1. Seaborn K, Miyake N P, Pennefather P, et al. Voice in Human–Agent Interaction: A Survey[J]. ACM Computing Surveys (CSUR), 2021, 54(4): 1-43.
2. 中泰证券研究所. Cerence：人工智能助力未来出行-车智能化海外公司系列报告[EB/OL]. (2021-04-25)[2021-12-25]. https:. pdf.dfcfw.com/pdf/H3\_AP202104261487921362\_1.pdf?1619458440000.pdf.
3. 陈艳华. 基于智能交互的车载语音系统的设计与实现[D]. 北京: 北京交通大学, 2020, 1-2.
4. Deriu J, Rodrigo A, Otegi A, et al. Survey on evaluation methods for dialogue systems[J]. Artificial Intelligence Review, 2021, 54(1): 755-810.
5. Rigoll G, Neukirchen C. A new approach to hybrid HMM/ANN speech recognition using mutual information neural networks[C]. NIPS. 1996: 772-778.
6. Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition[C]. Nips workshop on deep learning for speech recognition and related applications. 2009, 1(9): 39.
7. Abdel-Hamid O, Mohamed A, Jiang H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on audio, speech, and language processing, 2014, 22(10): 1533-1545.
8. Graves A. Connectionist temporal classification[M]. Supervised Sequence Labelling with Recurrent Neural Networks. Springer, Berlin, Heidelberg, 2012: 61-93.
9. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]. Advances in neural information processing systems. 2014: 3104-3112.
10. Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]. Proceedings of the 23rd international conference on Machine learning. 2006: 369-376.
11. Graves A. Sequence transduction with recurrent neural networks[J]. arXiv preprint arXiv:1211.3711, 2012.
12. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]. 2013 IEEE international conference on acoustics, speech and signal processing. Ieee, 2013: 6645-6649.
13. Graves A. Generating sequences with recurrent neural networks[J]. arXiv preprint arXiv:1308.0850, 2013.
14. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
15. Chorowski J, Bahdanau D, Cho K, et al. End-to-end continuous speech recognition using attention-based recurrent NN: First results[J]. arXiv preprint arXiv:1412.1602, 2014.
16. Bahar P, Zeyer A, Schlüter R, et al. On using 2d sequence-to-sequence models for speech recognition[C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5671-5675.
17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in neural information processing systems. 2017: 5998-6008.
18. Zhou S, Dong L, Xu S, et al. A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese[C]. International Conference on Neural Information Processing. Springer, Cham, 2018: 210-220.
19. Merboldt A, Zeyer A, Schlüter R, et al. An Analysis of Local Monotonic Attention Variants[C]. INTERSPEECH. 2019: 1398-1402.
20. Zeyer A, Irie K, Schlüter R, et al. Improved training of end-to-end attention models for speech recognition[J]. arXiv preprint arXiv:1805.03294, 2018.
21. Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented Transformer for Speech Recognition[J]. arXiv preprint arXiv:2005.08100, 2020.
22. Qin L, Xie T, Che W, et al. A survey on spoken language understanding: Recent advances and new frontiers[J]. arXiv preprint arXiv:2103.03095, 2021.
23. Ravuri S, Stolcke A. Recurrent neural network and LSTM models for lexical utterance classification[C]. Sixteenth Annual Conference of the International Speech Communication Association. 2015: 35-40
24. Yao K, Zweig G, Hwang M Y, et al. Recurrent neural networks for language understanding[C]. Interspeech. 2013: 2524-2528.
25. Mesnil G, He X, Deng L, et al. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding[C]. Interspeech. 2013: 3771-3775.
26. Mesnil G, Dauphin Y, Yao K, et al. Using recurrent neural networks for slot filling in spoken language understanding[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(3): 530-539.
27. Zhang X, Wang H. A joint model of intent determination and slot filling for spoken language understanding[C]. IJCAI. 2016, 16: 2993-2999.
28. Hakkani-Tür D, Tür G, Celikyilmaz A, et al. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm[C]. Interspeech. 2016: 715-719.
29. Goo C W, Gao G, Hsu Y K, et al. Slot-gated modeling for joint slot filling and intent prediction[C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 753-757.
30. Li C, Li L, Qi J. A self-attentive model with gate mechanism for spoken language understanding[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3824-3833.
31. Qin L, Che W, Li Y, et al. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 2078-2087.
32. Niu P, Chen Z, Song M. A novel bi-directional interrelated model for joint intent detection and slot filling[J]. arXiv preprint arXiv:1907.00390, 2019.
33. Zhang C, Li Y, Du N, et al. Joint Slot Filling and Intent Detection via Capsule Neural Networks[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5259-5267.
34. Liu Y, Meng F, Zhang J, et al. CM-Net: A Novel Collaborative Memory Network for Spoken Language Understanding[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 1051-1060.
35. Zhang L, Ma D, Zhang X, et al. Graph lstm with context-gated mechanism for spoken language understanding[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 9539-9546.
36. Qin L, Liu T, Che W, et al. A co-interactive transformer for joint slot filling and intent detection[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 8193-8197.
37. Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.
38. Chen Q, Zhuo Z, Wang W. Bert for joint intent classification and slot filling[J]. arXiv preprint arXiv:1902.10909, 2019.
39. Chen H, Liu X, Yin D, et al. A survey on dialogue systems: Recent advances and new frontiers[J]. Acm Sigkdd Explorations Newsletter, 2017, 19(2): 25-35.
40. McTear M F. Spoken dialogue technology: enabling the conversational user interface[J]. ACM Computing Surveys (CSUR), 2002, 34(1): 90-169.
41. 赵新颜. 基于深度学习的对话系统研究与应用[D]. 合肥: 中国科学技术大学, 2022, 13-14.
42. Ni J, Young T, Pandelea V, et al. Recent advances in deep learning based dialogue systems: A systematic survey[J]. Artificial intelligence review, 2022: 1-101.
43. Jorgensen P E T, Song M S, Tian J. Operator theory, kernels, and Feedforward Neural Networks[J]. arXiv preprint arXiv:2301.01327, 2023.
44. Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088): 533-536.
45. Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE transactions on neural networks and learning systems, 2021.
46. Hewamalage H, Bergmeir C, Bandara K. Recurrent neural networks for time series forecasting: Current status and future directions[J]. International Journal of Forecasting, 2021, 37(1): 388-427.
47. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
48. Yao S, Wan X. Multimodal transformer for multimodal machine translation[C]. Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 4346-4350.
49. Du Y, Pei B, Zhao X, et al. Deep scaled dot-product attention based domain adaptation model for biomedical question answering[J]. Methods, 2020, 173: 69-74.
50. Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]. 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017: 2227-2231.
51. Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition[C]. 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016: 4960-4964.
52. Karmakar P, Teng S W, Lu G. Thank you for attention: a survey on attention-based artificial neural networks for automatic speech recognition[J]. arXiv preprint arXiv:2102.07259, 2021.
53. Wang Y, Shen Y, Jin H. A bi-model based rnn semantic frame parsing model for intent detection and slot filling[J]. arXiv preprint arXiv:1812.10235, 2018.
54. Pham N Q, Nguyen T S, Niehues J, et al. Very deep self-attention networks for end-to-end speech recognition[J]. arXiv preprint arXiv:1904.13377, 2019.
55. Synnaeve G, Xu Q, Kahn J, et al. End-to-end asr: from supervised to semi-supervised learning with modern architectures[J]. arXiv preprint arXiv:1911.08460, 2019.
56. Luo H, Zhang S, Lei M, et al. Simplified self-attention for transformer-based end-to-end speech recognition[C]. 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021: 75-81.
57. Zhang S, Liu C, Jiang H, et al. Feedforward sequential memory networks: A new structure to learn long-term dependency[J]. arXiv preprint arXiv:1512.08301, 2015.
58. Mehta S, Ghazvininejad M, Iyer S, et al. Delight: Deep and light-weight transformer[J]. arXiv preprint arXiv:2008.00623, 2020.
59. Wen Q, Zhou T, Zhang C, et al. Transformers in time series: A survey[J]. arXiv preprint arXiv:2202.07125, 2022.
60. Misra D. Mish: A self regularized non-monotonic activation function[J]. arXiv preprint arXiv:1908.08681, 2019.
61. Lin T, Wang Y, Liu X, et al. A survey of transformers[J]. AI Open, 2022, 1(1): 1-16.
62. Wang D, Zhang X. Thchs-30: A free chinese speech corpus[J]. arXiv preprint arXiv:1512.01882, 2015.
63. Bu H, Du J, Na X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C]. 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA). IEEE, 2017: 1-5.
64. Rousseau A, Deléglise P, Esteve Y. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks[C]. LREC. 2014: 3935-3939.
65. Watanabe S, Hori T, Karita S, et al. Espnet: End-to-end speech processing toolkit[J]. arXiv preprint arXiv:1804.00015, 2018.
66. Malik M, Malik M K, Mehmood K, et al. Automatic speech recognition: a survey[J]. Multimedia Tools and Applications, 2021, 80: 9411-9457.
67. Whitehead N, Fit-Florea A. Precision & performance: Floating point and IEEE 754 compliance for NVIDIA GPUs[J]. rn (A+ B), 2011, 21(1): 18749-19424.
68. Higuchi Y, Inaguma H, Watanabe S, et al. Improved Mask-CTC for non-autoregressive end-to-end ASR[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 8363-8367.
69. Ghahremani P, BabaAli B, Povey D, et al. A pitch extraction algorithm tuned for automatic speech recognition[C]. 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2014: 2494-2498.
70. Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018: 66-71.
71. Povey D, Peddinti V, Galvez D, et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI[C]. Interspeech. 2016: 2751-2755.
72. Shan C, Weng C, Wang G, et al. Component fusion: Learning replaceable language model component for end-to-end speech recognition system[C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5361-5635.
73. Dong L, Xu S, Xu B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition[C]. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 5884-5888.
74. Xu M, Li S, Zhang X L. Transformer-based end-to-end speech recognition with local dense synthesizer attention[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 5899-5903.
75. Chen X, Zhang S, Song D, et al. Transformer with bidirectional decoder for speech recognition[J]. arXiv preprint arXiv:2008.04481, 2020.
76. Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.
77. Kannan A, Wu Y, Nguyen P, et al. An analysis of incorporating an external language model into a sequence-to-sequence model[C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 1-5828.
78. Wu T W, Su R, Juang B. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 4884-4896.
79. del Pino G E, Galaz H. Statistical applications of the inverse gram matrix: A revisitation[J]. Brazilian Journal of Probability and Statistics, 1995: 177-196.
80. Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
81. Qin L, Xu X, Che W, et al. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling[J]. arXiv preprint arXiv:2004.10087, 2020.
82. Hemphill C T, Godfrey J J, Doddington G R. The ATIS spoken language systems pilot corpus[C]. Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990. 1990.
83. Coucke A, Saade A, Ball A, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces[J]. arXiv preprint arXiv:1805.10190, 2018.
84. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
85. Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling[J]. arXiv preprint arXiv:1609.01454, 2016.
86. Gangadharaiah R, Narayanaswamy B. Joint multiple intent detection and slot labeling for goal-oriented dialog[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 564-569.
87. Chen L, Zhou P, Zou Y. Joint multiple intent detection and slot filling via self-distillation[C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7612-7616.
88. Süzen A A, Duman B, Şen B. Benchmark analysis of jetson tx2, jetson nano and raspberry pi using deep-cnn[C]. 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2020: 1-5.
89. Lee C H, Cheng H, Ostendorf M. Dialogue State Tracking with a Language Model using Schema-Driven Prompting[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 4937-4949.
90. Peng B, Li X, Gao J, et al. Deep Dyna-Q: Integrating Planning for Task-Completion Dialogue Policy Learning[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2182-2192.
91. Yu W, Jiang M, Hu Z, et al. Knowledge-Enriched Natural Language Generation[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts. 2021: 11-16.
92. Łańcucki A. Fastpitch: Parallel text-to-speech with pitch prediction[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6588-6592.