The Effect of Pretraining on Extractive Summarization for Scientific Documents

Yash Gupta[1] **Pawan Sasanka Ammanamanchi**[2] **Shikha Bordia**[3] **Arjun Manoharan**[3]
Deepak Mittal[3] **Ramakanth Pasunuru**[4] **Manish Shrivastava**[2]
Maneesh Singh[3] **Mohit Bansal**[4] **Preethi Jyothi**[1*]
[1]Indian Institute of Technology, Bombay, [2]International Institute of Information Technology, Hyderabad
[3]Verisk Analytics ,[4]University of North Carolina, Chapel Hill

Abstract

Large pretrained models have seen enormous success in extractive summarization tasks. We investigate, here, the influence of pretraining on a BERT-based extractive summarization system for scientific documents. We derive performance improvements using an intermediate pretraining step that leverages existing summarization datasets and report state-of-the-art results on a recently released scientific summarization dataset, SciTLDR. We systematically analyze the intermediate pretraining step by varying the size and domain of the pretraining corpus, changing the length of the input sequence in the target task and varying target tasks. We also investigate how intermediate pretraining interacts with contextualized word embeddings trained on different domains.

1 Introduction

Text summarization is a quintessential NLP task that involves generating a coherent and succinct summary of an article containing the most salient information from the original article. Summarization systems are particularly useful for scientific articles that tend to be long and rich in technical content. Summarization can arguably reduce information overload on researchers and facilitate the quick retrieval of relevant papers from vast amounts of scientific literature. Broadly, summarization techniques can be categorized as extractive or abstractive. While abstractive systems treat the summarization problem as a natural language generation task and produce new phrases and sentences directly in the summary, extractive techniques select salient phrases or sentences verbatim from the original document to create a summary. Maynez et al. (2020), Kryscinski et al. (2020), Huang et al. (2020) report factual hallucinations in abstractive summarization. Durmus et al. (2020) highlight the trade-off between faithfulness and abstractiveness. Since for the scientific summarization task, it is critical to be factually-accurate and be faithful to the source document, we focus on extractive summarization of scientific articles.

Large pretrained language models (e.g. BERT (Devlin et al., 2019)) have been successfully used for many NLP tasks including summarization (Liu and Lapata, 2019), using the following, now widely-adopted, two-step approach:

Pretraining. Start with a pretrained model like BERT and suitably adapt its architecture to fit the target task.

Finetuning. Finetune the model using a labeled dataset for the target task.

Recent work shows the benefits of interspersing the pretraining and finetuning steps with an intermediate pretraining step (Phang et al., 2018),(Vu et al., 2020). This intermediate step often involves supervised pretraining using labeled datasets from different domains for a task that is related to or is the same as the target task. While the efficacy of such pretraining approaches have been studied in prior work for natural language understanding tasks (like entailment, question answering, etc. (Vu et al., 2020)), the effect of pretraining on summarization has been far less explored.

In this work, we explore the benefits of intermediate pretraining using existing summarization datasets for a target task involving the summarization of scientific articles. We obtain improvements in performance over state-of-the-art extractive summarization baseline systems on a new sci-

*Correspondence to pjyothi@cse.iitb.ac.in

Proceedings of the Second Workshop on Scholarly Document Processing, pages 73–82
June 10, 2021.

entific summarization benchmark, SciTLDR (Cachola et al., 2020). We also make the following key observations:

- Intermediate pretraining using labeled summarization datasets (even when containing articles that are very different in domain from scientific articles) is very beneficial to low-resource target tasks like SciTLDR. We also derive additional benefits by filtering the intermediate pretraining data to only retain a subset of articles (based on a similarity metric) that best matches the target task.
- While starting with a BERT-based model pretrained on scientific articles (e.g., SciBERT(Beltagy et al., 2019)) offers a small advantage compared to the standard BERT-based model as an initialization, this advantage is eclipsed by the effect of intermediate pretraining which is much more significant.
- The benefits from intermediate pretraining diminish with access to sufficiently large amounts of finetuning data in the target task. We also observe a trend of diminishing returns with the intermediate pretraining, as we increase the amount of pretraining data.

2 Related Work

Transfer Learning Pretrained language models like BERT(Devlin et al., 2019) are trained on self-supervised training objectives over large amount of unlabelled text corpus. As shown in (Phang et al., 2018), (Zhang and Bowman, 2018), (Phang et al., 2020), the pretrained knowledge in these models can be leveraged by domain and task adaptive pretraining before finetuning the model to the desired target task. Gururangan et al. (2020), Chakrabarty et al. (2019), Beltagy et al. (2019) finetune language models on the domains of interest and show improvements on the respective in-domain tasks.

Summarization Recent works in summarization MatchSum (Zhong et al., 2020), BERTSUM (Liu and Lapata, 2019), STEPwise ETCSum (Narayan et al., 2020) use pretrained language models. BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020) use variants of self-supervised training objectives on massive amounts of text corpora and compute to achieve stellar performance on summarization tasks. While most of the recent works focus on improving state-of-the-art results on news datasets like CNN/DailyMail and XSum, improving summarization on scientific documents is an overlooked area.

Intermediate Pretraining Howard and Ruder (2018) first introduced the idea of intermediate pretraining in NLP and its benefits on the 6 tasks of classification. The benefits of this in summarization has been shown by Yu et al. (2021) where in they finetune BART (Lewis et al., 2020) on XSUM (Narayan et al., 2018) and show its results on low resource domain adaptation benchmark for summarization . We show the effects of intermediate pretraining in the context of scientific document summarization.

Scientific Summarization Cachola et al. (2020) introduce the SciTLDR task and benchmark a variety of summarization models such as MatchSum and BERTSUM on the task. Impressive results were reported by Pilault et al. (2020), Zaheer et al. (2020) on scientific datasets like Pubmed, arXiv using compute-intensive transformer based models. We report results on Pubmed and SciTLDR where our models use significantly less compute and achieve superior results on SciTLDR over BERTSUM and MatchSum.

Cross task learning Lebanoff et al. (2018), Mao et al. (2020) use methods that adapt single document summarization task to multi document summarization setup, namely using the CNN/Daily Mail (CNN/DM) dataset. While it is similar to the idea of the intermediate finetuning used in this paper, the end task is different and are tested over a different set of metrics. Zhong et al. (2019) conducts some experiments with supervised pretrained knowledge transfer, we do extensive experiments in the context of scientific summarization.

3 Base Model

In this paper, we base our experiments on the BERTSUM (Liu and Lapata, 2019) architecture that uses BERT embeddings and formulates extractive summarization as a sentence classification problem. The intermediate pretraining step uses data from a summarization task that is different from the target task and could also be from a different domain. We also experiment with replacing pretrained BERT embeddings with SciBERT embeddings (Beltagy et al., 2019).

BERTSUM Model We use the extractive model proposed by Liu and Lapata (2019) as our base model. It uses a BERT-based encoder (Devlin et al.,

2019) to obtain sentence level representations of a document using the [CLS] token at the beginning of each sentence. Several transformer layers are stacked to represent the discourse. These transformer layers are jointly fine-tuned with BERT on a sentence classification task with a sigmoid layer as the final output predicting whether or not each sentence in the input document should be in the summary. The loss of the model is a cross-entropy loss for binary classification.

Using SciBERT Embeddings Beltagy et al. (2019) finetune BERT-Base on scientific documents from the biomedical and computer science domains. To leverage the stylistic variation and adapt to domain knowledge specific to scientific articles, we examine the effects of replacing BERT embeddings in the BERTSUM model with SciBERT embeddings.

4 Experimental Setup

4.1 Summarization Datasets

We evaluate the models on two scientific summarization benchmark datasets— Pubmed (Cohan et al., 2018) and SciTLDR (Cachola et al., 2020). We use the CNN/DM (Hermann et al., 2015) dataset for intermediate pretraining.

SciTLDR. SciTLDR is a curated corpus containing computer science articles, with each article having one or more reference TLDR's or one-sentence summaries. The inputs could either be abstract-only (SciTLDR-A) or the abstract, introduction and conclusion sections of the article (SciTLDR-AIC). We present results for both settings and use the splits specified in (Cachola et al., 2020).

Pubmed. The Pubmed dataset consists of scientific articles from PubMed.org. We used the splits and preprocessing steps from (Zhong et al., 2020), wherein the introduction is used as the article and the abstract is used as the summary.

CNN/DM. The CNN/DM dataset consists of news articles and highlights from *CNN* and *Daily Mail* news articles, on diverse topics including sports, health, business, etc. The standard splits are used for training, validation and testing without anonymizing the entities. Appendix A contains more detailed statistics about all the three datasets used in this work.

For intermediate pretraining, we also experiment with a subset of articles from Pubmed and CNN/DM together (henceforth referred to as Mixed) that are most similar to our target tasks, SciTLDR-A and SciTLDR-AIC (Guo et al., 2020). We derive BERT-base embeddings for each Pubmed and CNN/DM article via [CLS] tokens. Then, we select 83K articles (roughly 35K and 48K articles from Pubmed and CNN/DM, respectively) with the smallest averaged L2 distance between embeddings of the Pubmed/CNN/DM articles and the SciTLDR target tasks.[1]

4.2 Models and Implementation Details

Our extractive summarization system uses the BERT-based architecture by (Liu and Lapata, 2019) described in Section 3. For intermediate pretraining, we use one of CNN/DM, Pubmed or Mixed. The finetuning step involves data from one of three target tasks, SciTLDR-A, SciTLDR-AIC and Pubmed. For all training steps, we set the dropout rate to 0.1 and learning rate to 2e-3, which are the reported parameters in (Liu and Lapata, 2019) for CNN/DM. We use a batch size of 3000 for all experiments involving CNN/DM during pretraining. The best model is selected on the basis of validation Rouge scores for one-line summaries on the validation set. This is done to select the model with the best "extreme" summarization capability. When evaluating on Pubmed, the number of sentences extracted is set to 6, as reported in (Zhong et al., 2020). For fine-tuning on SciTLDR-A as well as SciTLDR-AIC, the batch size is set to 100 and the number of extracted sentences to form the final summary is 1.

Evaluation Metrics. The SciTLDR tasks have multiple reference summaries for each test article. We compute Rouge scores between the summary generated by our system and each of the reference summaries. We consider the reference with the maximum Rouge-1 score as the main gold summary used in further evaluations. We choose Rouge-1 (R1), Rouge-2 (R2) and Rouge-L (RL) as our main evaluation metrics, as is typically done for summarization tasks. To determine the best possible performance from an extractive summarization system, we also compute oracle scores by choosing a sentence from each test article with the highest R1 score across all reference summaries and averaging these scores across the test articles.

[1]We select 83K articles in Mixed, which is the size of Pubmed, to examine the effect of varying pretraining corpora of a fixed size.

	SciTldr-A			SciTldr-AIC		
	R1	R2	RL	R1	R2	RL
Oracle	49.2	26.0	39.9	53.7	29.9	43.9
MatchSum† (BERT-base)	42.7	20.0	34.0	38.6	16.4	30.1
	Our Models					
Pretraining Datasets	**Using BERT**					
-	39.71	18.91	32.63	36.99	16.14	29.64
Pubmed (83K)	41.49	19.57	33.40	40.82	18.98	32.84
CNN/DM (83K)	41.69	19.55	33.44	41.93	20.10	33.95
Mixed (83K)	**42.32**	**20.50**	**34.30**	**42.78**	**21.06**	**34.83**
CNN/DM (Full)	42.26	20.32	34.09	42.21	20.24	34.19
	Using SciBERT					
-	39.93	18.50	32.32	37.16	15.94	29.65
CNN/DM (83K)	40.60	19.04	32.93	40.74	19.09	32.95
Pubmed (83K)	41.10	19.33	32.87	40.61	18.69	32.68
CNN/DM (Full)	40.66	19.08	32.59	41.25	19.40	33.37

Table 1: Max Rouge scores for SciTldr on test sets. † Results from (Cachola et al., 2020)

	Pubmed		
	R1	R2	RL
Oracle	45.12	20.33	40.19
MatchSum† (BERT-base)	41.21	14.91	36.75
	Our Models		
Pretraining Datasets	**Using BERT**		
-	40.65	14.85	36.18
CNN/DM (Full)	40.77	14.92	36.29
	Using SciBERT		
-	**41.08**	**15.16**	**36.59**
CNN/DM (Full)	40.59	14.76	36.12

Table 2: Mean Rouge scores for Pubmed test sets.† Results from (Zhong et al., 2020)

Dataset Size	R1	R2	RL
83K articles	41.93	20.10	33.95
176K articles	**42.27**	**20.37**	**34.32**
286K articles	42.21	20.24	34.19

Table 3: Results by varying the size of the pretraining dataset CNN/DM while finetuning on SciTldr-AIC.

Input Length	SciTldr-AIC R1	SciTldr-AIC R2	SciTldr-AIC RL	Pubmed R1	Pubmed R2	Pubmed RL
512	42.21	20.24	34.19	40.65	14.85	36.18
1024	42.21	20.34	34.35	42.44	16.39	37.86
1500	**42.23**	**20.65**	**34.41**	**42.65**	**16.59**	**38.03**

Table 4: Results by varying the input sequence length while finetuning. The pretraining dataset is CNN/DM for SciTldr-AIC and none for Pubmed.

5 Results and Discussion

Table 1 and Table 2 show our main results. In the first two rows, we present results from the state-of-the-art MatchSum system (Zhong et al., 2020) and oracle scores. The remaining rows show pretraining results using BERT and SciBERT embeddings in the BERTSUM model. Without any intermediate pretraining, SciBERT offers a small advantage over BERT on Pubmed and is statistically comparable to BERT on both SciTldr tasks. With pretraining and using BERT, we observe significant improvements in performance regardless of the pretraining corpora used. (We significantly outperform MatchSum on SciTldr-AIC.) With keeping the size of the pretraining corpus fixed at 83K articles, pretraining with Mixed gives the best results showing that it is beneficial to selectively choose articles in the pretraining corpus that best match the target tasks. Unlike for the low-resource SciTldr target tasks, intermediate pretraining does not benefit Pubmed showing that its effect diminishes when sufficient amounts of finetuning data are available for the target task.

With pretraining and replacing BERT with SciBERT, we observe a deterioration in performance indicated by the drop in Rouge scores (especially with CNN/DM). The SciBERT initialization appears to be counterproductive when using CNN/DM during intermediate pretraining. It is more beneficial to start with BERT, rather than SciBERT, and pretrain on CNN/DM before the final finetuning step.

Additionally, we undertake two ablation experiments. 1) We investigate the effect of varying amounts of pretraining data. We vary the size of CNN/DM to 83K, 176K and 286K articles and analyse the finetuning results on SciTldr-AIC with BERT embeddings. As shown in Table 3, R1, R2 and RL scores increase on moving from 83K to 176K articles but performance stagnates with a further increase in the size of the pretraining corpus. 2) During finetuning, we experiment with truncating the input sequence lengths of SciTLDR-AIC and Pubmed at 512, 1024 and 1500 tokens, as shown in Table 4. We initialize the model with BERT embeddings for the first 512 tokens and repeat the last set of weights for the remaining input tokens. We observe that the ROUGE scores improve with longer input lengths, with a sizeable boost for Pubmed.

6 Conclusions and Future Work

In this paper, we present a systematic investigation of the benefits of transfer learning via pretraining for extractive summarization of scientific articles. We show improvements in ROUGE scores for the SciTLDR benchmark using an intermediate pretraining that uses existing summarization datasets. We obtain additional benefits by filtering these existing datasets to construct a pretraining corpus that best matches the target task. This suggests the need for further explorations in future work on different criteria to be used for selective pretraining and how it could benefit both extractive and abstractive summarization.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. *ArXiv*, abs/2001.04362.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1737–1751, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Jakub Adamek, Daniele Pighin, Blaz Bratanic, and Ryan McDonald. 2020. Stepwise extractive summarization and planning with structured transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4143–4159, Online. Association for Computational Linguistics.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across nlp tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. Adaptsum: Towards low-resource domain adaptation for abstractive summarization.

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 sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Dataset Details

Corpus ($\mathcal{C}$)	$\lvert\mathcal{C}\rvert$ Train	Val	Test	Source of $\mathcal{C}$	# of Tokens Doc.	Sum.
SciTldr-A	1992 papers (1992 TLDRs)	619 papers (1452 TLDRs)	618 papers (1967 TLDRs)	OpenReview API	159	21
SciTLDR-AIC	1992 papers (1992 TLDRs)	619 papers (1452 TLDRs)	618 papers (1967 TLDRs)	OpenReview API	993	21
CNN/DM (Full) CNN/DM (83k)	287k 83k	-	-	News Articles	685	53
Pubmed	83233	4946	5025	Biomedical Literature	444	209

Table 5: Dataset details of summarization datasets. Unlike the other datasets, SCITLDR consists of multiple reference summaries for each article. SCITLDR-AIC has the highest compression ratio when compared to the other datasets.

A.1.1 SCITLDR

This dataset is built from a combination of TLDRS written by human experts and author-written TLDRS of computer science papers from OpenReview. OpenReview (https://openreview.net/) is one such example where authors are asked to submit TLDRs of their papers, which communicates to both reviewers and users of OpenReview the main content of the paper. SCITLDR has multiple reference summaries for each of the test and validation articles. The additional reference summaries (apart from the author written one) were obtained from human annotators. This is an "extreme" summarisation task as the compression ratio is very high compared to the other datasets i.e. around 47 for the AIC task. While the dataset is inherently abstractive in nature, the extractive oracle scores listed in Table **??** are quite high (in fact, they are much higher than existing abstractive and extractive SoTA scores), which implies there is a lot of scope for extractive summarisation.

A.1.2 CNN/DM

This dataset contains online news articles paired with multi-sentence summaries (which are highlights of the news articles). The dataset is fairly large and also has a high extractive oracle (with ROUGE-1 / ROUGE-2 / ROUGE-L scores of 52.59 / 31.24 / 48.87), although the summaries are not inherently extractive. The compression ratio is much lower compared to SCITLDR i.e. around 13.

A.1.3 Pubmed

This dataset is collected from scientific papers. It has a very low compression ratio i.e. around 2 (which is a direct consequence of using the introduction section as the document and the abstract as the corresponding summary). The summaries are relatively long, compared to SCITLDR and CNN/DM, with around 6 sentences per summary.

A.2 Qualitative Analysis

We present examples of SCITLDR articles and generated summaries to illustrate the effects of pretraining and other design choices (such as varying input lengths and BERT/SCIBERT initializations).

A.2.1 Effect of Input Sequence Length on SciTLDR-AIC

Article 1
Good representations facilitate transfer learning and few-shot learning. Motivated by theories of language and communication that explain why communities with large number of speakers have, on average, simpler languages with more regularity, we cast the representation learning problem in terms of learning to communicate.Our starting point sees traditional autoencoders as a single encoder with a fixed decoder partner that must learn to communicate. Generalizing from there, we introduce community-based autoencoders in which multiple encoders and decoders collectively learn representations by being randomly paired up on successive training iterations. Our experiments show that increasing community sizes reduce idiosyncrasies in the learned codes, resulting in more invariant representations with increased reusability and structure. The importance of representation learning lies in two dimensions. First and foremost, representation learning is a crucial building block of a neural model being trained to perform well on a particular task, i.e., representation learning that induces the "right" manifold structure can lead to models that generalize better, and even extrapolate. Another property of representation learning, and arguably the most important one, is that it can facilitate transfer of knowledge across different tasks , essential for transfer learning and few-shot learning among others BID0 . With this second point in mind, we can define good representations as the ones that are reusable, induce the abstractions that capture the "right" type of invariances and can allow for generalizing very quickly to a new task. Significant efforts have been made to learn representations with these properties; one frequently explored direction involves trying to learn disentangled representations BID12 BID6 BID5 BID17), while others focus on general regularization methods BID15 BID18 . In this work, we take a different approach to representation learning, inspired by successful abstraction mechanisms found in nature, to wit human language and communication.Human languages and their properties are greatly affected by the size of their linguistic community BID11 BID19 BID16 BID9 .....
Ground Truth Summaries
Motivated by theories of language and communication, we introduce community-based autoencoders, in which multiple encoders and decoders collectively learn structured and reusable representations. The authors tackle the problem of representation learning, aim to build reusable and structured represenation, argue co-adaptation between encoder and decoder in traditional AE yields poor representation, and introduce community based auto-encoders. The paper presents a community based autoencoder framework to address co-adaptation of encoders and decoders and aims at constructing better representations.
Input Length 512 (ROUGE-1: 18.18, ROUGE-2: 0.00, ROUGE-L: 12.12)
Good representations facilitate transfer learning and few-shot learning .
Input Length 1024 (ROUGE-1: 28.57, ROUGE-2: 0.00, ROUGE-L: 14.29)
Our starting point sees traditional autoencoders as a single encoder with a fixed decoder partner that must learn to communicate.
Input Length 1500 (ROUGE-1: 60.0, ROUGE-2: 49.99, ROUGE-L: 55.99)
Generalizing from there, we introduce community-based autoencoders in which multiple encoders and decoders collectively learn representations by being randomly paired up on successive training iterations.
Article 2
Generative models are important tools to capture and investigate the properties of complex empirical data. Recent developments such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) use two very similar, but *reverse*, deep convolutional architectures, one to generate and one to extract information from data. Does learning the parameters of both architectures obey the same rules? .We exploit the causality principle of independence of mechanisms to quantify how the weights of successive layers adapt to each other. Using the recently introduced Spectral Independence Criterion, we quantify the dependencies between the kernels of successive convolutional layers and show that those are more independent for the generative process than for information extraction, in line with results from the field of causal inference. In addition, our experiments on generation of human faces suggest that more independence between successive layers of generators results in improved performance of these architectures. Deep generative models have proven powerful in learning to design realistic images in a variety of complex domains (handwritten digits, human faces, interior scenes). In particular, two approaches have recently emerged: Generative Adversarial Networks (GANs) BID8 , which train an image generator by having it fool a discriminator that should tell apart real from artificially generated images; and Variational Autoencoders (VAEs) BID15 BID21) that learn both a mapping from latent variables to the data (the decoder) and the converse mapping from the data to the latent variables (the encoder), such that correspondences between latent variables and data features can be easily investigated.....
Ground Truth Summaries
We use causal inference to characterise the architecture of generative models . This paper examines the nature of convolutional filters in the encoder and a decoder of a VAE, and a generator and a discriminator of a GAN. This work exploits the causality principle to quantify how the weights of successive layers adapt to each other.
Input Length 512 (ROUGE-1: 25.92, ROUGE-2: 3.84, ROUGE-L: 14.81)
Using the recently introduced Spectral Independence Criterion, we quantify the dependencies between the kernels of successive convolutional layers and show that those are more independent for the generative process than for information extraction, in line with results from the field of causal inference.
Input Length 1024 (ROUGE-1: 38.46, ROUGE-2: 8.33, ROUGE-L: 23.08)
Generative models are important tools to capture and investigate the properties of complex empirical data.
Input Length 1500 (ROUGE-1: 82.05, ROUGE-2: 75.68, ROUGE-L: 82.05)
We exploit the causality principle of independence of mechanisms to quantify how the weights of successive layers adapt to each other.

Table 6: Articles 1 and 2 have 1068 and 1265 tokens, respectively. We see that increasing the length of the input sequence significantly improves the ROUGE scores.

A.2.2 Effect of Pretraining on SciTLDR-AIC

Article 1
Recent advances in neural Sequence-to-Sequence (Seq2Seq) models reveal a purely data-driven approach to the response generation task. Despite its diverse variants and applications, the existing Seq2Seq models are prone to producing short and generic replies, which blocks such neural network architectures from being utilized in practical open-domain response generation tasks. In this research, we analyze this critical issue from the perspective of the optimization goal of models and the specific characteristics of human-to-human conversational corpora. Our analysis is conducted by decomposing the goal of Neural Response Generation (NRG) into the optimizations of word selection and ordering. It can be derived from the decomposing that Seq2Seq based NRG models naturally tend to select common words to compose responses, and ignore the semantic of queries in word ordering. On the basis of the analysis, we propose a max-marginal ranking regularization term to avoid Seq2Seq models from producing the generic and uninformative responses. The empirical experiments on benchmarks with several metrics have validated our analysis and proposed methodology. Past years have witnessed the dramatic progress on the application of generative sequential models (also noted as seq2seq learning (Sutskever et Despite these promising results, current Sequence-to-Sequence (Seq2Seq) architectures for response generation are still far from steadily generating relevant and coherent replies. The essential issue identified by many studies is the Universal Replies: the model tends to generate short and general replies which contain limited information, such as "That's great!", "I don't know", etc. Nevertheless, most previous analysis over the issue are empirical and lack of statistical evidence. Therefore, in this paper, we conduct an in-depth investigation on the performance of seq2seq models on the NRG task....
Ground Truth Summaries
Analyze the reason for neural response generative models preferring universal replies; Propose a method to avoid it. Investigates the problem of universal replies plaguing the Seq2Seq neural generation models. The paper looks into improving the neural response generation task by deemphasizing the common responses using modification of the loss function and presentation the common/universal responses during the training phase.
Pubmed (ROUGE-1: 20.51, ROUGE-2: 0.00, ROUGE-L: 20.51)
In this research, we analyze this critical issue from the perspective of the optimization goal of models and the specific characteristics of human-to-human conversational corpora.
CNN/DM (ROUGE-1: 34.62, ROUGE-2: 8.00, ROUGE-L: 26.92)
Our analysis is conducted by decomposing the goal of Neural Response Generation (NRG) into the optimizations of word selection and ordering.
CNN/DM+Pubmed (ROUGE-1: 37.50, ROUGE-2: 0.0 , ROUGE-L: 31.25)
Therefore, in this paper, we conduct an in-depth investigation on the performance of seq2seq models on the NRG task.

Article 2
Graph convolutional networks (GCNs) have been widely used for classifying graph nodes in the semi-supervised setting. Previous works have shown that GCNs are vulnerable to the perturbation on adjacency and feature matrices of existing nodes. However, it is unrealistic to change the connections of existing nodes in many applications, such as existing users in social networks. In this paper, we investigate methods attacking GCNs by adding fake nodes. A greedy algorithm is proposed to generate adjacency and feature matrices of fake nodes, aiming to minimize the classification accuracy on the existing ones. In additional, we introduce a discriminator to classify fake nodes from real nodes, and propose a Greedy-GAN algorithm to simultaneously update the discriminator and the attacker, to make fake nodes indistinguishable to the real ones....
Ground Truth Summaries
non-targeted and targeted attack on GCN by adding fake nodes The authors propose a new adversarial technique to add "fake" nodes to fool a GCN-based classifier
Pubmed (ROUGE-1: 23.53, ROUGE-2: 0.0, ROUGE-L: 11.76)
Graph convolutional networks (GCNs) have been widely used for classifying graph nodes in the semi-supervised setting.
CNN/DM (ROUGE-1: 34.15, ROUGE-2: 5.13, ROUGE-L: 24.39)
A greedy algorithm is proposed to generate adjacency and feature matrices of fake nodes, aiming to minimize the classification accuracy on the existing ones.
CNN/DM+Pubmed (ROUGE-1: 52.17, ROUGE-2: 38.09, ROUGE-L: 52.17)
In this paper, we investigate methods attacking GCNs by adding fake nodes.

Table 7: For both the articles, we note an increasing trend in the ROUGE scores with pretraining on Pubmed, CNN/DM and MIXED (i.e., CNN/DM+Pubmed).

A.2.3 Bert vs SciBERT without pretraining on SciTLDR-AIC

Article 1
In this paper, we introduce a system called GamePad that can be used to explore the application of machine learning methods to theorem proving in the Coq proof assistant. Interactive theorem provers such as Coq enable users to construct machine-checkable proofs in a step-by-step manner. Hence, they provide an opportunity to explore theorem proving with human supervision. We use GamePad to synthesize proofs for a simple algebraic rewrite problem and train baseline models for a formalization of the Feit-Thompson theorem. We address position evaluation (i.e., predict the number of proof steps left) and tactic prediction (i.e., predict the next proof step) tasks, which arise naturally in tactic-based theorem proving. Theorem proving is a challenging AI task that involves symbolic reasoning (e.g., SMT solvers BID2) and intuition guided search. Recent work BID7 Loos et al., 2017; has shown the promise of applying deep learning techniques in this domain, primarily on tasks useful for automated theorem provers (e.g., premise selection) which operate with little to no human supervision. In this work, we aim to move closer to learning on proofs constructed with human supervision.We look at theorem proving in the realm of formal proofs. A formal proof is systematically derived in a formal system, which makes it possible to algorithmically (i.e., with a computer) check these proofs for correctness....
Ground Truth Summaries
We introduce a system called GamePad to explore the application of machine learning methods to theorem proving in the Coq proof assistant. This paper describes a system for applying machine learning to interactive theorem proving, focuses on tasks of tactic prediction and position evaluation, and shows that a neural model outperforms an SVM on both tasks. Proposes that machine learning techniques be used to help build proof in the theorem prover Coq.
Bert Output (ROUGE-1: 34.78, ROUGE-2: 4.55, ROUGE-L: 21.74)
We use GamePad to synthesize proofs for a simple algebraic rewrite problem and train baseline models for a formalization of the Feit-Thompson theorem.
SciBERT Output (ROUGE-1: 86.27, ROUGE-2: 81.63, ROUGE-L: 86.27)
In this paper, we introduce a system called GamePad that can be used to explore the application of machine learning methods to theorem proving in the Coq proof assistant.

Article 2
We propose a novel method that makes use of deep neural networks and gradient decent to perform automated design on complex real world engineering tasks. Our approach works by training a neural network to mimic the fitness function of a design optimization task and then, using the differential nature of the neural network, perform gradient decent to maximize the fitness. We demonstrate this methods effectiveness by designing an optimized heat sink and both 2D and 3D airfoils that maximize the lift drag ratio under steady state flow conditions. We highlight that our method has two distinct benefits over other automated design approaches. First, evaluating the neural networks prediction of fitness can be orders of magnitude faster then simulating the system of interest. Second, using gradient decent allows the design space to be searched much more efficiently then other gradient free methods. These two strengths work together to overcome some of the current shortcomings of automated design. Automated Design is the process by which an object is designed by a computer to meet or maximize some measurable objective. This is typically performed by modeling the system and then exploring the space of designs to maximize some desired property whether that be an automotive car styling with low drag or power and cost efficient magnetic bearings BID1 BID4 . A notable historic example of this is the 2006 NASA ST5 spacecraft antenna designed by an evolutionary algorithm to create the best radiation pattern (Hornby et al.) . More recently, an extremely compact broadband on-chip wavelength demultiplexer was design to split electromagnetic waves with different frequencies BID17 . While there have been some significant successes in this field the dream of true automated is still far from realized. The main challenges present are heavy computational requirements for accurately modeling the physical system under investigation and often exponentially large search spaces. These two problems negatively complement each other making the computation requirements intractable for even simple problems.Our approach works to solve the current problems of automated design in two ways. First, we learn a computationally efficient representation of the physical system on a neural network. This trained network can be used to evaluate the quality or fitness of the design several orders of magnitude faster. Second, we use the differentiable nature of the trained network to get a gradient on the parameter space when performing optimization. This allows significantly more efficient optimization requiring far fewer iterations then other gradient free methods such as genetic algorithms or simulated annealing....
Ground Truth Summaries
A method for performing automated design on real world objects such as heat sinks and wing airfoils that makes use of neural networks and gradient descent. Neural network (parameterization and prediction) and gradient descent (back propogation) to automatically design for engineering tasks. This paper introduces using a deep network to approximate the behavior of a complex physical system, and then design optimal devices by optimizing this network with respect to its inputs.
Bert Output (ROUGE-1: 16.67, ROUGE-2: 4.35, ROUGE-L: 12.49)
This allows significantly more efficient optimization requiring far fewer iterations then other gradient free methods such as genetic algorithms or simulated annealing.
SciBERT Output (ROUGE-1: 62.75, ROUGE-2: 40.82, ROUGE-L: 39.22)
We propose a novel method that makes use of deep neural networks and gradient descent to perform automated design on complex real world engineering tasks.

Table 8: For both the articles, we observe clear improvements in ROUGE scores with using SciBERT as opposed to BERT.