[42] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention net-

works for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Asso-

ciation for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San

Diego, California, 1480–1489. DOI:https://doi.org/10.18653/v1/N16-1174

[43] Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018. Modelling do-

main relationships for transfer learning on retrieval-based question answering systems in e-commerce. In Proceedings

of the 11th ACM International Conference on Web Search and Data Mining. 682–690.

[44] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip

Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Se-

quences. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Bal-

can, and H. Lin (Eds.). Vol. 33. Curran Associates, Inc., 17283–17297. https://proceedings.neurips.cc/paper/2020/file/

c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

[45] Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Pooling-

former: Long document modeling with pooling attention. In Proceedings of the 38th International Conference on Ma-

chine Learning, Marina Meila and Tong Zhang (Eds.). PMLR, 12437–12446. Retrieved from https://proceedings.mlr.

press/v139/zhang21h.html.

[39] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In Proceedings of the 2019 Conference on

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language

Processing. 11–20.

[40] Liu Yang, Qingyao Ai, Jiafeng Guo, and W. Bruce Croft. 2016. aNMM: Ranking short answer texts with attention-based

neural matching model. In Proceedings of the 25th ACM International on Conference on Information and Knowledge

Management. 287–296.

[41] Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-

depth transformer-based hierarchical encoder for long-form document matching. In Proceedings of the 29th ACM

International Conference on Information and Knowledge Management. 1725–1734.

[36] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differen-

tiable models by constraining their explanations. In Proceedings of the 26th International Joint Conference on Artificial

Intelligence. 2662–2670.

[37] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In Proceedings of

the International Conference on Machine Learning. 3319–3328.

[25] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The efficient transformer. In Proceedings of the

International Conference on Learning Representations.

[26] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the

International Conference on Machine Learning. 1188–1196.

[27] Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In Proceedings of

the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,

Florence, Italy, 5070–5081. DOI:https://doi.org/10.18653/v1/P19-1500

[28] Dongsheng Luo, Wei Cheng, Jingchao Ni, Wenchao Yu, Xuchao Zhang, Bo Zong, Yanchi Liu, Zhengzhang Chen,

Dongjin Song, Haifeng Chen, and Xiang Zhang. 2021. Unsupervised document embedding via contrastive augmenta-

tion. arXiv:2103.14542. Retrieved from https://arxiv.org/abs/2103.14542.

[29] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations

of text for web search. In Proceedings of the 26th International Conference on World Wide Web. 1291–1299.

[30] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image

recognition. In Proceedings of the AAAI Conference on Artificial Intelligence.

[32] Jiezhong Qiu, Hao Ma, Omer Levy, Wen-tau Yih, Sinong Wang, and Jie Tang. 2020. Blockwise self-attention for long

document understanding. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:

Findings. 2555–2565.

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu,

and Dilip Krishnan. 2020. Supervised Contrastive Learning. In Advances in Neural Information Processing Systems,

H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.). Vol. 33. Curran Associates, Inc., 18661–18673.

https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

[16] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval.

In Proceedings of the 25th ACM International Conference on Information and Knowledge Management. 55–64.

[17] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. 2019. Axial attention in multidimensional

transformers. arXiv:1912.12180. Retrieved from https://arxiv.org/abs/1912.12180.

[18] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional Neural Network Architectures for

Matching Natural Language Sentences. In Advances in Neural Information Processing Systems, Z. Ghahramani, M.

Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.). Vol. 27. Curran Associates, Inc. 2042–2050. https:

//proceedings.neurips.cc/paper/2014/file/b9d487a30398d42ecff55c228ed5652b-Paper.pdf.

[22] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast

autoregressive transformers with linear attention. In Proceedings of the International Conference on Machine Learning.

5156–5165.

[1] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification.

arXiv:1904.08398. Retrieved from https://arxiv.org/abs/1904.08398.

[2] David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box

sequence-to-sequence models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Pro-

cessing. 412–421.

[5] Minmin Chen. 2017. Efficient vector representation for documents through corruption. In Proceedings of the ICLR

(Poster).

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive

learning of visual representations. In Proceedings of the International Conference on Machine Learning. 1597–1607.

[7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers.

arXiv:1904.10509. Retrieved from https://arxiv.org/abs/1904.10509.

[8] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter

Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2020. Re-

thinking attention with performers. In Proceedings of the International Conference on Learning Representations.

[9] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL:

Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association

for Computational Linguistics. 2978–2988.

[11] Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. CogLTX: Applying BERT to Long Texts. In

Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan,

and H. Lin (Eds.). Vol. 33. Curran Associates, Inc., 12792–12804. https://proceedings.neurips.cc/paper/2020/file/

96671501524948bc3937b4b30d0e57b9Paper.pdf.

[12] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational

Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.

552.

[13] Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: A

case study on natural language inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Lan-

guage Processing. 4952–4957.

[14] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised

textual representations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and

the 11th International Joint Conference on Natural Language Processing. 879–895.

[15] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures.