Text Summarization

1st Fehmi Ayberk Uçkun
EECS
KTH Royal Institute of Technology
Stockholm, Sweden
uckun@kth.se

2nd Fredrik Gölman
EECS
KTH Royal Institute of Technology
Stockholm, Sweden
golman@kth.se

***Abstract*—In this mini-project, we tested some of the State Of The Art (SOTA) transformer models on the sequence-to-sequence generation task of abstract text summarization using multiple datasets. Evaluations are done using ROUGE metrics. We compare the models by their architectures, pre-training and success on specific datasets under resource constraints. We present comparative examples in the end.**

***Index Terms*—Abstractive Text Summarization, Transformers, Transfer Learning, ROUGE**

I. Introduction

Text summarization is one of the most challenging tasks in Natural Language Processing (NLP). In this report we attempt to investigate how well different state-of-the-art methods fare over several different datasets. These models belong to the class of abstractive models which means that the models actually generate the summaries instead of extracting the pre-existing words in the text. The results are evaluated both measuring the loss during the training process as well as through different variations of the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics. We briefly also discuss these abstractive models in relation to the less sophisticated extractive models.

We proceed to discuss related work in section 2, primarily in the context of the actual models and datasets we use. In section 3, we describe our models, datasets and evaluation methods. Section 4, explains the implementation details of the experiments. In section 5 we present our results and then finally in section 6 a discussion with regards to our results and what conclusions that follows from them ensues.

II. Related work

The recent success of transfer learning was ignited in 2018 by GPT, ULMFiT, ELMo, and BERT, and we saw the development of a huge diversity of new methods like XLNet, RoBERTa, ALBERT, Reformer, and MT-DNN [1]. Wıth the proven promising future of the field, many other architectures are proposed for their ability to be used on downstream tasks like sequence classification, token classification, sequence generation and machine translation. Also to open up future progress in the field many different corpora are developed. The development of Huggingface Hub [2] made all of this new progress accessible by many and usable for a wide range of researchers.

The rate of progress in the field has made it difficult to evaluate which improvements are most meaningful and how effective they are when combined thus many surveys are published that evaluate proposed architectures with novel improvements on various tasks. In a similar evaluation performed on the Amazon Fine Food dataset using LSTM networks, the authors investigate how different configurations of global and local attention affect the ROUGE scores and also show some reasonable looking actual summaries [3]. The authors proceed to conclude that local attention spans yield greater ROUGE-2 scores compared to global attention which instead yields greater ROUGE-1 scores compared to local attention spans.

III. Method

A reasonable amount of research was made prior to attempting to derive actual solutions to the problem at hand. The most optimal solution was using the transformer network models that are usually provided by big companies already pre-trained by spending a considerable amount of computation resources. These pre-trained architectures, whether encoder-only, decode-only or encoder-decoder, can be used on downstream tasks by fine-tuning on much smaller datasets and requiring much fewer iterations [4]. This concept is called transfer learning. It relies on the network to learn meaningful representations of the input space using a huge amount of unlabeled data by training for a pre-text task. Before the transformers become widespread, transfer learning was a huge success in the image domain but these days it has shown to be very useful in the field of NLP, including text summarization [5] [6].

A. Models

While we do incorporate two extractive methods in our web application used for demonstration purposes in Lex Rank and Latent Semantic Analysis (LSA) our efforts in terms of evaluation have exclusively been spent on transformer models and these models have then further been fine-tuned on the specific dataset under scrutiny. We used 3 different architectures, BART (Bidirectional Auto-Regressive Transformers) [7], T-5 (Text-to-Text-Transfer-Transformer) [6], and a custom encoder-decoder architecture that uses BERT (Bidirectional Encoder Representations from Transformers) [8] for the encoder and GPT-2 (Generative Pre-trained Transformer) [9] for the decoder part.

1) BART: BART is a denoising autoencoder published by Facebook. Bart's architecture is composed of a bidirectional encoder (like BERT) and a left-to-right (autoregressive) decoder (like GPT) [7]. This is the standard sequence-to-sequence architecture used in practice. The base model we used consists of 6 encoder and 6 decoder transformer layers while the large version has 12 of them. BART-base approximately 140 million parameters. The pre-training task involves token masking, token deletion, token infilling, document rotation and sentence permutation. It has pre-trained on the same corpus as RoBERTa, which includes BOOKCORPUS, CC-NEWS, OPENWEBTEXT and STORIES.

2) T-5: T-5 is an encoder-decoder model published by Google. The version we used "T5-small" uses 6 layers on both encoder and decoder parts and it has around 60 million parameters. Much smaller than BART-base. The model was pre-trained on a multi-task mixture of unsupervised and supervised tasks. For the denoising objective C4 and Wiki-DPR datasets and for the supervised text-to-text language modelling objective 14 different datasets are used [6]. The suggested usage on downstream tasks is adding a prefix text to the input, in our case "Summarize: ".

3) BERT2GPT2: This architecture is composed of BERT encoder and GPT-2 decoder. BERT is a bidirectional transformer pre-trained using a combination of masked language modelling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia. BERT-base has 12 layers and around 110 million parameters. The version we have used is "distilBERT. DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than bert-base-uncased, and runs 60% faster while preserving over 95% of BERT's performance as claimed by the authors of [10]. GPT-2 is an autoregressive decoder architecture published by OpenAI. It has pre-trained on the next word prediction task on a very large English corpus (consisting of 8 million web pages) in a self-supervised manner. The original model has 1.5 billion parameters while our version "distilGPT-2" which is trained by the supervision of the original GPT-2 on OpenWebTextCorpus has only 82 million parameters and 6 layers.

B. Datasets

For the evaluation of these models, the following datasets have been utilized, mostly of smaller size given computational constraints; SciTLDR [11], SAMSum [12], BillSum [13], WikiLingua [14], Reddit TIFU [15] and finally the quite large Amazon Fine Foods dataset [16]. Datasets are specifically chosen from different domains in order to get a grasp on how models behave in different situations.

1) SciTLDR: SciTLDR is a scientific paper summarization dataset. It consists of 3.2K papers from the computer science field and their multiple summaries both written by authors and experts.

2) SAMSum: SAMSum contains 16k messenger-like conversations with summaries. These conversations are created by linguists reflecting the real-life messenger conversations. Conversations could be informal, semi-formal or formal, they may contain slang words, emoticons and typos.

3) BillSum: BillSum is the summarization of US Congressional and California state bills. The corpus consists of bills from 1993-2018 sessions of Congress. The texts are 5,000 to 20,000 characters in length and their summaries were written by their Legislative Counsel. It consists of around 20k bills.

4) WikiLingua: WikiLingua consists of article-summary pairs in 18 languages from WikiHow, a collaborative resource of how-to guides on a diverse set of topics. We used the English subset of this corpus which consists of 58k samples and we only used the last instructions and their summaries.

5) Reddit TIFU: Reddit TIFU consists of posts from the subreddit /r/tifu (Today I Fucked Up). It has 42k posts and their respective TLDRs as summaries.

6) Amazon Fine Foods: Amazon Fine Foods is a dataset consisting of roughly 568k reviews from 1999 to 2012 from about 260k users, and there are roughly 74k products being reviewed in total. It includes the product, the review, the title of the review, and some user information.

C. Evaluation

While the cross-entropy or negative log-likelihood loss is used to evaluate models' fine-tuning performance on labels, the actual quality of the generated summaries is better captured by ROUGE metrics. ROUGE is proposed in 2004 by Xhin-Yew Lin [17]. It includes techniques to automatically determine the quality of a summary by comparing it to reference summaries. The variants we used in this project are ROUGE-N and ROUGE-L. The n-suffix stands for some n-gram, i.e. ROUGE-1 for unigrams and ROUGE-2 for bi-grams, and l-suffix stands for the Longest Common Subsequence (LCS). While ROUGE-N tries to measure overlapping n-grams, ROUGE-L considers the sentence level structure similarity by not looking at subsequent word pairs but looking at their order in the sentence. Even though the limits of the ROUGE metrics are known [18] like the impossibility of 100% score for good quality datasets, it is still the most common automatic measure for these types of tasks.

IV. Implementation Details

All base models are downloaded from Huggingface [2]. For BART, "Facebook/bart-base", for T-5 "t5-small", For BERT "distilbert-base-uncased" and for GPT-2 "distilgpt2" models are used. The small versions of the models are chosen because of the lack of computational resources but even the small/distilled version of these models are proven to achieve good results [10]. All models were trained with 512 maximum input token sizes and 128 output token sizes. Longer inputs and outputs are truncated. For BART and T-5 models, default configurations are used. For the custom BERT2GPT2 architecture, the same configurations (token ids, vocab size etc.) of the BERT tokenizer are applied. For the decoding $num\_beams = 4, no\_repeat\_ngram\_size = 3, early\_stopping = True$ are set.

For the training process, we used an adaptive optimizer AdamW with 0.01 weight decay and generally configured our training parameters with a start learning rate of $5e-5$ and a batch size of 32. The numbers of training epochs have varied among the different datasets, a bit depending on performance, but this has generally been a predefined number in the range between 5 and 15 epochs. The exception is being the Amazon Fine Food dataset, which was very large and the number of epochs did not exceed 1. The models are evaluated based on train and test loss as well as ROUGE metrics.

The experiment environment was Python 3.8.10 and included PyTorch 1.11.0+cu113 and Transformers 4.20.0.dev0 packages. All training except the Amazon Fine Food dataset training was done with NVIDIA GeForce RTX 3070 Laptop GPU. The Amazon Fine Food dataset was trained on Google Colab GPUs with Python 3.7.13, Transformers 4.19.2 and PyTorch 1.11.0+cu113.

Further numerical details are presented in section 5 when we present our results.

V. Results

This section will be split into subsections for each dataset on which models have been evaluated. For the general impression of the evaluations please refer to the Table I.

A. SciTLDR

Since the BERT2GPT2 model is a merged version of separate encoder and decoders, the model doesn't have any pre-trained cross-attention layers, thus, in fine-tuning, they are randomly initialized and require more training compared to other models to give good results. In this dataset, while other models are trained for 10 epochs, the BERT2GPT2 model was trained for 15 epochs but still left quite behind in scores. Since we had limited computational resources, we couldn't train the model until it reaches its best performance but the progress of the training can be inspected in Figure 1. Since this architecture would require the same higher computation cost, we didn't test it on further datasets.

For BART and T-5 performances are similar but BART is better by a small margin. Additionally, in the original paper of T-5, the suggested usage of the model by adding a prefix text that is describing the task of the model to the input to the model. In our case, this prefix text is ”Summarize: ”. We run the fine-tuning with and without this prefix and observe a small increase in the performance when the prefix is added, see Figure 2.

SciTLDR is a multi-target dataset and it is possible to find the best matching summary for the generated text but in our experiment we only used the first available reference summary. Also, both only ”Abstract” and ”Abstract-Introduction-Conclusion” versions of the full text are available and we observed slightly better performance, see Figure 3, when we trained with the ”AIC” version even though the total length of the text is longer than the maximum accepted length of 512 and most of the ”IC” parts of the text are truncated.

The outputs of the 3 model are compared in Table V.

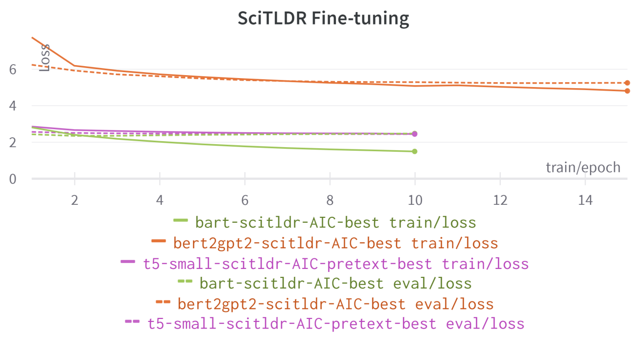

Fig. 1. Fine-tuning of different models on SciTLDR dataset.

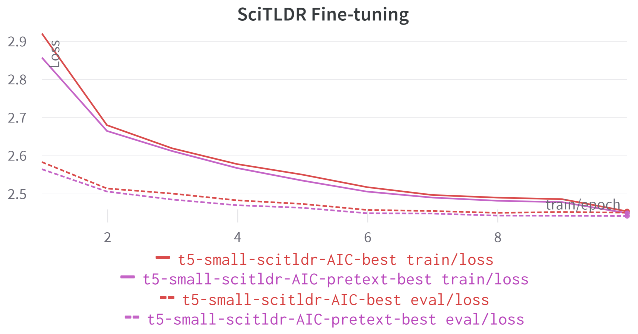

Fig. 2. Performance of T-5 model with and without prefix-text that explains the task at hand.

B. SAMSum

For this dataset, we fine-tuned BART and T-5 models. Again, BART model achieved better results but this time with a higher margin. The most successful results of the whole evaluation are achieved in SAMSum dataset for the BART model. One visible difference in the training can be seen in Figure 4. Even though both models' test errors converge to the approximately same point, T-5's training error remains a lot higher compared to BART's.

Some outputs of the model compared in Table VI.

C. BillSum

The rest of the datasets are only used to train the BART model. In Table II, all metrics of the training are shown in detail. For this dataset a clear high precision is distinct compared to other dataset results. Since this dataset includes state bills and those bills generally are very formal and structured we are good at predicting the right words but we are not good at generating novel words that are specific to that bill.

Some outputs of the model compared in Table VII.

D. WikiLingua

As a general trend among whole datasets, we can see that the R2 score is generally lower but R1 and RL are higher. We can deduce that models are good at predicting

	SciTLDR			SAMSum			BillSum			WikiLingua			Reddit TIFU			Amazon Fine Food		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
BART	35.17	16.48	30.60	**48.00**	**24.23**	**40.25**	23.78	19.43	23.11	31.11	13.30	26.54	25.38	7.99	20.72	18.69	7.27	18.40
T-5	34.10	15.27	29.46	42.34	18.92	35.19	-	-	-	-	-	-	-	-	-	-	-	-
BERT2GPT2	20.47	4.22	17.49	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

The numbers presented are the f1-scores.

TABLE I
MODELS AND THEIR EVALUATIONS ON THE DATASETS.

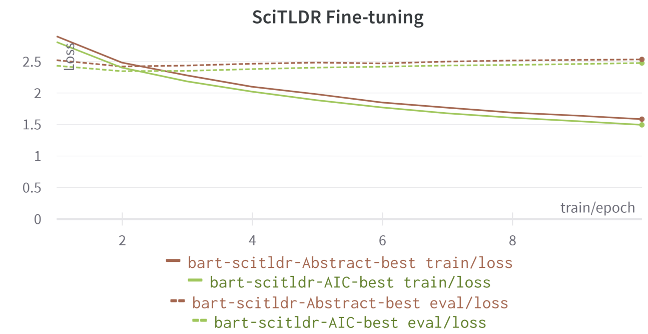

Fig. 3. Fine-tuning of BART model with only Abstract and with "AIC" version of SciTLDR dataset.

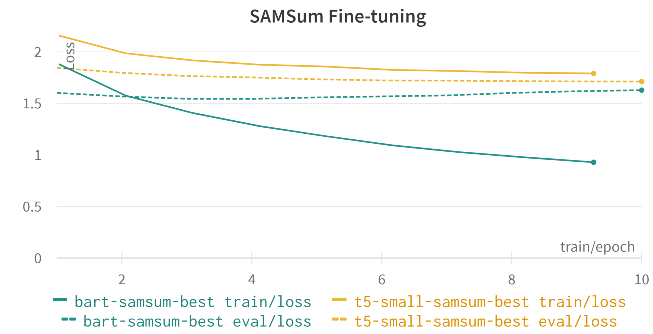

Fig. 4. Fine-tuning of T-5 and BART on SAMSum dataset.

the correct words and predicting them in the correct order that is understanding the sentence structure but words are not necessarily generated in pairs with their neighbour words.

Some outputs of the model compared in Table VIII.

E. Reddit TIFU

One remarkable outcome of this evaluation was seeing that the model is pretty family friendly. As a characteristic of this dataset, many of texts include inappropriate language. However, our model managed to stay clear of these words but still generated sensible meaning.

Some outputs of the model compared in Table IX.

F. Amazon Fine Food

The Amazon Fine Food dataset was so large that we opted to only evaluate it on one model, the BART model. The dataset

	R1			R2			RL		
	precision	recall	f1	precision	recall	f1	precision	recall	f1
BART	89.23	14.11	23.78	77.63	11.42	19.43	86.96	13.70	23.11

TABLE II
ALL METRICS FOR BART MODEL ON BILLSUM.

	R1			R2			RL		
	precision	recall	f1	precision	recall	f1	precision	recall	f1
BART	53.44	23.98	31.10	23.47	10.22	13.30	45.18	20.58	26.53

TABLE III
ALL METRICS FOR BART MODEL ON WIKILINGUA.

was pre-processed only in the manner to replace missing label values with empty strings. The model was fine-tuned with a start learning rate of $2e - 5$ and was otherwise using the parameters mentioned previously in the subsection 3.4. It is a bit curious that the validation loss is lower than the training loss, but one can observe that the training loss is decreasing quicker whereas the validation loss seems to stagnate, and would presumably start to increase, as can be seen in Figure 5. As would be the case with nearly all datasets, the ROUGE-1 and ROUGE-l scores are higher than the ROUGE-2 scores as shown in Figure 6. That ROUGE-2 is lower than ROUGE-1 is hardly a surprise for obvious reasons. Given the nature of ROUGE-l and the reasonably high score this model attains suggests that it may manage to summarize a text in a reasonably fluent manner compared to the ground truth.

Some sample summaries were presented in [3]. These samples along with our model's summaries can be seen in X. They will be referred to as *LSTM G*, *LSTM L5*, *LSTM L7*, where *G* stands for Global attention, and *L5* and *L7* for Local attention with window size 5 and 7 respectively. *BART* and *BART constrained* refers to our fine-tuned model where constrained refers to a configuration we use, with a minimum length of 40 and a maximum length of 150 as well as a penalty for longer outputs, when generating summaries to provide more expressive summaries.

As can be seen in our sample outputs, when the model is forced to output some minimum length it may start to reiterate over similar expressions, but they are still mostly very reasonable.

	R1			R2			RL		
	precision	recall	f1	precision	recall	f1	precision	recall	f1
BART	32.76	23.18	25.38	10.52	7.21	7.99	26.48	19.14	20.72

TABLE IV
ALL METRICS FOR BART MODEL ON REDDIT TIFU.

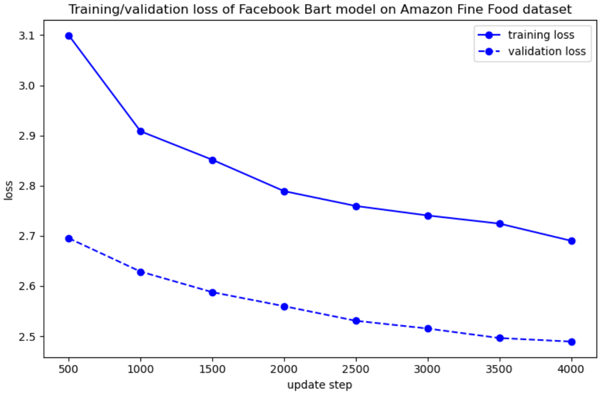

Fig. 5. Training process of Facebook Bart model on the Amazon Fine Food dataset

VI. CONCLUSION

As has been previously mentioned computational resources has been an issue, but we also had another limiting factor in the form of available time to perform the experiments. These factors have imposed constraints on the scope of our evaluation. Nevertheless, we believe these experiments have successfully showed results of text summaries of high quality. As was brought up in section 2 in a similar evaluation on the Amazon Fine Food dataset using LSTM networks [3], and further directly comparing outputs of their model and our model in section 4F, we can observe that our ROUGE scores are more than three times higher for ROUGE-1 and for ROUGE-2 roughly a hundred times higher. Naturally, these scores are not comparable straight up for several reasons. For one, they perform stricter pre-processing in that they filter any data point whose text does not span between 25 and 300 tokens.

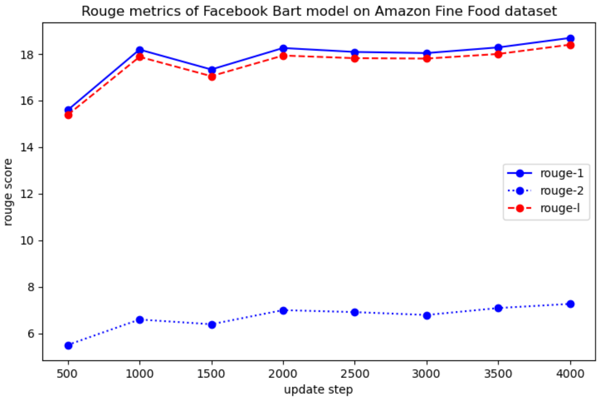

Fig. 6. Rouge scores during training on the Amazon Fine Food dataset

In the end we found focusing on a few different datasets and models was most important to our evaluation. Briefly discussing our results between different datasets we can see that the ROUGE scores varies quite a bit. For instance, one could argue that the Amazon Fine Food dataset is not an ideal one to perform summaries. The reviews may often be reasonably coherent, but at times they may be a mess and a combination of misspelled words, poor grammar, slang and other less desirable things. The labels may be even worse as they may sometimes not be representative of the actual review, and sometimes even missing entirely. On the other hand, some other datasets contain actual articles combined with proper summaries, and one would have to think are generally easier for the model to handle. We ultimately conclude that, despite occasionally generating poor summaries for some inputs, the models largely have success in summarizing our textual input. This despite the fact that a relatively limited amount of fine-tuning was done.

Lastly we saw that for some summaries even though their ROUGE metrics suffers from low scores, they are actually quite representative of the original text. One problem in this domain is that nearly none of summaries written by different people %100 matches and we shouldn't expect this to be the case for the models either, thus, evaluating automatic summarizations systems is not a straightforward task.

VII. FUTURE WORK

The effect of pre-processing on model performance is one aspect that clearly could be further investigated, but is seemingly something that could have a positive effect on the model's performance. There can also be significant differences with regards to the vocabulary and how well the input is encoded into numerical vector representations. It goes without saying that the Facebook BART model used in our evaluations, which is not even the largest one published, most certainly has been extensively pre-trained. Still, it is our opinion that this big discrepancy in ROUGE scores reflects the power of transfer learning and utilization of these big pre-trained transformer networks for various NLP tasks. Other areas that are worth further exploration is hyper-parameter tuning such as start learning rate, weight decay, and perhaps optimizer. One thing we observed first hand during our experiments and which is extensively discussed by Xiong et al. [19] in what they refer to as ”warm-up learning rate” is that these transformer networks require significantly lower start learning rate than many other training processes, otherwise it becomes unstable. Lower start learning rate subsequently results in slower convergence, and that is something they also discuss and suggest a less computationally heavy alternative to which yields similar final performance results. While the Adam optimizer suffers from poor generalization capability [1] which has then been addressed in the AdamW optimizer we are using, exploring the results of other optimizers such as Stochastic Gradient Descent (SGD), both with regards to training time and model performance, would be interesting.

References

[1] N. S. Keskar and R. Socher, "Improving generalization performance by switching from adam to sgd," *arXiv preprint arXiv:1712.07628*, 2017.

[2] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. V. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing." [Online]. Available: https://github.com/huggingface/

[3] P. M. Hanunggul and S. Suyanto, "The impact of local attention in lstm for abstractive text summarization," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2019, pp. 54–57.

[4] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds. Cham: Springer International Publishing, 2018, pp. 270–279.

[5] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 15–18. [Online]. Available: https://aclanthology.org/N19-5004

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension."

[8] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "Bert: Pre-training of deep bidirectional transformers for language understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners." [Online]. Available: https://github.com/codelucas/newspaper

[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter." [Online]. Available: https://github.com/huggingface/transformers

[11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 2020-December, 6 2020. [Online]. Available: https://arxiv.org/abs/2006.11477v3

[12] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "Samsum corpus: A human-annotated dialogue dataset for abstractive summarization," 2019.

[13] A. Kornilova and V. Eidelman, "Billsum: A corpus for automatic summarization of us legislation." [Online]. Available: http://leginfo.legislature.ca.gov

[14] F. Ladhak, E. Durmus, C. Cardie, and K. Mckeown, "Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization." [Online]. Available: https://www.wikihow.com

[15] B. Kim, H. Kim, and G. Kim, "Abstractive summarization of reddit posts with multi-level memory networks." [Online]. Available: http://vision.snu.ac.kr/projects/reddit-tifu

[16] Amazon fine food reviews — kaggle. [Online]. Available: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

[17] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[18] N. Schluter, "The limits of automatic summarisation according to ROUGE," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 41–45. [Online]. Available: https://aclanthology.org/E17-2007

[19] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 10 524–10 533. [Online]. Available: https://proceedings.mlr.press/v119/xiong20b.html

Appendix

A. GitHub

Our implementations can be found at our GitHub repository.

Model	Summary
Source	FearNet is a memory efficient neural-network, inspired by memory formation in the mammalian brain, that is capable of incremental class learning without catastrophic forgetting.
BART	We propose FearNet , a brain-inspired model that does not store previous examples , making it memory efficient.
T-5	Incremental class learning involves sequentially learning classes in bursts of examples from the same class . This violates the assumptions that underlie methods for training standard deep neural networks , and will cause them to suffer from catastrophic forgetting .
BERT2GPT2	we propose a new method for learning method that can be used to solve the training of learning.
Model	Summary
Source	Multi-view learning improves unsupervised sentence representation learning
BART	Multi-view learning can provide self-supervision when different views are available of the same data
T-5	Multi-view learning can provide self-supervision when different views are available of the same data . Distributional hypothesis provides another form of useful self-overpower from adjacent sentences which are plentiful in large unlabelled corpora .
BERT2GPT2	we propose an unsupervised multi - task learning framework for unsu alignvised learning

TABLE V
COMPARISON BETWEEN BART, T5 AND BERT2GPT2 ON SCITLDR

Model	Summary
Source	Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.
BART	Betty called Larry last time they were at the park together. Amanda can't find Betty's number. Hannah suggests Amanda texted Larry.
T5	Larry called Betty last time she was at the park together. He's very nice, but she doesn't know him well. Amanda will text him.
Model	Summary
Source	Emma will be home soon and she will let Will know.
BART	Emma will be home soon and she will tell Will when she gets home.
T5	Emma will be home soon. Will will pick her up when she gets home. Emma is not hungry, but she's not worried about cooking.

TABLE VI
SUMMARY EXAMPLES FOR BART AND T-5 ON SAMSUM

Model	Summary (300+ words)
Source	Amends the Water Resources Development Act of 1999 to: (1) authorize appropriations for FY 1999 through 2009 for implementation of a long-term resource monitoring program with respect to the Upper Mississippi River Environmental Management Program (currently, such funding is designated for a program for the planning, construction, and evaluation of measures for fish and wildlife habitat rehabilitation and enhancement); (2) (...) (8) authorize and provide for an authorization of appropriations for the existing program for the safety and operations expenses of the Federal Railroad Administration, and make available for obligation funds currently appropriated for such program.
BART	Amends the Water Resources Development Act of 1992 to authorize the Secretary of the Interior to make grants to: (1) Jackson County, Mississippi; (2) Manchester, New Hampshire; (3) Paterson, Passaic County, and Passaic Valley, New Jersey; and (4) the North Hudson Sewerage Authority for the elimination or control of combined sewer overflows.
Model	Summary
Source	Prescription Drug Monitoring Act of 2016 This bill requires a state that receives grant funds under the prescription drug monitoring program (PDMP) or the controlled substance monitoring program to comply with specified requirements. The Department of Justice (DOJ) or Department of Health and Human Services may withhold grant funds from a state that fails to comply. Additionally, the bill requires a state to share its PDMP data with other states through a data-sharing hub established by DOJ.
BART	Prescription Drug Monitoring Act of 2016 This bill requires each state to require each prescribing practitioner within the state or their designee, who shall be licensed or registered healthcare professionals or other employees who report directly to the practitioner before initiating treatment with a prescription for a controlled substance listed in schedule II, III, or IV of the Controlled Substances Act (CSA) and every three months thereafter as long as the treatment continues, to provide proactive notification to a practitioner when patterns indicative of controlled substance misuse, including opioid misuse, are detected, and to report to the PDMP within 24 hours after the controlled substance is dispensed to the patient

TABLE VII
SUMMARY EXAMPLES FOR BART ON BILLSUM

Model	Summary
Source	Click the Services tab. Locate a service you want to disable. Un-check the service. Click OK. Click Restart when prompted.
BART	Click the System Configuration tab. Select a service. Click the checkbox next to the service you want to disable. Click Restart.
Model	Summary
Source	Get some sheets of paper in different colors. Finished.
BART	Find a piece of paper. Cut the paper in half lengthwise. Cut out the colors of the rainbow. Finished..

TABLE VIII
SUMMARY EXAMPLES FOR BART ON WIKILINGUA

Model	Summary
Source	booked an appointment at a physio place and got a ride from my mom, walked in to see three women wearing see through shirts and sexy lingerie - mom thinks i'm a perv now.
BART	tried to get a late morning appointment at a physio/chiropractor , ended up getting caught in the middle of the act
Model	Summary
Source	tried to impress professor with historical reference, ended up looking like a jerkoff, whole class hates me now
BART	made a joke about immigrants eating babies , professor thought i was trying to disrupt class .

TABLE IX
SUMMARY EXAMPLES FOR BART ON REDDIT TIFU

Model	Summary (or source)
Source text 1	My wife bought me this fruit cake this past Holiday Season, never have I had a better fruit cake, I am in my late 50's and I have had many thru the years. If you never had a good fruit cake try this one, but you have to order early in the season or you may miss out, only so many to go round.
Source summary	Best Cake Ever
LSTM G	Great !
LSTM L5	Best ! ! ! ! ! ! ! ! ! ! ! ! ! ! !
LSTM L7	Great !
BART	Best Fruit Cake
BART constrained	Best fruit cake I have ever had, and I love this one. I love the fruit cake. My wife loves the fruit cakes. I love it. This is the best.
Model	Summary (or source)
Source text 2	This is the only tea in my house. I love its peachy taste and the ginger is very subtle. I have a cup every evening to put a soothing touch on my day. I got the pack of 6 so I would never run out! Haven't tried the other ones those are next but wanted to make sure I had enough of this one.
Source summary	Love it!
LSTM G	I
LSTM L5	yummy!!!!!!!!!!!!!!!
LSTM L7	Delicious!
BART	The only tea in my house
BART constrained	The only tea in my house that I have a cup of tea to put a soothing touch on my day. I love this tea. I love it. Love this one. Love it!
Model	Summary (or source)
Source text 1	Keurig is amazing and Green Mountain coffee is just as amazing. This is coffee that will open your eyes in the morning, as well as provide a welcome break during the day. It's not bitter, its smooooooooooooth and mellow. No aftertaste. Love it!!!!!!!!!!!
Source summary	Smoooooooooth
LSTM G	<UNK>
LSTM L5	Great!!!!!!!!!!!
LSTM L7	<UNK>
BART	Green Mountain Coffee
BART constrained	Green Mountain coffee is the best!! The best! The BEST! I love Green Mountain coffee. I LOVE Green Mountain! This is the BEST!! the BEST!

TABLE X
COMPARISON BETWEEN OUR BART MODEL AND ANOTHER SIMILAR EXPERIMENT WITH AN LSTM BASED NETWORK