.kondra@gmail.com*

Abstract

This paper explores the burgeoning potential of Bidirectional Encoder Representations from Transformers (BERT) for various multimodal tasks. BERT's ability to capture contextual relationships in text empowers it to create richer representations for data beyond just language.

The paper explores how BERT can be integrated with visual and auditory information for applications in video, image analysis and audio processing. Different approaches for this integration are discussed, including early fusion and multimodal transformers, where BERT collaborates with models specialized in other modalities to achieve a deeper understanding of the content.

BERT's capabilities extend to audio data as well. It can be employed for tasks like speech recognition, where it can improve word prediction accuracy by leveraging its understanding of language context. Additionally, BERT holds promise for sentiment analysis in audio, enabling the analysis of emotional tones and speaker intent.

Furthermore, the combination of BERT with Graph Neural Networks (GNNs) presents promising results for tasks involving relational data and text, as seen in recent work with Graph-BERT.

The paper also highlights the potential of multilingual BERT models (mBERT) for tasks in multiple languages. The versatility of mBERT and other multilingual models allows for the exploration of tasks beyond individual languages.

In conclusion, BERT's versatility in multimodal tasks opens new possibilities for data interaction and understanding.

Keywords:

Bidirectional Encoder Representations from Transformers; Multimodal; Multilingual; Graph Neural Networks; Natural Language Processing

Abbreviations:

BERT: Bidirectional Encoder Representations from Transformers; NLP: Natural Language Processing; GNN: Graph Neural Networks; ASR: Automatic Speech Recognition

Introduction

The Bidirectional Encoder Representations from Transformers (BERT) [15] model has become a cornerstone of natural language processing (NLP), since its inception. Its ability to capture contextual relationships between words has propelled advancements in tasks like sentiment analysis, question answering, and text summarization. However, the world around us is not confined to text alone. Images, videos, audio, and other sensory data often coexist, offering a wealth of contextual information that can significantly enrich our understanding. This paper explores the rapidly growing domain of mixed models, which combine the strengths of BERT with various modalities to create powerful composite models.

This paper is structured into five BERT-based Multimodal Applications:

• **BERT-based Video Analysis Techniques**: This subsection explores how BERT can be combined with visual features extracted from videos to enhance tasks like action recognition, video summarization, and caption generation. Different fusion approaches (early, late, and multimodal transformers) are discussed.

• **BERT-based Image Analysis Techniques:** We will investigate how BERT can be integrated with image data for tasks like visual question answering, image captioning, and image retrieval.

• **BERT with Audio Data**: This subsection delves into the potential applications of BERT for processing audio data. We discuss tasks like audio classification, automatic speech recognition, audio event detection, audio captioning/summarization, speaker identification, and sentiment analysis from audio.

• **BERT with Graph Neural Networks**: While BERT is primarily designed for sequential data, this subsection explores how it can be combined with Graph Neural Networks (GNNs) to improve performance on tasks involving textual and relational data.

• **Multilingual BERT models**: We explore the capabilities of multilingual BERT models like mBERT [15] and XLM-RoBERTa [24], which enable tasks in multiple languages without language-

specific pre-training.

This exploration of BERT's versatility in handling multimodal data highlights its potential to bridge the gap between different information modalities, leading to a deeper understanding of the world around us.

- **BERT**: Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) [15] has revolutionized the field of Natural Language Processing (NLP) since its introduction in 2018. BERT's core strength lies in its pre-training on massive text corpora using masked language modeling and next sentence prediction tasks. This pre-training allows the model to capture deep contextual relationships between words, enabling it to excel in various NLP tasks. BERT's effectiveness extends to a broad landscape of NLP tasks, including: text classification, sentiment analysis, topic modeling, spam detection, question answering, text summarization, natural language generation, machine translation, information retrieval and many more.

BERT leverages the Transformer architecture [28], a powerful neural network structure known for its ability to capture long-range dependencies within text data. Unlike traditional sequential models that process text word-by-word, BERT trains using a masked language modeling (MLM) objective. In MLM, some words in the input sequence are masked, and the model is tasked with predicting the masked words based on the surrounding context. This bidirectional training allows BERT to learn deep contextual representations of words, capturing not just their individual meaning but also their relationships with other words in the sentence.

Core Strengths of BERT lies in,

- **Pre-trained on Massive Text Corpora**: BERT is pre-trained on a colossal dataset of text and code, allowing it to acquire a rich understanding of general language patterns. This pre-training provides a strong foundation for various downstream NLP tasks.

- **Contextual Word Representation:** BERT excels at understanding the meaning of a word based on its surrounding context. This is crucial for tasks like sentiment analysis, where the sentiment of a word can vary depending on the context.

- **Versatility through Fine-tuning**: BERT's pre-trained weights can be fine-tuned for a wide range of NLP tasks. This flexibility makes it a valuable tool for researchers and developers working on various NLP applications.

These strengths and BERT's versatality can be leveraged by multimodal, non-textual data in the development of efficient applications. The subsequent sections will explore how BERT can be integrated with other powerful techniques to unlock even more sophisticated mixed models.

- **BERT-based Video Analysis Techniques**

BERT [15] helps extract richer representations of videos by considering not just the visual elements, but also the context provided by the captions. This empowers tasks like action recognition and video summarization. By leveraging the power of BERT's pre-trained knowledge on language, models can generate more accurate and descriptive captions that better reflect the content of the video. Common BERT-based approaches for video analysis and captioning can be categorized under early fusion, late fusion and multimodal transformers.

Early Fusion: This approach combines visual features extracted from video frames (e.g., using pre-trained convolutional neural networks) with textual information (e.g., video title or transcript) and feeds them into a single BERT model. The model then learns a joint representation that captures the interplay between the visual and textual domains. VideoBERT [1] model utilizes this approach, combining Automatic Speech Recognition (ASR) outputs, vector-quantized visual features, and BERT for joint representation learning. It achieves state-of-the-art results in action classification and video captioning. Vector quantization of VideoBERT [1] loses the fine-grained information of the video frames and hence researchers addressed this by replacing softmax loss with noise contrastive estimation (NCE) [27] and uses BERT [15] model for text sequences.

Late Fusion: The phenomenon of Late Fusion involves the independent processing of visual and textual features by separate models. The resulting outputs are then fused together using mechanisms such as attention to prioritize the most pertinent information from each modality. VL-BERT [2] introduces a unified architecture based on transformers [28]. Visual and linguistic content are fed directly into VL-BERT [2], allowing for earlier and more seamless interaction between the modalities. This model demonstrates strong performance on tasks like visual question answering and image captioning. Few other unified single-stream architectures are VisualBert [8], B2T2 [9], and Unicoder-VL [10].

Multimodal Transformers: This approach leverages transformer architectures specifically designed for handling multimodal data. These models inherently learn the relationships between visual and textual information throughout their training process. M-BERT [30] (Multimodal-BERT) uses the pre-trained BERT [15] as baseline model, and injects audio-visual information to fine-tuning in presence of nonverbal behaviors. This is done by gated-shifting of the input embedding of the BERT [15] model using word-level representations of nonverbal behaviors [11]. M-BERT [29] sets a new SOTA results CMU-MOSI dataset [30] of multimodal sentiment analysis.

Training and running BERT-based video analysis models can be resource-intensive and computationally expensive. Researchers are exploring techniques such as model compression and knowledge distillation to make BERT-based video analysis models more efficient and cost-effective. By reducing the size of the model through methods like eliminating redundant parameters and using low-rank factorization [32] or transferring

knowledge from a larger model to a smaller one [31], researchers are working towards making these models more accessible for resource-intensive tasks.

The other known challenge with training video analysis is the availability of large, high-quality video datasets with well-annotated captions for training effective models. Continued efforts are needed in data collection and annotation.

Despite these challenges, the integration of BERT with video analysis and captioning holds immense promise. As research progresses, we can expect even more sophisticated models capable of generating detailed, informative, and engaging video descriptions.

3. BERT-based Image Analysis Techniques

In recent years, BERT [15] has also shown promising results in image analysis tasks, such as image classification, object detection, and semantic segmentation. This intersection between BERT and image analysis has opened new possibilities for image understanding and has gained tremendous attention from both the computer vision and natural language processing communities.

ViLBERT [4] consists of two parallel BERT-style models operating over image regions and text segments to learn joint representations of language and the corresponding visual content. ViLBERT [4] reported SOTA results in visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval tasks.

Lxmert [5] (Learning Cross-Modality Encoder Representations from Transformers) framework consists of three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder. The model is pre-trained with large amounts of image-and-sentence pairs to learn both intra-modality and inter-modality relationships. This model achieves SOTA results on two visual question answering datasets VQA[6] and GQA[7].

In ViL BERT [4] and LXMERT [5], the network architectures are of two separate single-modal networks applied on input sentences and images respectively, followed by a cross-modal Transformer combining information from the two sources.

Image BERT [12] is based on the BERT architecture and is trained on large-scale image-text data. It consists of a stack of transformer layers that are trained by predicting the masked tokens in the input text and image data. The pre-training process of ImageBERT [12] allows it to learn rich image-text representations that capture the complex relationships between visual and textual data. This pre-trained ImageBERT [12] model can then be fine-tuned for downstream tasks without the need for extensive data or resource requirements.

The ability of BERT to understand the contextual relationships between words and objects within an image has greatly enhanced the accuracy and efficiency of image analysis. Its ability to incorporate both text and image data has also led to more comprehensive and holistic understanding of images. Beyond traditional computer vision methods, BERT-based image analysis techniques have also shown potential in fields such as medical imaging [33], where the combination of text and image features can aid in more accurate diagnoses and treatment. However, there is still room for improvement and further research is needed to fully explore the capabilities of BERT in image analysis. Some challenges, such as ensuring the interpretability of the model and handling large-scale image datasets, will need to be addressed.

4. BERT with Audio data

While research in this area is less extensive compared to image analysis, there are initial explorations into using BERT for audio data processing. Some potential applications include audio classification, automatic speech recognition, audio event detection, audio captioning and summarization, speaker identification, sentiment analysis from audio and multimodal tasks with audio and text.

Audio Classification: BERT [15] can be used to analyze the content of audio data (e.g., music genre classification, speech act recognition) by incorporating additional pre-processing steps to convert audio features into a suitable format for BERT. Like video analysis, audio features extracted from spectrograms or mel-frequency cepstral coefficients (MFCCs) can be combined with textual information (e.g., captions, transcripts) and fed into a BERT model. This allows the model to learn the relationships between audio characteristics and the corresponding meaning. In a study by Korzack and Weller [34], BERT was used to classify speech acts in political speeches, achieving an accuracy of 81%. The authors of [35] propose a lightweight version self-supervised speech representation model called Audio ALBERT. They demonstrate the effectiveness of this model in achieving comparable performance with powerful pre-trained networks while using 91% fewer parameters. The lightweight model is applied to two downstream tasks, speaker classification and phoneme classification, and shows promising results.

Automatic Speech Recognition (ASR): Standard ASR systems convert spoken language into text. BERT's [15] strength in understanding language context can significantly enhance ASR performance. Huang et al. (2021) [36] involved fine-tuning pre-trained BERT models on large speech datasets. The authors of [37] aimed to improve performance in automatic speech recognition by adapting BERT for the N-best list rescoring task without the need for fine-tuning. This was achieved by considering only the masked language modeling within a single sentence.

Audio Event Detection: Identifying and classifying events within audio recordings holds applications in areas like video surveillance or automatic content creation. Miyazaki et al. (2020) [38] introduces a new sound event detection (SED) approach that combines self-attention from the Transformer with a tag token for weak label prediction. This method outperforms the baseline method and effectively incorporates both local and global context

information.

Audio Captioning/Summarization: Similar to image captioning, BERT [15] could be used to generate textual descriptions or summaries of audio content. Liu, Xubo, et al [19] proposes a technique for generating captions from audio recordings using a pre-trained BERT model.

Speaker Identification and Diarization: Distinguishing between speakers in a recording and segmenting the audio based on who is speaking are crucial tasks in tasks like meeting summarization or forensics. While not directly applying BERT itself, some research leverages BERT's pre-training for speaker identification. MPC-BERT [40] is a pre-trained model designed to improve multi-party conversation understanding by considering the complex structure and semantics of conversations. It incorporates various self-supervised tasks specifically for modeling interlocutor structure and utterance semantics and has shown superior performance on addressee recognition, speaker identification, and response selection tasks compared to previous methods on two benchmark datasets. BERTphone [39] is a Transformer encoder that uses two objectives, acoustic and phonetic representation learning, to create phonetically aware contextual representation vectors for speaker and language recognition. Pretrained BERTphone models are used as feature extractors in DNNs, achieving a state-of-the-art result on language and speaker recognition tasks.

Multimodal Sentiment Analysis from Audio and Text: By combining BERT with audio processing techniques, researchers aim to analyze the emotional tone and sentiment conveyed in spoken language. Extracting emotional cues from speech recordings is valuable for applications ranging from customer service evaluations to mental health analysis. Sentiment analysis in speech can be more complex than text due to factors like tone of voice, sarcasm, and cultural nuances. CM-BERT [17] uses a combination of text and audio modality to fine-tune the pre-trained BERT. The model includes a masked multimodal attention component that dynamically adjusts the weight of words using information from both modalities. MF-BERT [18] introduces a multimodal fusion BERT model that incorporates nonverbal information to improve sentiment analysis. The model also incorporates an internal updating mechanism with two different optimizers to prevent overfitting and achieve better results compared to previous methods. Experiments on public datasets demonstrate the superior performance of this model.

Multimodal Tasks with Audio and Video: The power of BERT can be extended to multimodal tasks where both audio and visual information contribute to understanding the content. M-BERT [29] (Multimodal-BERT) uses the pre-trained BERT as baseline model, and injects audio-visual information to fine-tuning in presence of nonverbal behaviors. This is done by gated-shifting of the input embedding of the BERT model using word-level representations of nonverbal behaviors [12]. M-BERT [29] sets a new SOTA results CMU-MOSI dataset [30] of multimodal sentiment analysis.

Extracting meaningful representations from audio data for BERT processing remains an active area of research. Additionally, limited datasets of labeled audio data can hinder model development. Martin Morato et al. [20] discusses diversity and bias in audio datasets.

5. BERT with Graph Neural Networks

While BERT excels at processing sequential data like text, graph neural networks (GNNs) have emerged as powerful tools for handling relational data represented as graphs. Combining the strengths of both approaches holds immense potential for tasks involving textual and structural information. This section explores recent advancements in integrating BERT with GNNs, highlighting their synergistic capabilities.

G-BERT [13] combines Graph neural networks (GNN) and BERT [15] for medical/ diagnostic code representation and recommendation. Medical codes internal hierarchical structures are encoded using GNNs and the representations are fed as input during pre-training phase of BERT along with single-visit Electronic Health Records (HER) data. It is then finetuned for downstream medical prediction tasks.

Yuxuan Liang et al. (2022) [21] explores using BERT to enhance text graph neural networks for classification tasks. They propose a framework where BERT pre-trains node representations based on textual information associated with nodes in the graph. These pre-trained representations are then fed into a GNN for classification. Their approach demonstrates improvements on benchmark datasets involving text-rich graphs.

Jiawei Zhang et al. (2020) [14] proposes Graph-BERT, a novel GNN architecture that solely relies on attention mechanisms for learning graph representations. It departs from traditional GNNs that use graph convolutions and aggregation operators. The model demonstrates effectiveness on graph classification and clustering tasks, achieving competitive results compared to existing GNNs.

While pre-training BERT for GNNs shows promise, researchers see even greater potential in jointly training both models together. Additionally, designing specialized architectures for specific tasks involving text and structure could lead to significant improvements. Finally, overcoming limitations with handling massive graphs through efficient message passing and scalable architectures is crucial for real-world applications. By tackling these challenges, BERT and GNNs have the potential to become game-changers in dealing with complex data that combines text and relationships.

6. Multilingual BERT models

Multilingual BERT (mBERT) [15], also BERT [15], revolutionized the field of NLP by enabling tasks in multiple languages without the need for language-specific pre-training. It is pre-trained on a massive dataset of text in 104 languages. The authors demonstrate that mBERT achieves state-of-the-art performance on zero-shot cross-lingual transfer tasks, where a model trained in one language can be applied to another language without any

fine-tuning.

Pires et al. (2019) [23] investigated the true extent of mBERT's [15] multilingual capabilities. They show that mBERT [15] performs well on zero-shot tasks, even for languages not explicitly included in its pre-training data. This highlights the model's ability to learn generalizable linguistic representations.

Baptista et al. (2021) [25] used mBERT [15] to explore universal speech tagging, the task of assigning grammatical tags to words in spoken language. Their work demonstrates that mBERT[15] can be effectively used for this task across multiple languages, even with limited training data.

Cross-Lingual BERT Transformation (CLBT)[22] is an offline approach developed to generate cross-lingual contextualized word embeddings based on publicly available

pre-trained BERT [15] models performing zero-shot cross-lingual transfer parsing. It takes pretrained monolingual word embeddings of different languages as input and projects them into a shared semantic space. It showed outperforming results over the static word embeddings.

Conneau et al. (2020) [24] proposed XLM-RoBERTa, a multilingual model based on the RoBERTa [41] architecture, known for its improved performance over BERT. XLM-RoBERTa [24] is pre-trained on a massive dataset of 100 languages and demonstrates strong performance on various multilingual NLP tasks.

Though, multilingual BERT [15] showed success, researchers created other world language BERT models. Wang et al. [42] propose a way to extend M-BERT (E-BERT) to include any new language, resulting in improved performance in Named Entity Recognition (NER) tasks for both languages already in M-BERT and new languages. They conducted experiments on 27 languages, with an average increase of approximately 6% F1 for languages already in M-BERT and 23% for new languages. Virtanen et al. [26] finetuned BERT for Finnish language that outperformed multilingual BERT [15]. The authors of [43] show that monolingual models perform better than massively multilingual models. In this study, two trilingual BERT-like models were trained for Finnish, Estonian, and English, and Croatian, Slovenian, and English. These models, named FinEst BERT and CroSloEngual BERT, outperformed multilingual BERT [15] on all tasks in both monolingual and cross-lingual settings. Multilingual Representations for Indian Languages (MuRIL) [44] trained solely on Indian text corpora, outperforms the current state-of-the-art model on cross-lingual tasks and is also effective at handling transliterated data.

Conclusion

BERT's versatility extends beyond traditional text-based NLP tasks. By incorporating BERT into multimodal learning frameworks, researchers are achieving significant advancements in areas like video and image analysis, audio processing, and tasks involving graphs and relational data.

The ability to combine visual, auditory, and textual information empowers models to gain richer and more nuanced understandings of the content they process. This translates to improved performance in tasks like video captioning, image retrieval, audio sentiment analysis, and speaker identification.

While challenges remain in areas like efficient model training, data availability, and handling complex data modalities like audio, the potential of BERT-based multimodal applications is vast. As research continues to explore new integration techniques and address existing limitations, we can expect even more powerful models that bridge the gap between different data types, leading to a deeper and more comprehensive understanding of the world around us.

The multimodal research works we highlighted in this paper are not fully exhaustive list in the field of BERT and non-textual objectives. This field is constantly evolving. As researchers delve deeper, we can expect even more innovative applications that unlock the potential of BERT for multimodal data analysis.

Conflict of interest

There are no conflicts of interest.

References

1. Sun, Chen, et al. "Videobert: A joint model for video and language representation learning." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
2. Su, Weijie, et al. "Vl-bert: Pre-training of generic visual-linguistic representations." arXiv preprint arXiv:1908.08530 (2019).
3. Shen, T., et al. "Bert-based denoising and reconstructing data of distant supervision for relation extraction." CCKS2019-shared task (2019).
4. Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in neural information processing systems 32 (2019).
5. Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." arXiv preprint arXiv:1908.07490 (2019).
6. Goyal, Yash, et al. "Making the v in vqa matter: Elevating the role of image understanding in visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
7. Hudson, Drew A., and Christopher D. Manning. "Gqa: A new dataset for real-world visual reasoning and compositional question answering." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
8. Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." arXiv preprint arXiv:1908.03557 (2019).
9. Alberti, Chris, et al. "Fusion of detected objects in text for visual question answering." arXiv preprint arXiv:1908.05054 (2019).
10. Li, Gen, et al. "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.
11. Wang, Yansen, et al. "Words can shift: Dynamically adjusting word

representations using nonverbal behaviors." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

12. Qi, Di, et al. "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data." arXiv preprint arXiv:2001.07966 (2020).
13. Shang, Junyuan, et al. "Pre-training of graph augmented transformers for medication recommendation." arXiv preprint arXiv:1906.00346 (2019).
14. Zhang, Jiawei, et al. "Graph-bert: Only attention is needed for learning graph representations." arXiv preprint arXiv:2001.05140 (2020).
15. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
16. Koroteev, M. V. "BERT: a review of applications in natural language processing and understanding." arXiv preprint arXiv:2103.11943 (2021).
17. Yang, Kaicheng, Hua Xu, and Kai Gao. "Cm-bert: Cross-modal bert for text-audio sentiment analysis." Proceedings of the 28th ACM international conference on multimedia. 2020.
18. He, Jiaxuan, and Haifeng Hu. "MF-BERT: Multimodal fusion in pre-trained BERT for sentiment analysis." IEEE Signal Processing Letters 29 (2021): 454-458.
19. Liu, Xubo, et al. "Leveraging pre-trained bert for audio captioning." 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022.
20. Martin Morato, Irene, and Annamaria Mesaros. "Diversity and bias in audio captioning datasets." (2021).
21. Yuxuan Liang et al., "Bert-Enhanced Text Graph Neural Network for Classification," MDPI Applied Sciences 12.7 (2022): 3322.
22. Wang, Yuxuan, et al. "Cross-lingual BERT transformation for zero-shot dependency parsing." arXiv preprint arXiv:1909.06775 (2019).
23. Pires, Telmo, Eva Schlinger, and Dan Garrette. "How multilingual is multilingual BERT?." arXiv preprint arXiv:1906.01502 (2019).
24. Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).
25. Baptista, R., et al. (2021). Universal Speech Tagging with Multilingual BERT. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 1563-1573). Association for Computational Linguistics.
26. Virtanen, Antti, et al. "Multilingual is not enough: BERT for Finnish." arXiv preprint arXiv:1912.07076 (2019).
27. Gutmann, Michael, and Aapo Hyvärinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models." Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010.
28. Ashish, Vaswani. "Attention is all you need." Advances in neural information processing systems 30 (2017): I.
29. Rahman, Wasifur, et al. "M-bert: Injecting multimodal information in the bert structure." arXiv preprint arXiv:1908.05787 (2019).
30. Zadeh, Amir, et al. "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages." IEEE Intelligent Systems 31.6 (2016): 82-88.
31. Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).
32. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." science 313.5786 (2006): 504-507.
33. Linna, Nathaniel, and Charles E. Kahn Jr. "Applications of natural language processing in radiology: A systematic review." International Journal of Medical Informatics 163 (2022): 104779.
34. Korzack, M.S. and Weller, A., 2020. Political speech act classification using pre-trained BERT. In Proceedings of the 23rd International Conference on Speech and Computer (pp. 516-527). Springer, Berlin, Heidelberg.
35. Chi, Po-Han, et al. "Audio albert: A lite bert for self-supervised learning of audio representation." 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021.
36. Huang, Wen-Chin, et al. "Speech recognition by simply fine-tuning BERT." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.
37. Shin, Joonbo, Yoonhyung Lee, and Kyomin Jung. "Effective sentence scoring method using bert for speech recognition." Asian Conference on Machine Learning. PMLR, 2019.
38. Miyazaki, Koichi, et al. "Weakly-supervised sound event detection with self-attention." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
39. Ling, Shaoshi, et al. "Bertphone: Phonetically-aware encoder representations for utterance-level speaker and language recognition." arXiv preprint arXiv:1907.00457 (2019).
40. Gu, Jia-Chen, et al. "MPC-BERT: A pre-trained language model for multi-party conversation understanding." arXiv preprint arXiv:2106.01541 (2021).
41. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
42. Wang, Zihan, Stephen Mayhew, and Dan Roth. "Extending multilingual BERT to low-resource languages." arXiv preprint arXiv:2004.13640 (2020).
43. Ulčar, Matej, and Marko Robnik-Šikonja. "FinEst BERT and CroSloEngual BERT: less is more in multilingual models." Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23. Springer International Publishing, 2020.
44. Khanuja, Simran, et al. "Muril: Multilingual representations for indian languages." arXiv preprint arXiv:2103.10730 (2021).