Faster Transformers for Document Summarization

Zaid Nabulsi, Dian Ang Yap, Vineet Kosaraju

Background

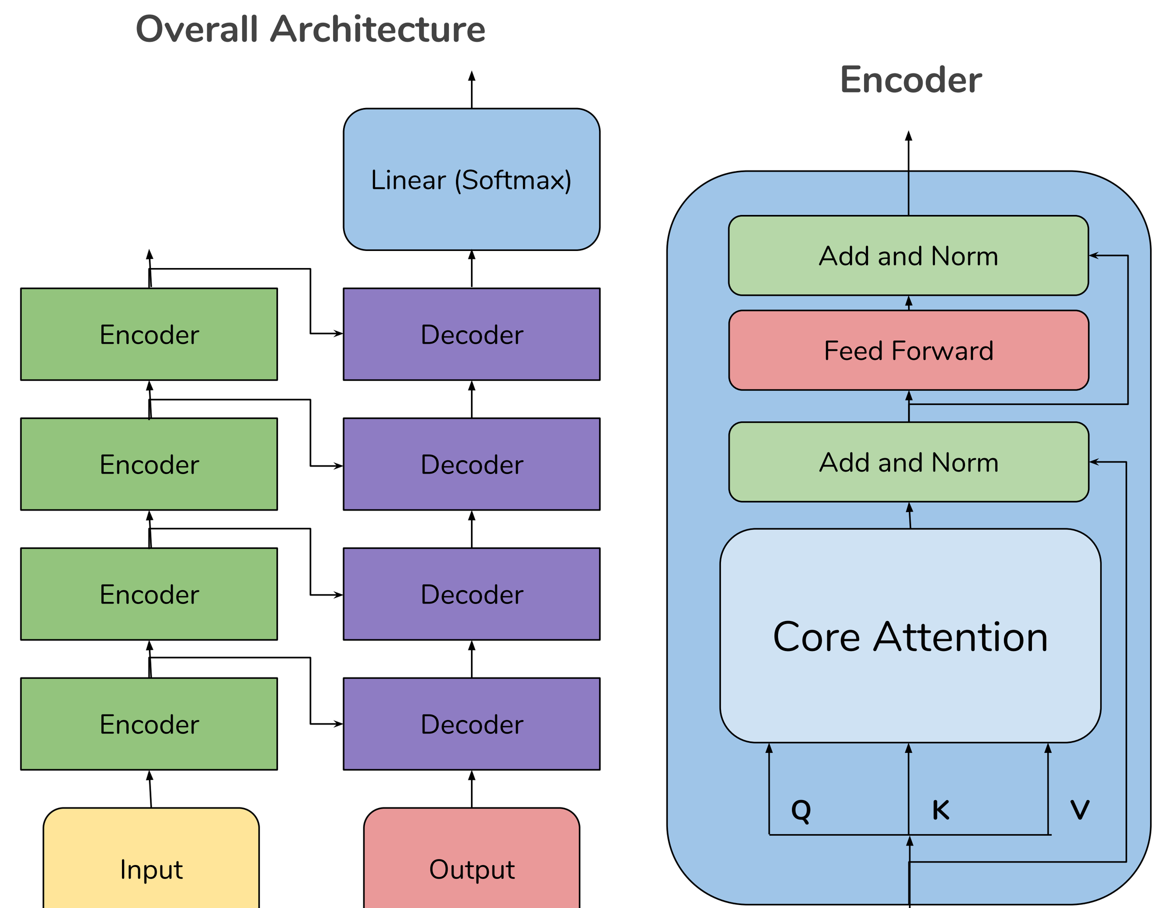
Task: Long Document Summarization

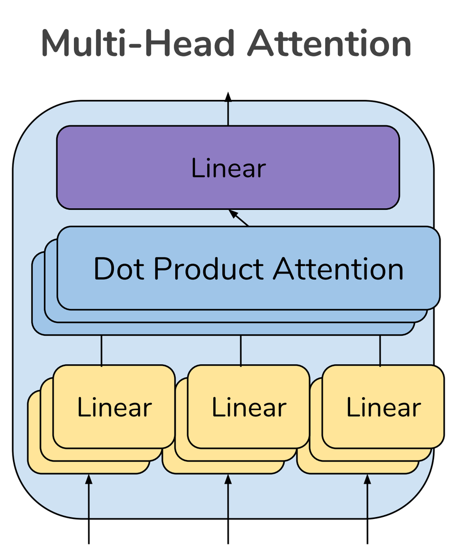
CNN Daily Mail

Document summarization has been done through vanilla RNNs, RL agents, and transformers. Transformers are very promising but are difficult to train as there attention layers serve as bottleneck. We present architectural design modifications to improve both efficiency and performance.

Approach & Methods

Transformer Architecture

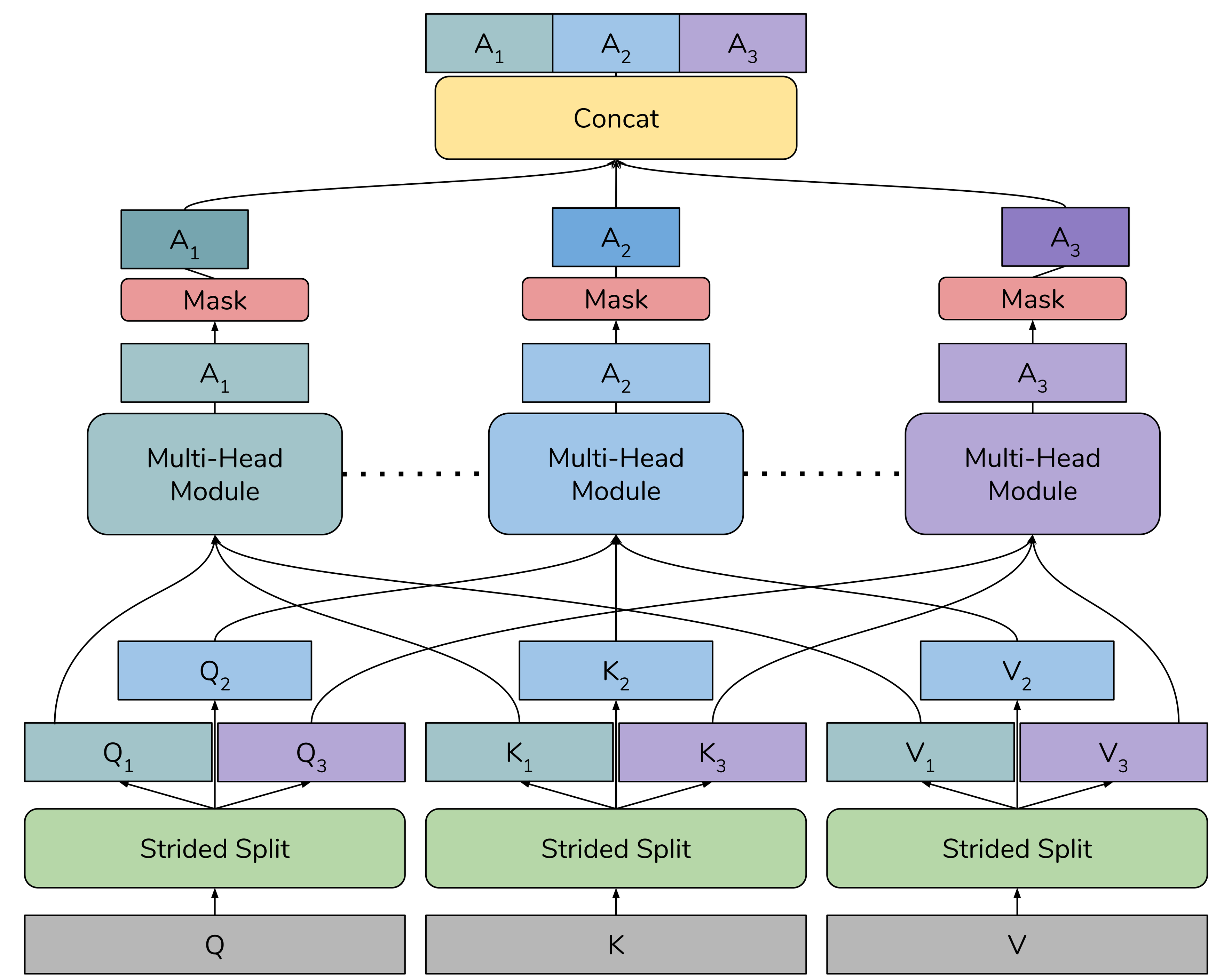




$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

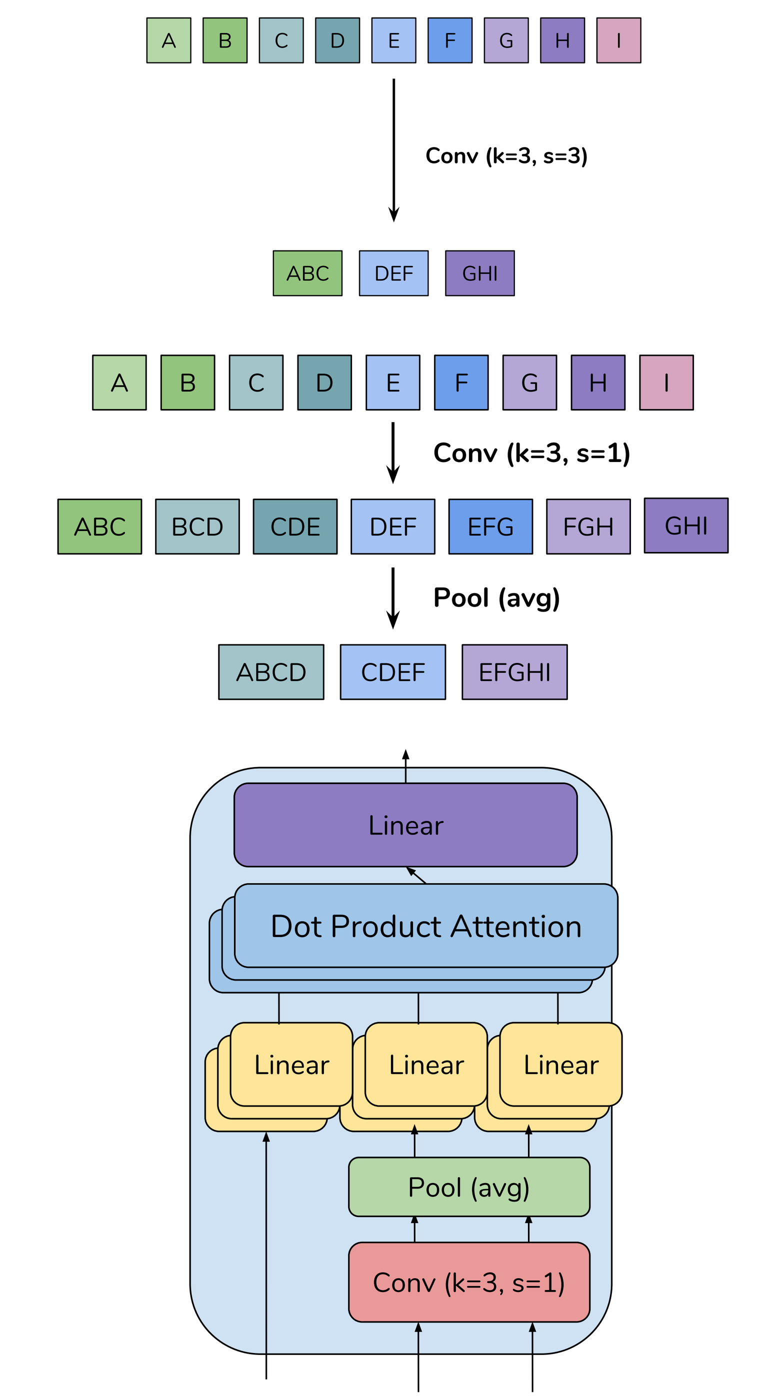
Layer	Runtime
RNN (Recurrent)	$O(n \cdot d^2)$
Transformer (Baseline)	$O(n^2 \cdot d)$

Strided Neighborhood Attention

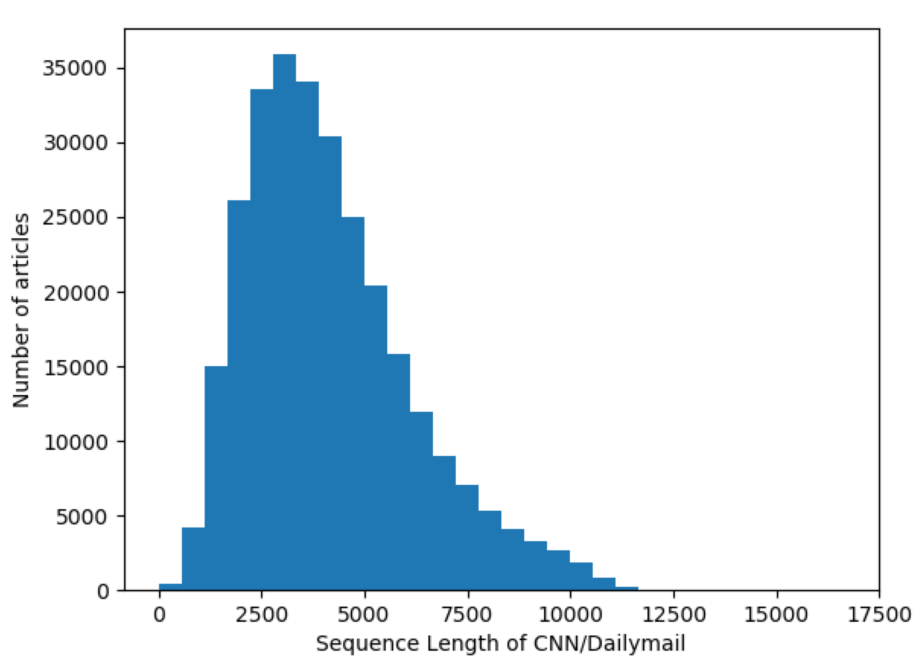


Convolutional Attention

$$\text{MultiHead}(Q, \text{Conv1D}(K^\top)^\top, \text{Conv1D}(V^\top)^\top)$$



Dataset

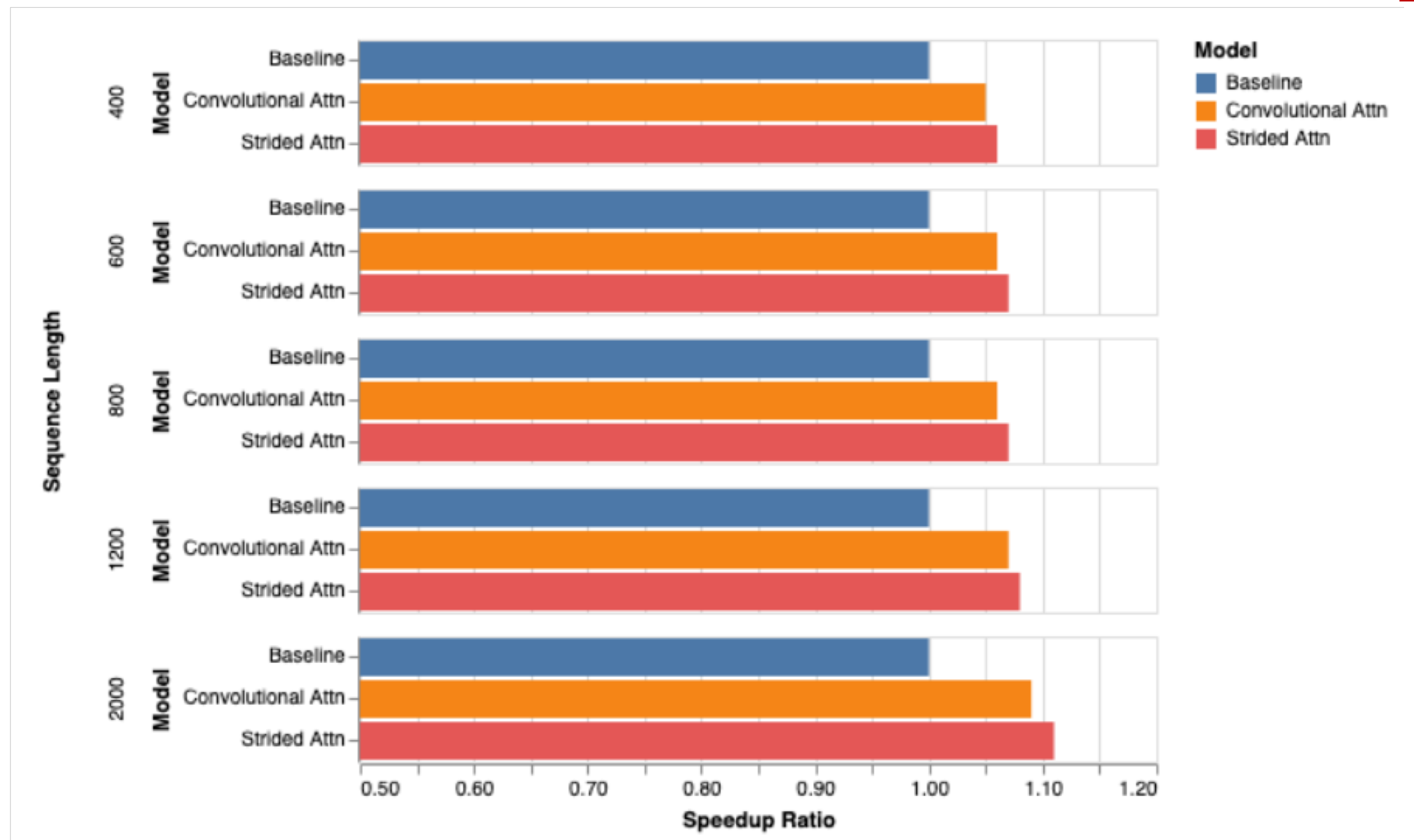


Data was split into train/val/test with a **92/4/4** ratio. The sequence length in the dataset ranges from 250 tokens to 16652 tokens, with a mean of 4153 tokens and a standard deviation of 2014 tokens.

Results

Attention	Accuracy		Perplexity		Speed (Tokens/s)		Theoretical Runtime
	Training	Validation	Training	Validation	Training	Inference	
Baseline	56.12	56.55	7.65	9.26	5.96	54.48	$O(n^2)$
Conv.	56.51	56.12	7.50	9.41	6.28	57.70	$O((\frac{n}{s})^2)$
Strided	**56.62**	**56.77**	**7.49**	**9.01**	**6.29**	**58.18**	$O(\frac{n^2}{c})$

	Baseline Attention	Convolutional Attention	Strided Attention
ROUGE-1 Recall	38.60	**39.26**	37.79
ROUGE-1 Precision	41.90	41.19	**42.33**
ROUGE-1 F Score	**38.82**	38.77	38.53
ROUGE-2 Recall	16.64	**16.86**	16.40
ROUGE-2 Precision	18.47	18.06	**18.83**
ROUGE-2 F Score	16.89	16.80	**16.91**
ROUGE-L Recall	35.62	**36.38**	34.93
ROUGE-L Precision	38.76	38.27	**39.24**
ROUGE-L F Score	35.87	**35.97**	35.67



Document to Summarize (model input)	Baseline	Convolutional	Strided
five americans who were monitored for three weeks ... almost all the deaths have been in guinea, liberia and sierra leone. more than 10,000 ...died... ebola is spread by direct contact... .	the last of 17 ... released ... more than 10,000 people died...	the last of 17 ... released ... more than 10,000 people died...	more than 10,000 people ... in guinea, liberia and sierra leone.

Conclusions

- Presented two novel models with architectural improvements to transformers that allow for more efficient training while maintaining (and even exceeding) comparable metrics to existing state-of-the-art methods on document summarization.
- As next steps, combining the models might result in even better performance.

References

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 06 2017.

P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences. International Conference on Learning Representations, 01 2018.

S. Gehrmann, Y. Deng, and A. M. Rush. Bottom-up abstractive summarization. CoRR, abs/1808.10792, 2018

We would like to acknowledge the help of Kevin Clark and Abigail See in mentoring us on this project and helping us brainstorm theoretical contributions, as well as giving feedback on datasets, experiments, visualizations, and our milestone reports.