Transformers and Applications
Attention is all you need!

Prof. Dr. Do Phuc
University of Information Technology
Vietnam National University Ho Chi Minh City
Email: phucdo@uit.edu.vn
Web: Tri Nhan Data Science & Application R&D Group
https://trinhansg.github.io/
July 2023

Short Bio

- Prof. Dr. Do Phuc is currently a member of the University of Information Technology, Vietnam National University - Ho Chi Minh City.
- Prof. Dr. Do Phuc has published over 100 scientific papers, more than10 books of Computer Science and Applications. He has been the principal investigator for many projects related to data mining, text mining, computational biology, social network analysis, big data processing and knowledge graph based QA system.
- Prof. Dr. Do Phuc served as the Vice Rector of the University of Information Technology and the Director of the International Relations Department of the Vietnam National University - Ho Chi Minh City from 2011 to 2013.

Outline

- Introduction
- Applications of the transformer in NLP
- Applications of the transformer in CV
- Applications of the transformer in Speech
- Challenges and Future research
- References

Introduction

- Transformers are a famous and powerful neural network architecture.
- Introduced by Vaswani et al. in 2017.
- Quickly became a core technology in many modern AI applications.

Characteristics of Transformers

- Utilize self-attention mechanism to understand the relationships between words/objects.
- Ability to process information in parallel across different parts of the input data.
- Easily scalable to handle complex language tasks.

Applications of Transformers in Natural Language Processing (NLP)

Introduction of transformers and NLP Application

- Natural Language Processing (NLP) involves understanding and generating human language using computational methods.
- Transformers have revolutionized NLP by achieving state-of-the-art performance in various tasks.

Applications of Transformers in Natural Language Processing (NLP)

- Pre-trained language models like BERT, GPT have made significant advancements in NLP.
- Used for tasks such as machine translation, text summarization, text classification, and question answering.

Machine Translation

- Transformers have greatly advanced machine translation systems.
- Examples include Google's Neural Machine Translation (GNMT) and Facebook's Fairseq.
- Transformers capture long-range dependencies and improve translation quality.

Language Modeling

- Transformers excel in language modeling tasks.
- Models like GPT (Generative Pre-trained Transformer) generate coherent and contextually relevant text.
- Language models trained on vast amounts of data achieve impressive performance.

Text Summarization

- Transformers have been applied to extractive and abstractive summarization tasks.
- Extractive summarization involves selecting important sentences from the source text, while abstractive summarization generates new sentences.
- Models like "BART" (Bidirectional and Auto-Regressive Transformers) have achieved state-of-the-art performance in summarization tasks.

Sentiment Analysis

- Transformers have proven effective in sentiment analysis, determining the sentiment (positive, negative, or neutral) expressed in a piece of text.
- Models like "BERT" and "GPT" have been fine-tuned for sentiment analysis tasks with high accuracy.

Named Entity Recognition (NER)

- Transformers have shown promising results in NER tasks, which involve identifying and classifying named entities in text.
- They can capture contextual information and learn to recognize various types of entities like person names, organizations, locations, and more.
- Models like "BERT" and "RoBERTa" have achieved state-of-the-art performance in NER tasks.

Chatbots and Conversational AI

- Transformers have been used to build powerful chatbots and conversational AI systems.
- By leveraging transformers, chatbots can understand user queries and generate human-like responses.
- OpenAI's "GPT-3" is an example of a large-scale transformer model used for conversational AI.

Introduction to Natural Language Inference (NLI)

- Natural Language Inference (NLI) involves determining the logical relationship between two given statements: the premise and the hypothesis.
- NLI is a fundamental task in NLP and finds applications in question answering, information retrieval, and dialogue systems.

- Slide 5: Inference Tasks in NLI
- Transformers can be applied to various NLI tasks, including:
  - Recognizing Textual Entailment (RTE)
  - Paraphrase Identification
  - Natural Language Understanding in dialogue systems
- Slide 6: Recognizing Textual Entailment (RTE)
- RTE is the task of determining whether the meaning of the hypothesis is entailed or contradicted by the premise.
- Transformers can encode both the premise and hypothesis, capture their relationships, and provide a prediction of the entailment relationship.
- Slide 7: Paraphrase Identification
- Paraphrase identification involves determining whether two sentences convey the same meaning, even if the wording differs.
- Transformers can leverage their contextual understanding and semantic representations to identify paraphrases accurately.

Inference Tasks in NLI

- Transformers can be applied to various NLI tasks, including:
  - Recognizing Textual Entailment (RTE)
  - Paraphrase Identification
  - Natural Language Understanding in dialogue systems
- Recognizing Textual Entailment (RTE)
  - RTE is the task of determining whether the meaning of the hypothesis is entailed or contradicted by the premise.
  - Transformers can encode both the premise and hypothesis, capture their relationships, and provide a prediction of the entailment relationship.

Conclusion
Transformers in NLP

- Transformers have become a cornerstone of NLP, enabling breakthroughs in various tasks.
- Their ability to model long-range dependencies and capture context has significantly advanced the field.
- Transformers continue to drive innovation and research in NLP, offering exciting possibilities for future applications.

Applications of Transformers in Computer Vision (CV)

Introduction to Transformers in CV

- Transformers, originally developed for natural language processing (NLP), have also gained significant success in the field of Computer Vision.
- Transformers have revolutionized various CV tasks by effectively modeling spatial relationships and capturing global context.

Object Detection

- Transformers have been successfully employed in object detection tasks, which involve locating and classifying objects within an image.
- Models such as DETR (DEtection TRansformer) utilize transformers to handle the object detection pipeline, eliminating the need for handcrafted components like anchor boxes

Image Classification

- Transformers have been applied to image classification tasks, where the goal is to assign a label to an input image.
- Unlike traditional convolutional neural networks (CNNs), transformers process images globally, capturing long-range dependencies.
- Models like Vision Transformer (ViT) have achieved competitive performance in image classification benchmarks.

Semantic Segmentation

- Transformers have shown promise in semantic segmentation tasks, where the objective is to assign a label to each pixel in an image.
- By utilizing self-attention mechanisms, transformers capture contextual information and capture long-range dependencies, improving segmentation accuracy.
- Models like Uformer and TransUNet have achieved competitive results in semantic segmentation benchmarks.

Instance Segmentation

- Transformers have also been applied to instance segmentation, which involves identifying and delineating individual objects within an image.
- Models like Mask-RCNN, which combines transformers with region proposal networks, have achieved state-of-the-art performance in instance segmentation tasks.

Image Generation

- Transformers have been used for image generation tasks, such as generating realistic images from textual descriptions or sketches.
- Models like DALL-E and CLIP have demonstrated the ability to generate high-quality images conditioned on text prompts.

Video Understanding

- Transformers have been extended to video understanding tasks, including action recognition, video captioning, and video object segmentation.
- By leveraging spatio-temporal relationships, transformers capture the motion and temporal context in videos, improving performance in video-related tasks.

- Vision Transformers (ViT) have achieved remarkable success in image classification.
- Used for tasks such as image segmentation, object detection, and face recognition.

Conclusion transformer in CV

- Transformers have significantly impacted the field of Computer Vision by offering a powerful alternative to traditional CNN-based approaches.
- Their ability to model global context, capture long-range dependencies, and handle spatial relationships has advanced various CV tasks.
- Transformers continue to drive innovation and research in Computer Vision, opening up new possibilities for image and video analysis.

Applications of Transformers in Speech and Audio Processing

Introduction to Transformers in Speech and Audio Processing

- Transformers, known for their success in natural language processing (NLP) and computer vision, have also found applications in speech and audio processing tasks.
- Transformers excel at modeling sequential data, making them valuable for analyzing speech and audio signals.

Automatic Speech Recognition (ASR)

- Transformers have been applied to automatic speech recognition, which involves converting spoken language into written text.
- Transformers can model the temporal dependencies in speech signals and capture contextual information to improve transcription accuracy.
- Models like Conformer and ESPnet Transformer ASR have achieved competitive results in ASR benchmarks.

Speech Synthesis

- Transformers have shown promise in speech synthesis, also known as text-to-speech (TTS) conversion.
- By leveraging attention mechanisms, transformers can generate natural-sounding speech from text inputs.
- Models like Transformer TTS and FastSpeech have demonstrated high-quality and expressive speech synthesis capabilities.

Speaker Recognition

- Transformers have been used in speaker recognition tasks, which involve identifying and verifying individuals based on their unique vocal characteristics.
- Transformers can learn discriminative representations for speaker embeddings, enabling accurate speaker recognition.
- Models like X-vector-Transformer have achieved state-of-the-art performance in speaker verification benchmarks.

Music Generation

- Transformers have been employed for music generation tasks, including composing new melodies or generating complete musical pieces.
- By modeling sequential patterns and dependencies in music data, transformers can generate harmonious and coherent musical compositions.
- Models like Music Transformer and PerformanceRNN have demonstrated the ability to generate diverse and creative music.

Environmental Sound Classification

- Transformers have been applied to environmental sound classification tasks, where the goal is to identify and classify sounds from the surrounding environment.
- By capturing the temporal relationships in audio signals, transformers can learn meaningful representations for sound classification.
- Models like Sound-Transformer have achieved competitive performance in environmental sound classification benchmarks.

Conclusion of the application of Transformers in Speech and audio

- Transformers have made significant contributions to the field of speech and audio processing by enabling accurate transcription, synthesis, recognition, and generation of speech and audio signals.
- Their ability to model sequential data and capture temporal dependencies has opened up new possibilities for advancing speech and audio-related applications.
- Transformers continue to drive innovation and research in the field, offering exciting opportunities for further advancements.

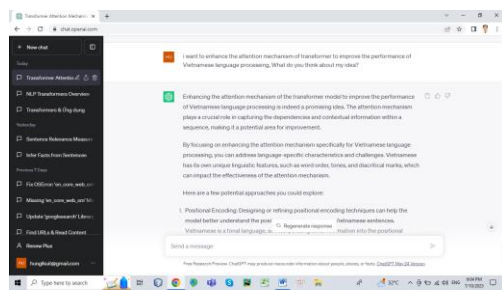
Audio Source Separation

- Transformers have shown promise in audio source separation tasks, which involve isolating individual sound sources from a mixed audio signal.
- Transformers can learn to separate sources by modeling the spectrogram representations of audio signals.
- Models like Conv-TasNet and DPTNet have combined transformers with convolutional neural networks to achieve state-of-the-art results in audio source separation.

Challenges and Future Research

- Training Transformers requires large amounts of data and computational resources.
- Optimizing and compressing Transformers to fit resource-constrained applications.
- Research on transfer learning techniques to utilize Transformers on smaller tasks.

Study about attentionmechanism



Conclusions

- Transformers have revolutionized the way we approach complex AI tasks.
- There are numerous successful applications of Transformers in NLP, CV, and speech processing.
- There is still significant potential to explore and research in the future.

References for transformers and NLP

References

- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3058. [Google Scholar]
- Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805. [Google Scholar]
- Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *OpenAI* **2018**. [Google Scholar]
- Chowdhary, K.; Chowdhary, K. Natural language processing. *Fundam. Artif. Intell.* **2020**, *1*, 603–649. [Google Scholar]
- Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [Google Scholar] [CrossRef]
- Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [Google Scholar] [CrossRef]
- Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, 3–10 March 2021; pp. 610–623. [Google Scholar]
- Dang, N.C.; Moreno-García, M.N.; De la Prieta, F. Sentiment analysis based on deep learning: A comparative study. *Electronics* **2020**, *9*, 483. [Google Scholar] [CrossRef]

References

- Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; Sen, P. A survey of the state of explainable AI for natural language processing. *arXiv* **2020**, arXiv:2010.0071. [Google Scholar]
- Alyafeai, Z.; AlShaibani, M.S.; Ahmad, I. A survey on transfer learning in natural language processing. *arXiv* **2020**, arXiv:2007.04239. [Google Scholar]
- Wu, L.; Chen, Y.; Shen, K.; Guo, X.; Gao, H.; Li, S.; Pei, J.; Long, B. Graph neural networks for natural language processing: A survey. *Found. Trends® Mach. Learn.* **2023**, *16*, 119–328. [Google Scholar] [CrossRef]
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9. [Google Scholar]
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. [Google Scholar]
- Shin, T.; Razeghi, Y.; Logan IV, R.L.; Wallace, E.; Singh, S. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv* **2020**, arXiv:2010.15980. [Google Scholar]
- Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv* **2019**, arXiv:1901.02860. [Google Scholar]

References

- Krause, B.; Kahembwe, E.; Murray, I.; Renals, S. Dynamic evaluation of transformer language models. *arXiv* **2019**, arXiv:1904.08378. [Google Scholar]
- Ziegler, D.M.; Stiennon, N.; Wu, J.; Brown, T.B.; Radford, A.; Amodei, D.; Christiano, P.; Irving, G. Fine-tuning language models from human preferences. *arXiv* **2019**, arXiv:1909.08593. [Google Scholar]
- Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv* **2020**, arXiv:2001.04451. [Google Scholar]
- Zhu, C.; Zeng, M.; Huang, X. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv* **2018**, arXiv:1812.03593. [Google Scholar]
- Garg, S.; Vu, T.; Moschitti, A. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 7780–7788. [Google Scholar]
- Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 3088. [Google Scholar]
- Wu, S.; Cotterell, R.; Hulden, M. Applying the transformer to character-level transduction. *arXiv* **2020**, arXiv:2005.10213. [Google Scholar]
- Zhu, J.; Xia, Y.; Wu, L.; He, D.; Qin, T.; Zhou, W.; Li, H.; Liu, T.Y. Incorporating bert into neural machine translation. *arXiv* **2020**, arXiv:2002.06823. [Google Scholar]
- Yasunaga, M.; Leskovec, J.; Liang, P. Linkbert: Pretraining language models with document links. *arXiv* **2022**, arXiv:2203.15827. [Google Scholar]
- Hosseini, P.; Broniatowski, D.A.; Diab, M. Knowledge-augmented language models for cause-effect relation classification. In Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022), Dublin, UK, 27 May 2022; pp. 43–48. [Google Scholar]
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551. [Google Scholar]
- Liu, Q.; Chen, Y.; Chen, B.; Lou, J.G.; Chen, Z.; Zhou, B.; Zhang, D. You impress me: Dialogue generation via mutual persona perception. *arXiv* **2020**, arXiv:2004.05388. [Google Scholar]
- Guo, T.; Gao, H. Content enhanced bert-based text-to-sql generation. *arXiv* **2019**, arXiv:1910.07179. [Google Scholar]
- Wang, Y.; Wang, W.; Joty, S.; Hoi, S.C. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv* **2021**, arXiv:2109.00859. [Google Scholar]
- Clive, J.; Cao, K.; Rei, M. Control prefixes for parameter-efficient text generation. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Online, 7–11 December 2022; pp. 363–382. [Google Scholar]
- Xiong, W.; Gupta, A.; Toshniwal, S.; Mehdad, Y.; Yih, W.T. Adapting Pretrained Text-to-Text Models for Long Text Sequences. *arXiv* **2022**, arXiv:2209.10052. [Google Scholar]
- Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150. [Google Scholar]
- Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput.*

References

- *Linguist.* **2020**, *8*, 726–742. [Google Scholar] [CrossRef]
- Xiao, W.; Beltagy, I.; Carenini, G.; Cohan, A. Primer: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv* **2021**, arXiv:2110.08499. [Google Scholar]
- Baumel, T.; Eyal, M.; Elhadad, M. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv* **2018**, arXiv:1801.07704. [Google Scholar]
- Ghalandari, D.G.; Hokamp, C.; Ifrim, G. Efficient Unsupervised Sentence Compression by Fine-tuning Transformers with Reinforcement Learning. *arXiv* **2022**, arXiv:2205.08221. [Google Scholar]
- Wang, X.; Jiang, Y.; Bach, N.; Wang, T.; Huang, Z.; Huang, F.; Tu, K. Automated concatenation of embeddings for structured prediction. *arXiv* **2020**, arXiv:2010.05006. [Google Scholar]
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [Google Scholar]
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 24–28 June 2022; pp. 10684–10695. [Google Scholar]
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125. [Google Scholar]
- Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv* **2023**, arXiv:2301.02111. [Google Scholar]
- Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *arXiv* **2022**, arXiv:2205.12005. [Google Scholar]
- Plepi, J.; Kacupaj, E.; Singh, K.; Thakkar, H.; Lehmann, J. Context transformer with stacked pointer networks for conversational question answering over knowledge graphs. In Proceedings of the The Semantic Web: 18th International Conference, ESWC 2021, Online, 6–10 June 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 356–371. [Google Scholar]
- Oguz, B.; Chen, X.; Karpukhin, V.; Peshterliev, S.; Okhonko, D.; Schlichtkrull, M.; Gupta, S.; Mehdad, Y.; Yih, S. Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering. *arXiv* **2020**, arXiv:2012.14610. [Google Scholar]
- Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv* **2022**, arXiv:2208.10442. [Google Scholar]

References for transformers and CV

References of the transformers application in CV

- *Linguist.* **2020**, *8*, 726–742. [Google Scholar] [CrossRef]
- Xiao, W.; Beltagy, I.; Carenini, G.; Cohan, A. Primer: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv* **2021**, arXiv:2110.08499. [Google Scholar]
- Baumel, T.; Eyal, M.; Elhadad, M. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv* **2018**, arXiv:1801.07704. [Google Scholar]
- Ghalandari, D.G.; Hokamp, C.; Ifrim, G. Efficient Unsupervised Sentence Compression by Fine-tuning Transformers with Reinforcement Learning. *arXiv* **2022**, arXiv:2205.08221. [Google Scholar]
- Wang, X.; Jiang, Y.; Bach, N.; Wang, T.; Huang, Z.; Huang, F.; Tu, K. Automated concatenation of embeddings for structured prediction. *arXiv* **2020**, arXiv:2010.05006. [Google Scholar]
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [Google Scholar]
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 24–28 June 2022; pp. 10684–10695. [Google Scholar]
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125. [Google Scholar]

- Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural Codec Language Models are Zero-Shot Text to

References of the transformers application in CV

- Speech Synthesizers. *arXiv* **2023**, arXiv:2301.02111. [Google Scholar]
- Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *arXiv* **2022**, arXiv:2205.12005. [Google Scholar]
- Plepi, J.; Kacupaj, E.; Singh, K.; Thakkar, H.; Lehmann, J. Context transformer with stacked pointer networks for conversational question answering over knowledge graphs. In Proceedings of the The Semantic Web: 18th International Conference, ESWC 2021, Online, 6–10 June 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 356–371. [Google Scholar]
- Oguz, B.; Chen, X.; Karpukhin, V.; Peshterliev, S.; Okhonko, D.; Schlichtkrull, M.; Gupta, S.; Mehdad, Y.; Yih, S. Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering. *arXiv* **2020**, arXiv:2012.14610. [Google Scholar]
- Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv* **2022**, arXiv:2208.10442. [Google Scholar]

References of the transformers and speech and audio speech

References of the transformers and speech and audio speech

- S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang et al., “A comparative study on transformer vs RNN in speech applications,” in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp. 449–456. [2] T. Lin, Y. Wang, X. Liu, and X. Qiu, “A survey of transformers,” AI Open, 2022. [3] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, “Efficient transformers: A survey,” ACM Computing Surveys, vol. 55, no. 6, pp. 1–28, 2022. [4] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, “Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss,” in ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7829–7833. [5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz et al., “Huggingface’s transformers: State-of-the-art natural language processing,” arXiv preprint arXiv:1910.03771, 2019. [6] ——, “Transformers: State-of-the-art natural language processing,” in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45. [7] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, “Learning deep transformer models for machine translation,” arXiv preprint arXiv:1906.01787, 2019. [8] L. Dong, S. Xu, and B. Xu, “Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition,” in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018, pp. 5884–5888. [9] Q. Song, B. Sun, and S. Li, “Multimodal sparse transformer network for audio-visual speech recognition,” IEEE Transactions on Neural Networks and Learning Systems, 2022. [10] A. Yang, A. Miech, J. Sivic, I. Laptev, and

References of the transformers and speech and audio speech

- C. Schmid, “Just ask: Learning to answer questions from millions of narrated videos,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1686–1697. [11] F. Dang, H. Chen, and P. Zhang, “Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 6857–6861. [12] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, “Dawn of the transformer era in speech emotion recognition: closing the valence gap,” arXiv preprint arXiv:2203.07378, 2022. [13] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation,” in ICASSP 2021- 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 21–25. [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010. [15] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, “Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned,” arXiv preprint arXiv:1905.09418, 2019. [16] S. Latif, H. Cuayahuitl, F. Pervez, F. Shamshad, H. S.

References of the transformers and speech and audio speech

- Ali, and ´ E. Cambria, “A survey on deep reinforcement learning for audio-based applications,” Artificial Intelligence Review, vol. 56, no. 3, pp. 2193– 2240, 2023. [17] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, “Deep learning for environmentally robust speech recognition: An overview of recent developments,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 9, no. 5, pp. 1–28, 2018. [18] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” IEEE access, vol. 7, pp. 19 143–19 165, 2019. [19] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech emotion recognition using deep learning techniques: A review,” IEEE Access, vol. 7, pp. 117 327–117 345, 2019. [20] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, “Survey of deep representation learning for speech emotion recognition,” IEEE Transactions on Affective Computing, 2021. [21] L. Deng, “Deep learning: from speech recognition to language and multimodal processing,” APSIPA Transactions on Signal and Information Processing, vol. 5, p. e1, 2016. [22] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, “Survey on deep neural networks in speech and vision systems,” Neurocomputing, vol. 417, pp. 302–321, 2020. [23] A. M. Bras¸oveanu and R. Andonie, “Visualizing transformers for nlp: a brief survey,” in 2020 24th International Conference Information Visualisation (IV). IEEE, 2020, pp. 270–279. [24] X. Tan, T. Qin, F.

References of the transformers and speech and audio speech

- Soong, and T.-Y. Liu, “A survey on neural speech synthesis,” arXiv preprint arXiv:2106.15561, 2021. [25] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. Alturki, F. Alshehri, and M. Almojil, “Automatic speech recognition: Systematic literature review,” IEEE Access, vol. 9, pp. 131 858–131 876, 2021. [26] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey,” Multimedia Tools and Applications, vol. 80, pp. 9411–9457, 2021. [27] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, “A survey of visual transformers,” arXiv preprint arXiv:2111.06091, 2021. [28] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal learning with transformers: A survey,” arXiv preprint arXiv:2206.06488, 2022. [29] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, “Transformers in medical imaging: A survey,” arXiv preprint arXiv:2201.09873, 2022. [30] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, “Transformer models for text-based emotion detection: a review of BERT-based approaches,” Artificial Intelligence Review, vol. 54, no. 8, pp. 5789– 5829, 2021.

Q&A

- Thank you
- Discussion