Day 124: NLP Papers Summary – TLDR: Extreme Summarization Of Scientific Documents

By Ryan 3rd May 2020 No Comments

Objective and Contribution

Introduced the TLDR generation task and SCITLDR, a new extreme summarisation dataset, where researchers can use to train models to generate TLDR for scientific papers. Introduced an annotation protocol for creating different ground-truth summaries using peer review comments, allowing us to scale our dataset and for the first time, there are multiple summaries link to a single source document. Lastly, we proposed a multi-task training strategy that's d on

TLDR and title generation to adapt our pre-trained language model BART. This has sho to outperform extractive and abstractive baselines.

Introduction to TLDR Generation Task

The TLDR generation task aims to generate TLDRs that leave out background or methodology details and focus more on key aspects such as the contributions of the paper. This requires the model to have background knowledge as well as the ability to understand domain-specific language. Figure below showcase an example of the TLDR task as well as a list of categories of the type of information that appears in TLDR.

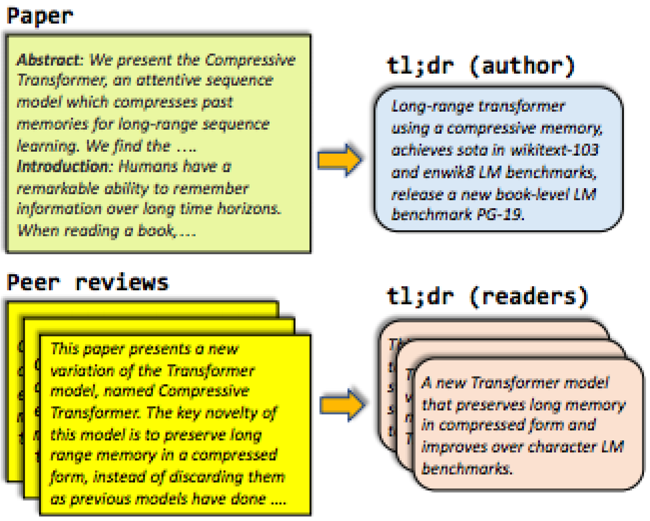
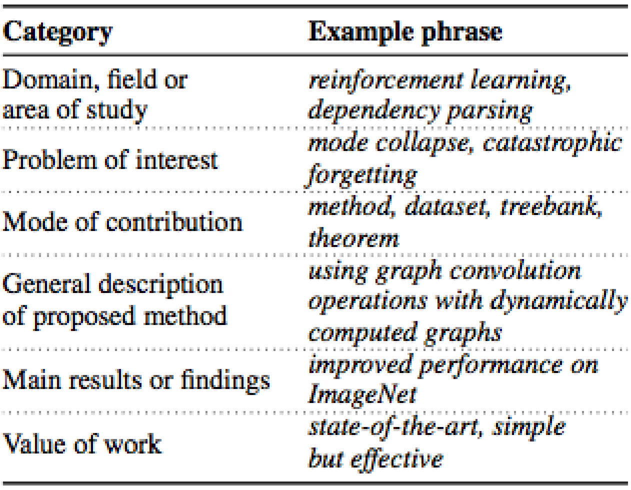

Paper
Abstract: We present the Compressive Transformer, an attentive sequence model which compresses past memories for long-range sequence learning. We find the ....
Introduction: Humans have a remarkable ability to remember information over long time horizons. When reading a book, ...
tl;dr (author)
Long-range transformer using a compressive memory, achieves sota in wikitext-103 and enwik8 LM benchmarks, release a new book-level LM benchmark PG-19.
Peer reviews
This paper presents a new variation of the Transformer model, named Compressive Transformer. The key novelty of this model is to preserve long range memory in a compressed form, instead of discarding them as previous models have done ....
tl;dr (readers)
A new Transformer model that preserves long memory in compressed form and improves over character LM benchmarks.

Figure 1: A TLDR is an extreme summary of a scientific paper. Our SCITLDR dataset includes multiple TLDRs for each paper. One written by the authors of papers and others are manually rewritten from peer reviews.

SCITLDR Dataset

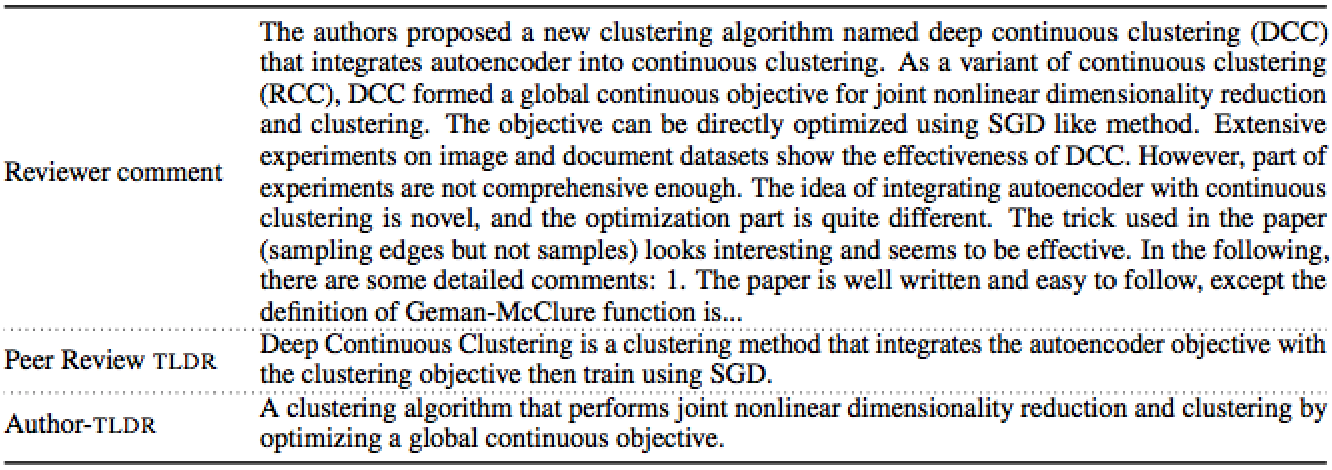
SCITLDR has 3935 TLDRs in computer science scientific documents. SCITLDR includes TLDRs written by both the original author of the paper and peer reviews. However, the key difference here is that authors and peer reviews are writing TLDR based on reviewer comments and not the original research paper. This method assumes readers to have a good background



Category	Example phrase
Domain, field or area of study	*reinforcement learning, dependency parsing*
Problem of interest	*mode collapse, catastrophic forgetting*
Mode of contribution	*method, dataset, treebank, theorem*
General description of proposed method	*using graph convolution operations with dynamically computed graphs*
Main results or findings	*improved performance on ImageNet*
Value of work	*state-of-the-art, simple but effective*

Table 1: Example categories of informati TLDR might contain

knowledge to follow the general research areas and so our TLDRs can leave out common concepts. In addition, the reviewer comments are written by experts in the field and so they are high quality summaries. Figure below showcase an example of the annotation process.



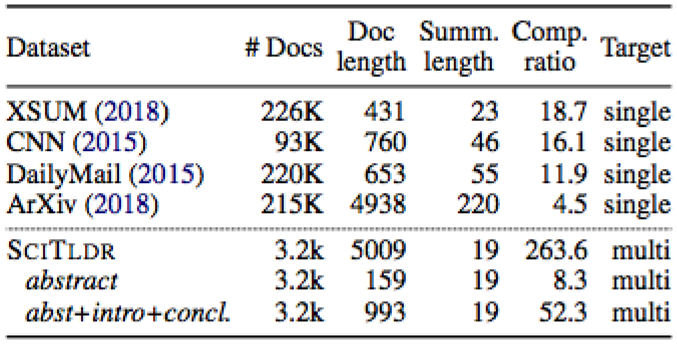
Reviewer comment	The authors proposed a new clustering algorithm named deep continuous clustering (DCC) that integrates autoencoder into continuous clustering. As a variant of continuous clustering (RCC), DCC formed a global continuous objective for joint nonlinear dimensionality reduction and clustering. The objective can be directly optimized using SGD like method. Extensive experiments on image and document datasets show the effectiveness of DCC. However, part of experiments are not comprehensive enough. The idea of integrating autoencoder with continuous clustering is novel, and the optimization part is quite different. The trick used in the paper (sampling edges but not samples) looks interesting and seems to be effective. In the following, there are some detailed comments: 1. The paper is well written and easy to follow, except the definition of Geman-McClure function is...
Peer Review TLDR	Deep Continuous Clustering is a clustering method that integrates the autoencoder objective with the clustering objective then train using SGD.
Author-TLDR	A clustering algorithm that performs joint nonlinear dimensionality reduction and clustering by optimizing a global continuous objective.

Table 2: Example of a reviewer's comment rewritten as a TL;DR

One of the uniqueness of SCITLDR is that each paper in the test set is map to multiple ground-truth TLDRs, one written by the original author and the rest by peer reviews. This would a) allow us to better evaluate our generated summaries as there are now multiple ground-truth summaries to compute ROUGE scores for, and b) having both the author and reader's TLDR allows us to capture the variation in summaries based on the reader's perspective.

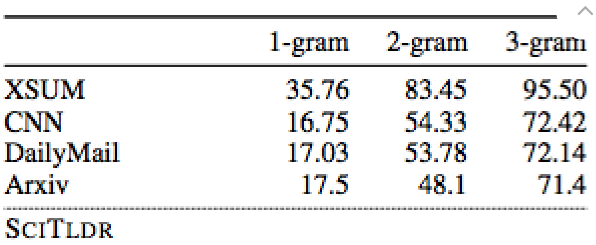
DATASET ANALYSIS

First of, SCITLDR is a much smaller dataset, with only 3.2K papers due to manual data collection and annotations. Secondly, SCITLDR has an extremely high compression ratio compared to other datasets. The average document length is 5009 and it's being compressed into an average summary length of 19. This makes the summarisation very challenging. Table 3 showcase these summary statistics. SCITLDR has at least two ground-truth TLDRs for each paper in the test set and so we investigate the ROUGE score difference between different ground-truth TLDRs. There is a low ROUGEE-1 overlap (27.40) between author-generated TLDRs and PR-generated TLDRs. Author-generated TLDRs has a ROUGE-1 of 34.1 with the title of the paper. PR-generated TLDRs only has ROUGE-1 of 24.7. This showcase the importance of multiple ground-truth TLDRs in summarisation as one source document could have multiple relevant summaries.



Dataset	# Docs	Doc length	Summ. length	Comp. ratio	Target
XSUM (2018)	226K	431	23	18.7	single
CNN (2015)	93K	760	46	16.1	single
DailyMail (2015)	220K	653	55	11.9	single
ArXiv (2018)	215K	4938	220	4.5	single
SciTldr	3.2k	5009	19	263.6	multi
abstract	3.2k	159	19	8.3	multi
abst+intro+concl.	3.2k	993	19	52.3	multi

Table 3: Comparison of datasets. The lengths are in tokens. Target summary can be single (only one gold summary per document) or multi (multiple gold summaries for each document). Compression ratio shows the ratio of document length to summary length.



	1-gram	2-gram	3-gram
XSUM	35.76	83.45	95.50
CNN	16.75	54.33	72.42
DailyMail	17.03	53.78	72.14
Arxiv	17.5	48.1	71.4
SciTldr			

Experimental Setup and Results

MODEL TRAINING

We finetuned BART model to generate TLDR. However, there are few limitations. First of, the size of our training data. We have a small dataset for training neural networks. This has led us to collect additional 20K paper-title pairs from arXiv and up sampling our SCITLDR to match the new volume. The reason we are collecting titles is because it often contains important information about the paper and we believe if we train the model to perform title generation too, it will learn how to select important information from the paper. With the new information, we are ready to train our model. First, we train BART-large model on XSUM dataset, which it's an extreme summarisation dataset on general news domain. Then, we would finetune our BART model on our SCITLDR and title dataset.

The second limitation we face is that BART has a limitation on input length and so we put BART under two settings: BART_abstract (SCITLDR_Abst) and BART_abstract_intro_conclusion (SCITLDR_AIC). Those are the different inputs used to generate title/TLDR. Existing works have shown that the most important information in a research paper is in the abstract, introduction, and conclusion.

MODELS COMPARISON

1. *Extractive models*. PACSUM (unsupervised extension of TextRank) and BERTSUMEXT (supervised)

2. *Abstractive models*. Different variations of BART

We used the ROUGE metric for evaluation. We would compute the ROUGE score for each ground-truth TLDRs and select the maximum.

RESULTS

The extractive oracle provides an upper bound performance. In table 6, we can see a continuous increase in ROUGE scores as the input space increases. Specifically, there are a 5 ROUGE score improvement when including introduction and conclusion as input, showcasing their importance in generating a useful summary. Although there are ROUGE score improvement from AIC to full text, the improvement is not big suggesting that the value added of other sections in the paper are not as high as AIC.

In table 5, we can see that BART finetuned on the original SCITLDR is enough to outperformed the other extractive and abstractive baselines. Further improvement is shown when pretraining BART on XSUM, however, this improvement only applies to SCITLDR_AIC. Our multitask learning strategy has outperformed all the baseline models and achieved further improvement on top of BART + XSUM. This showcase the value added of training the model for both tit nd TLDR generation. Figure below showcase a qualitative example of summaries generated by different models.

Conclusion and Future Work

Potential future work could make use of the information from the whole paper, capturing more context. In addition, we could explicitly model the background knowledge of the reader, creating TLDRs based on who the reader is. Lastly, we could apply our annotation process to other datasets and convert any peer review comments to TLDRs summaries.

Source: https://arxiv.org/pdf/2004.15011.pdf

Ryan

Data Scientist

Previous Post

Next Post

Day 123: NLP Papers Summary - Context-aware Embedding for Targeted Aspect-based Sentiment Analysis

Day 125: NLP Papers Summary - A2N: Attending to Neighbors for Knowledge Graph Inference

Data Science Natural Language Processing NLP Papers Summary

Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding

Ryan

30th December 2020

30th December 2020

Data Science Implementation Natural Language Processing

Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

Ryan
29th December 2020

Data Science Ryan's PhD Journey

Day 363: Ryan’s PhD Journey – Literature Review – Knowledge Acquisition – 1st Passes XIV

Ryan
28th December 2020