Notes on Attention in Deep Learning, Entropic Transport Interpretation and Related Causal Models

compiled by D. Gueorguiev, 8/22/25

# Introductory Notes

We will follow generally the organization of the excellent summary on the topic of Attention by D. Benveniste in [1].

## Attention in Statistical Machine Translation

Notable paper credited with the application of Attention in NLP tasks for SMT is

# References

Attention in Deep Learning

[1] [All You Need To Know About The Self-Attention Layer, Damien Benveniste, The AiEdge, 2025](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/attention/All_You_Need_To_Know_About_The_Self-Attention_Layer_Beneveniste_2025.pdf)

[2] [Attention is All You Need, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit et al, NIPS, 2017](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Attention-is-all-you-need-NIPS-2017.pdf)

[3] [Attention in Natural Language Processing, Andrea Galassi, Marco Lippi, Paolo Torroni, 2019](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/AttentionInNaturalLanguageProcessing.pdf)

[4] [A Structured Self-Attentive Sentence Embedding, Z. Lin et al, 2017](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/attention/A_Structured_Self-attentive_Sentence_Embedding_Lin_2017.pdf)

[5] [Attention and Augmented Recurrent Neural Networks, Chris Olah, Shan Carter, Google Brain blog, 2016](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/AttentionandAugmentedRecurrentNeuralNetworks.pdf)

[6] [Long Short-Term Memory-Networks for Machine Reading, J. Cheng et al, 2016](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/attention/Long_Short-Term_Memory-Networks_for_Machine_Reading_Cheng_2016.pdf)

[7] [A Decomposable Attention Model for Natural Language Inference, Ankur Parikh et al, 2016](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/attention/A_Decomposable_Attention_Model_for_Natural_Language_Inference_Parikh_2016.pdf)

[8] [Neural Machine Translation by Jointly Learning To Align and Translate, Dzmitry Bahdanau, K. Cho, Yoshua Bengio, 2016](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/NeuralMachineTranslationByJointlyLearningToAlignAndTranslateBahdanau2015.pdf)

[9] [Recurrent models of visual attention, V. Mnih et al, Google DeepMind, 2014](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Recurrent_Models_of_Visual_Attention_Mnih_2014.pdf)

[10] [Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks, Juergen Schmidhuber et al, 1992](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/attention/Learning_to_Control_Fast-Weight_Memories_1992-schmidhuber.pdf)

Information-theoretical Interpretation of Attention

[] [Scaled-Dot-Product Attention as One-Sided Entropic Optimal Transport, Elon Litman, 2025](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/attention/Scaled-Dot-Product_Attention_as_One-Sided_Entropic_Optimal_Transport_Littman_2025.pdf)

Causal Modeling and its connections to Attention

[] [The Mathematics of Causality, Miquel Noguer i Alonso, 2025](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/attention/The_Mathematics_of_Causality_Noguer_i_Alonso_2025.pdf)

Other Interpretations of Attention

[] [Deriving Machine Attention from Human Rationales, Y. Bao et al, CSAIL-MIT, IBM Watson, 2018](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Deriving_Machine_Attention_from_Human_Rationales_Bao_IBM-Watson_2018.pdf)

Statistical Machine Translation, RNN, LSTM, Autoencoders

[] [Bidirectional Recurrent Neural Networks, Mike Schuster, Kuldip Paliwal, IEEE Transactions on Signal Processing, 1997](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Bidirectional_Recurrent_Neural_Networks_Schuster_Paliwal_1997.pdf)

[] [Neural Networks for Pattern Recognition, Christopher M. Bishop, 1995](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Neural_Networks_for_Pattern_Recognition-Bishop_1995.pdf)

[] [Recurrent Continuous Translation Models, Nal Kalchbrenner, Phil Blunsom, Oxford U., 2013](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Recurrent_continuous_translation_models_Kalchbrenner_Blunsson_OxfordU_2013.pdf)

[] [On The Properties of Neural Machine Translation: Encoder-Decoder Approaches, K. Cho, B. van Merrienboer, Dzmitry Bahdanau, 2014](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/OnthePropertiesOfNeuralMachineTranslationEncoderDecoderApproaches.pdf)

[] [Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, K. Cho, B. van Merrienboer, C. Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, 2014](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/LearningPhraseRepresentationsUsingRNNEncoderDecoderForStatisticalMachineTranslation.pdf)

[] [Autoencoders, Dor Bank, Noam Koenigstein, Raja Giryes, 2021](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Autoencoders.pdf)

[] [Transforming Auto-encoders, G.E. Hinton, A. Krizhevsky, S.D. Wang, U of Toronto, 2010](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/TransformingAutoencodersHinton.pdf)

[] [Sequence to Sequence Learning with Neural Networks, Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Google Research, 2014](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/SequencetoSequenceLearningwithNeuralNetworksSutsekver2014.pdf)

[] [Generating Sequences With Recurrent Neural Networks, Alex Graves, U of Toronto, 2014](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/Generating_Sequences_With_Recurrent_Neural_Networks_Graves_2014.pdf)

[] [A Tutorial on Training Recurrent Neural Networks, covering BPPT, RTRL, EKF and the “echo state network” approach, Herbert Jaeger, Fraunhofer Institute for Autonomous Intelligent Systems, 2003](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/TutorialOnRNNAndBPTTJaeger2002.pdf)

[] [Understanding LSTM: a Tutorial into Long Short-Term Memory Recurrent Neural Networks, Ralf C. Staudemeyer, Eric Rothstein Morris, 2019](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/TutorialOnLongShortTermMemory2019.pdf)

[] [Long Short-Term Memory, Sepp Hochreiter, Juergen Schmidhuber, Neural Computation 9(8), 1997](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/LongShortTermMemory.pdf)

[] [Sequence Modeling With Neural Networks (Part 1): Language & Seq2Seq, Nathan Lintz’s blog, 2016](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/SequenceModelingWithNeuralNetworksPart1_Seq2Seq.pdf)

[] [Sequence Modeling With Neural Networks (Part 2): Attention Models, Nathan Lintz’s blog, 2016](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/SequenceModelingwithNeuralNetworksPart2_AttentionModels.pdf)

[] [Learning to combine foveal glimpses with a third-order Boltzmann machine, H. Larochelle and G. Hinton, 2010](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/NIPS-2010-learning-to-combine-foveal-glimpses-with-a-third-order-boltzmann-machine-Paper.pdf)

[] [Sequence to Sequence Learning with Neural Networks, Sutskever et al, Google Research, 2014](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/SequencetoSequenceLearningwithNeuralNetworksSutsekver2014.pdf)