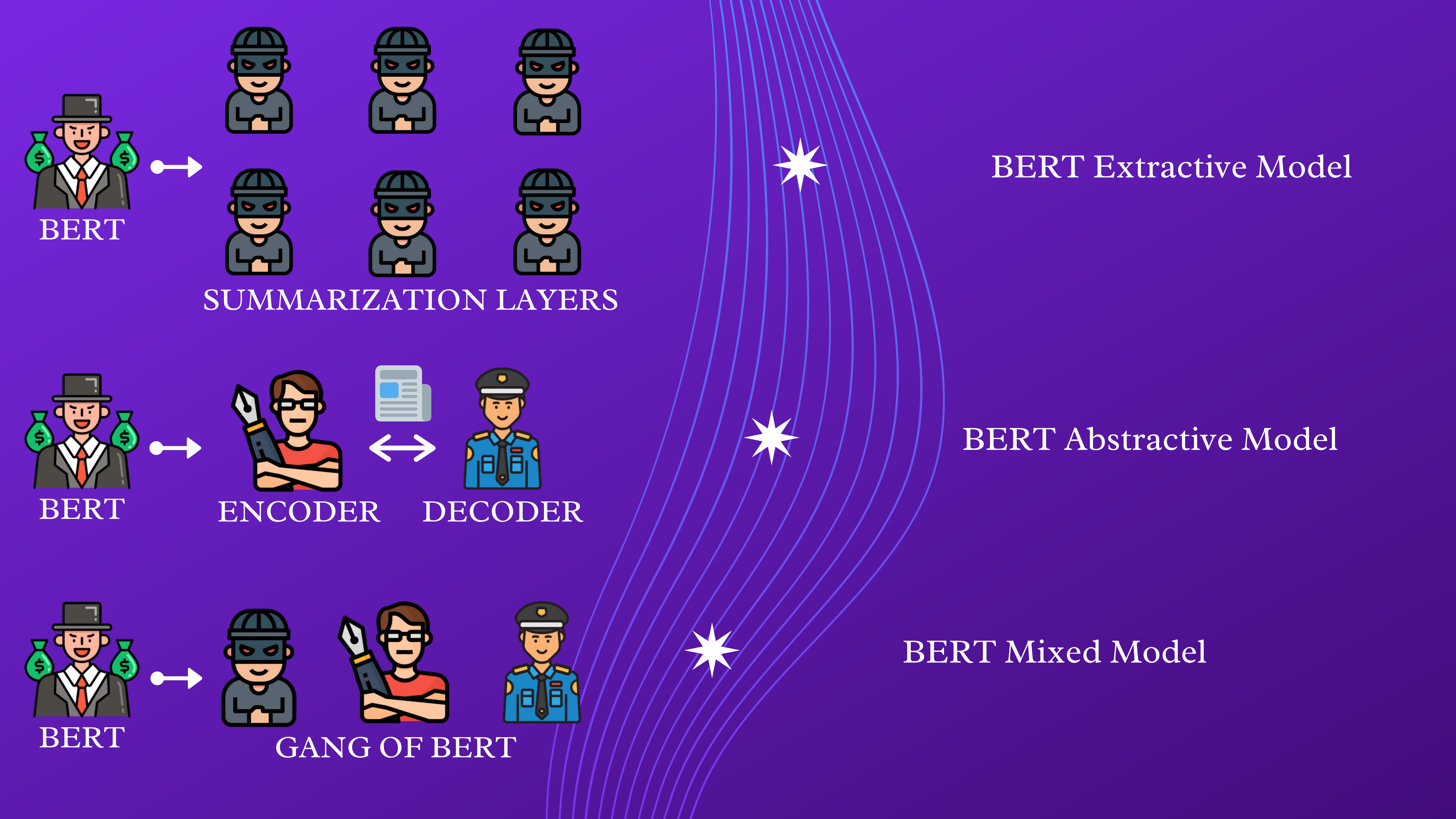
Automated News Summary with BERT

by Joshua Aje, Ole Teutloff & Dang Ngoc Huy

Mixed Summary

Abstractive Summary

Extractive Summary


BERT
SUMMARIZATION LAYERS
BERT Extractive Model
BERT
ENCODER
DECODER
BERT Abstractive Model
BERT
GANG OF BERT
BERT Mixed Model

Results

System	R-1	R-2	R-L
Extractive			
Lede-3 Baseline (by [3])	32.02	21.08	29.59
BERT (full testset)	28.74	17.76	26.15
Abstractive			
Seq2Seq + Attention (by [9])	5.99	0.37	5.41
BERT Model [4]	7.04	0.32	6.78
Mixed			
Modified P-G (by [10])	39.91	28.38	36.87
BERT Model	30.01	17.77	27

BERTSUM vs. State-of-the-art

Project Poster

Hertie School of Governance

AUTOMATED NEWS SUMMARIZATION WITH BERT

BY JOSHUA AJE, OLE TEUTLOFF & DANG NGOC HUY

MOTIVATION

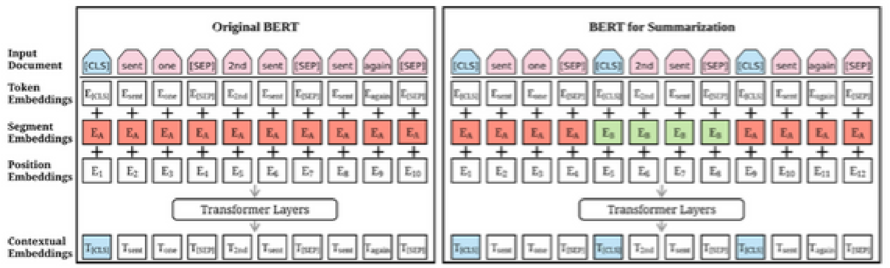
Automated text summarization is one of the central challenges of Natural Language Processing, helping to reduce complexity of documents, weeding out redundant information, decreasing the time to process large amounts of text, ultimately increasing productivity.

Our research adapts the BERTSUM model developed by Yang Liu and Mirella Lapata and apply it on the large and diverse Cornell Newsroom dataset and contribute to the literature on news summarization.

METHODOLOGY

BERTSUM utilizes pre-trained neural network to create word embeddings as features for modeling with 3 approaches:

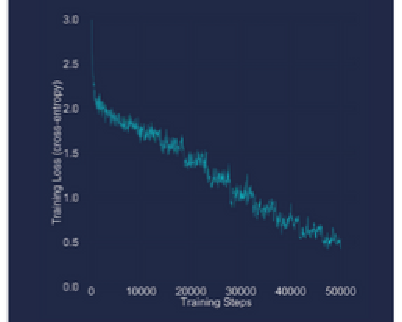
- **Extractive** (summary taken directly from source text): summarization-orientated layers are stacked and fine-tuned on top of BERT encoders.
- **Abstractive** (summary generated by the model): BERTSUM encodes input text and feed results to decoder to learn how to generate text;
- **Mixed**: weights of final extractive model are used for retraining model in abstractive setting:



RESULTS

Models are evaluated with ROUGE score, which calculates the overlapped units (n-grams, word pairs or word sequences) between the automated summaries and the gold standard (often human-generated).

All three models obtained fairly close results with state-of-the-art models for the Newsroom dataset, by training only on a fraction (7%) of the full training set (60,000 articles out of 1.3 million) and test on 12% of the full test set (20,000 articles out of 163477)



Cross-entropy loss over lifetime of training for extractive model

CONCLUSION

The success of these models presents an exciting venue of exploration in application of BERSUM for automated summary generation and can be explored much further to gauge the full extent of their utility.

System	R-1	R-2	R-L
Extractive			
Lede-3 Baseline (by [3])	32.02	21.08	29.59
BERT (full testset)	28.74	17.76	26.15
Abstractive			
Seq2Seq + Attention (by [9])	5.99	0.37	5.41
BERT Model [4]	7.04	0.32	6.78
Mixed			
Modified P-G (by [10])	39.91	28.38	36.87
BERT Model	30.01	17.77	27

Model results in comparison to SOTA