CiteSum: Citation Text-guided Scientific Extreme Summarization and Domain Adaptation with Limited Supervision

Yuning Mao, Ming Zhong, Jiawei Han
University of Illinois Urbana-Champaign
{yuningm2, mingz5, hanj}@illinois.edu

Abstract

Scientific extreme summarization (TLDR) aims to form ultra-short summaries of scientific papers. Previous efforts on curating scientific TLDR datasets failed to scale up due to the heavy human annotation and domain expertise required. In this paper, we propose a simple yet effective approach to automatically extracting TLDR summaries for scientific papers from their citation texts. Based on the proposed approach, we create a new benchmark CiteSum without human annotation, which is around 30 times larger than the previous human-curated dataset SciTLDR. We conduct a comprehensive analysis of CiteSum, examining its data characteristics and establishing strong baselines. We further demonstrate the usefulness of CiteSum by adapting models pre-trained on CiteSum (named CITES) to new tasks and domains with limited supervision. For scientific extreme summarization, CITES outperforms most fully-supervised methods on SciTLDR without any fine-tuning and obtains state-of-the-art results with only 128 examples. For news extreme summarization, CITES achieves significant gains on XSum over its base model (not pre-trained on CiteSum), *e.g.*, +7.2 ROUGE-1 zero-shot performance and state-of-the-art few-shot performance. For news headline generation, CITES performs the best among unsupervised and zero-shot methods on Gigaword.[1]

1 Introduction

Scientific summarization typically regards paper abstract as the ground-truth summary, as it is written by the authors themselves with relatively high quality and readily available in most scientific documents. However, paper abstract may not always be the ideal summary because it often involves certain details such as task description, background information, and experiment results (cf. the abstract of this paper). As a result, recent work (Cachola et al.,

---

Paper Abstract: We study the problem of transferring a sample in one domain to an analog sample in another domain. Given two related domains, S and T, we would like to learn a generative function G that maps an input sample from S to the domain T, such that the output of a given function f, which accepts inputs in either domains, would remain unchanged. Other than the function f, the training data is unsupervised and consist of a set of samples from each domain. The Domain Transfer Network (DTN) we present employs a compound loss function that includes a multiclass GAN loss, an f-constancy component, and a regularizing component that encourages G to map samples from T to themselves. We apply our method to visual domains including digits and face images and demonstrate its ability to generate convincing novel images of previously unseen entities, while preserving their identity.

Citation Text: Taigman et al. [8] proposed the Domain Transfer Network (DTN) to map a sample from one domain to an analog sample in another domain and achieved favorable performance on small resolution face and digit images.

Table 1: An example showing that the citation texts of a paper can often be used as its ultra-short summary.

2020) has studied the problem of scientific extreme summarization, which aims at forming ultra-short summaries (usually one sentence) of the papers, namely the TLDR[2] summaries.

However, unlike paper abstracts, ultra-short paper summaries are far from being universally available. Only certain scientific venues such as OpenReview.net support a *TLDR* field during paper submission, which is completely optional, and not all submitted papers provide such information. In addition, human-annotated summaries of scientific documents are rather costly and require domain expertise. As a consequence, the previous SciTLDR dataset (Cachola et al., 2020), using a combination of author-provided TLDR and human-annotated TLDR (rephrased from paper reviews on OpenRe-

---

[1]Our dataset and code can be found at https://github.com/morningmoni/CiteSum.

[2]"TLDR" (or "TL;DR") is short for "too long; didn't read", often used in online discussions about scientific papers.

view), only collected around 2,000 examples for training and 600 for testing.

In this paper, we argue that citation texts can often serve as high-quality short summaries of the cited papers. In Table 1, we show the abstract of one paper and its citation sentence in a follow-up paper. We observe that the citation sentence introduces the cited method and its contributions in a concise and accurate manner. Motivated by such observations, we propose a simple yet effective approach to locating, extracting, and filtering citation texts from scientific papers. We then treat the processed citation texts as ground-truth summaries of the cited papers. Based on the proposed approach, we create a large-scale scientific extreme summarization benchmark, CiteSum, which is automatically derived from citation texts and around 30 times larger than the previous human-annotated dataset SciTLDR (Cachola et al., 2020).

We conduct a comprehensive analysis of CiteSum regarding its data characteristics and quality, meanwhile establishing strong baselines as the reference for future studies. We further verify the usefulness of CiteSum by demonstrating that models pre-trained on CiteSum, which we name as CITES (**C**itation **T**ext-guided **S**ummarizer), exhibit superior generalizability during low-resource adaptation to new tasks and domains.

On the human-annotated scientific extreme summarization dataset SciTLDR (Cachola et al., 2020), our zero-shot BART-based (Lewis et al., 2020) CITES, without any fine-tuning, performs better than most *fully-supervised* baselines, including the fully-supervised BART model (without pre-training on CiteSum). Our few-shot CITES achieves state-of-the-art performance with only 128 labeled examples from SciTLDR. In addition, CITES outperforms its base model (BART) on two more diverse scientific tasks – discipline classification and title generation. When transferring to news extreme summarization, despite the domain mismatch, CITES achieves significantly better zero-shot performance than BART and PEGA-SUS (Zhang et al., 2020) (*e.g.*, +7.2 ROUGE-1) and state-of-the-art few-shot performance on the XSum dataset (Narayan et al., 2018). Furthermore, CITES performs the best among unsupervised and zero-shot methods on the Gigaword news headline generation dataset (Rush et al., 2015).

Contributions. (1) We propose a simple yet effective approach to automatically extracting ultra-

---

Citation Example 1: We take the publicly available Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020) as the source for data creation.

Citation Example 2: Unlike WikiTransfer (Fabbri et al., 2021), CITES does not involve any downstream task-specific data selection or model tuning.

Table 2: Examples (in our paper) showing that citation texts have different intents and cannot *always* be used as summaries of the cited paper.

short paper summaries from citation texts. (2) Based on the proposed approach, we create a large-scale scientific extreme summarization benchmark CiteSum and conduct a comprehensive analysis of it. (3) We further verify the quality and usefulness of CiteSum by demonstrating that models pre-trained on CiteSum perform very well on new tasks and domains such as news extreme summarization and headline generation with limited training.

2 CiteSum: A Large-scale Scientific Extreme Summarization Benchmark

2.1 Data Creation

Data Source We take the publicly available Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020) as the source for data creation. In the latest version of S2ORC, there are 136M scientific papers from different academic disciplines and the number of papers with full-text access is 12M. We further remove papers without citation information, resulting in 9M papers as the candidates.

Quality Control Not all citation texts are of high quality and can be used as summaries of the cited papers. In Table 2, we show two examples (in our paper) where the citation sentence simply (1) describes the data source or (2) introduces the difference of the citing paper from the cited paper. We note that prior studies on citation text generation (Chen et al., 2021; Ge et al., 2021) often do not filter these citation texts and simply treat all paragraphs/sentences with citations as the ground-truth labels, as their goals are not on paper summarization but writing assistance.

To ensure data quality, we carefully locate, extract, and filter the citation texts of papers in the following manner. First, we only take citation texts in the Related Work section of a paper, which largely ensures that they describe the content of the cited paper instead of irrelevant information, such as task

Dataset	Train / Val / Test	len$_{src}$	len$_{summ}$	Automatic?
Gigaword (Rush et al., 2015)	3,803,957 / 189,651 / 1,951	32	9	✓
XSum (Narayan et al., 2018)	204,045 / 11,332 / 11,334	431	23	✓
arXiv (Cohan et al., 2018)	203,037 / 6,436 / 6,440	4.9K	220	✓
SciSummNet (Yasunaga et al., 2019)	1,000 / - / -	4.7K	150	✗
TalkSumm (Lev et al., 2019)	1,716 / - / -	4.8K	150	✓
SciTLDR (Cachola et al., 2020)	1,992 / 619 / 618	159	21	✗
CiteSum	83,304 / 4,721 / 4,921	255	23	✓

Table 3: Statistics of relevant summarization datasets showing the number of samples per data split, the average number of words in the source document (src) and reference summary (summ), and whether dataset creation is automatic without human annotation. SciSummNet (Yasunaga et al., 2019) and TalkSumm (Lev et al., 2019) do not contain validation/test set as their model evaluation was done on another dataset (Jaidka et al., 2016).

background in the Introduction section or the implementation details in the Experiment section. After filtering papers without a Related Work section, there are around 288K papers left.

Second, we only keep citation sentences that cite a single paper, since those with multiple citations typically discuss one line of work and cannot be used as the summary of a specific paper. In total, we obtain about 426K citation sentences.

Next, we measure the similarity between the citation texts and the cited papers and filter dissimilar pairs. Intuitively, if a citation sentence can serve as a high-quality summary, certain amount of its content should be from the cited paper. Prior work (Lu et al., 2020) also showed that authors tend to cite a paper using the information in the abstract of the cited paper. We thus calculate the overlap between paper abstracts and their citation sentences, and filter those below a threshold $\mathcal{T}$. We set $\mathcal{T}$ to 50/20/40 for ROUGE-1/2/L recall through manual examination, resulting in a ROUGE-1/2/L recall of 73.1/39.4/58.5 after filtering.[3] As a reference, the ROUGE-1/2/L recall between paper abstracts and reference summaries on SciTLDR (Cachola et al., 2020) is 81.1/38.9/62.0 and 65.2/17.9/45.7 for author-provided (SciTLDR-Auth) and peer review-derived (SciTLDR-PR) TLDR, respectively. That is, the abstraction level of CiteSum is between SciTLDR-Auth and SciTLDR-PR. This filtering step is rather strict as we prefer quality to quantity of the data and only 93K of the 426K examples (21.8%) are kept.

We further replace each citation span (*e.g.*, "Taigman et al. [8]") with a special token "REF" as they vary in different papers but essentially have the same meaning (*i.e.*, referring to a cited paper).

Dataset Split After data filtering and preprocessing, there are 92,946 examples in the final *citation text-guided summarization* dataset, which we name as CiteSum. We take about 5% of the data as the validation and test sets respectively, and the remaining 90% as the training set. As one paper may be cited multiple times in different papers, we ensure that there is no label leakage by excluding papers used for evaluation from the training set.

2.2 Data Analysis

Dataset Statistics In Table 3, we show the data statistics of CiteSum and other relevant summarization datasets. In terms of data size, CiteSum is about half the size of other automatically constructed datasets like XSum (Narayan et al., 2018) and arXiv (Cohan et al., 2018) due to the availability of citation texts and our strict quality control. On the other hand, the size of CiteSum is much larger than *human-annotated* datasets on paper summarization (Yasunaga et al., 2019; Cachola et al., 2020) – almost 30 times larger than the SciTLDR dataset (Cachola et al., 2020).

When compared to SciTLDR, the average length of source documents in CiteSum is longer, while that of the reference summaries is similar as the majority of summaries in SciTLDR also involve one sentence. When compared to XSum, the summary length in CiteSum is also quite similar. However, the inputs in XSum are news articles instead of scientific papers and the input lengths also vary. As for Gigaword (Rush et al., 2015), a news headline generation dataset, both its source input and target output are much shorter than CiteSum. Despite such differences, we observe that our models pre-trained on CiteSum transfer very well to these datasets in zero-shot and few-shot settings (Sec. 4).

---

[3]We also experimented with semantic metrics such as BERTScore (Zhang et al., 2019) but they did not function as well as ROUGE-based metrics in our human evaluation.
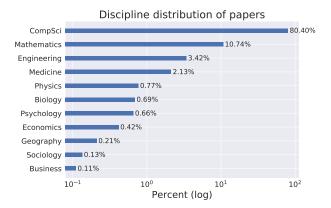
Figure 1: Discipline distribution of papers in CiteSum. Log scale is used for clearer illustration. Disciplines with lower than 0.1% distribution are omitted.

Figure 2: Citation distribution of papers in CiteSum.

Discipline Analysis   In Fig. 1, we show the discipline distribution of papers in CiteSum. The discipline information is derived from the field of study in Microsoft Academic Graph (MAG) (Shen et al., 2018). We take the top field of study for each paper if there are multiple. We note that the discipline distribution in CiteSum is quite different from its data source S2ORC (Lo et al., 2020) where medicine and biology dominate. In contrast, most papers in CiteSum are in computer science. The shift in discipline distribution is because we explicitly keep papers with a Related Work section, where around 82.8% are computer science papers. We then take the citation texts in the above papers, which largely lead to papers in similar disciplines. As a result, most papers in CiteSum are from computer science, mathematics, and engineering.

Citation Analysis   In Fig. 2, we show the average number of citations for papers in CiteSum. Note that the citation count shown does NOT reflect the total number of citations due to data filtering, but how many times a paper appears in CiteSum as examples (with the same input and different citation sentences as target output). In total, there are 59,707 unique papers in CiteSum with an average citation of 1.56, and 98% of the papers have fewer than 5 citations. Compared to prior work, we do not only target popularly cited papers (Yasunaga et al., 2019) and use different citation texts as different training examples instead of multiple reference summaries (Cachola et al., 2020).

Human Evaluation   We randomly sample 50 examples from CiteSum and ask two human annotators with a background in computer science to examine whether the citation sentences can serve
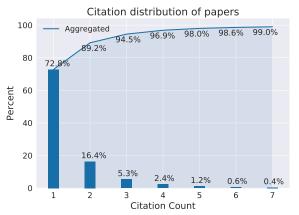
as high-quality summaries of the papers. Similar to Cachola et al. (2020), we use a 4-point scale for evaluation with 1 - false or misleading, 2 - partially accurate, 3 - mostly accurate, and 4 - accurate. The rating distribution is listed in Table 4. 80% citation sentences are considered (mostly) accurate to be used as summaries of the cited papers. On the other hand, there are still 10% misleading summaries, which we argue is quite common in automatically created summarization datasets (Mao et al., 2020a). We show 4 examples corresponding to each rating in App. A. We will further verify the quality of CiteSum by adapting models pre-trained on it to new tasks and domains (Sec. 4).

Rating	1	2	3	4
Percentage	10%	20%	28%	42%

Table 4: Ratings of citation sentences in CiteSum regarding whether they can serve as high-quality summaries of the cited papers.

3   Experiments on CiteSum

In this section, we experiment on CiteSum with state-of-the-art baselines and analyze their performance under different setups to provide references for future studies. Implementation and training details are provided in App. B.

3.1   Examined Methods

We use BART-large (Lewis et al., 2020) and PEGASUS-large (Zhang et al., 2020) as the base models as they are the state-of-the-art methods on multiple summarization datasets. We examine the base models with different inputs such as paper abstract (Abs), abstract+introduction+conclusion

(AIC), and abstract+title. In addition to using the TLDR (citation text) as the only generation target, we evaluate two multi-task settings with paper title and discipline (Disci)[4] as the targets, where different prefix tokens are added to the input such that the model can generate different targets given the same paper abstract as input (Cachola et al., 2020).

We further evaluate the following extractive baselines. EXT-LEAD: a method that takes the first sentence of the paper abstract, which performs fairly well in news summarization. EXT-HEURISTIC: a heuristic method that looks for the first sentence containing "propose", "introduce", or "in this paper", as such sentences likely reflect the contribution of the paper. It falls back to EXT-LEAD if no such sentences are found. EXT-ORACLE: an upper bound that matches each sentence in the paper abstract with the reference summary and takes the sentence with the highest ROUGE-2 F1.

3.2 Results

In Table 5, we show the results of various baseline methods on CiteSum. When given paper abstract as the source document, PEGASUS performs worse than BART and we thus use BART as the major model in the following experiments. Further adding paper introduction and conclusion to the model input slightly improves model performance, at the expense of longer training time and increased memory usage. The gains brought by adding title and discipline information to model input are marginal, while using them for multi-task learning does not lead to clearly better results. The fact that methods proposed by recent studies such as multi-task learning (Cachola et al., 2020) perform ineffectively on CiteSum indicates that CiteSum remains an unresolved and challenging scenario.

For the extractive baselines, EXT-LEAD performs significantly worse than that in the news domain (Mao et al., 2020a). EXT-HEURISTIC improves upon EXT-LEAD drastically and yet lags behind state-of-the-art methods by a large margin. EXT-ORACLE performs the best, the performance of which is generally consistent with the numbers on the human-annotated SciTLDR dataset (Cachola et al., 2020). On the other hand, the fact that abstractive methods have approached the extractive upper bound indicates that more abstraction is needed to further improve model performance on

---

[4]Here, we cast discipline classification as a seq2seq task (Mao et al., 2021). We found that all generated outputs are valid discipline names in our experiments.

Method		R-1	R-2	R-L
Source	**Target**			
Abs	TLDR	41.86	19.36	33.72
Abs	TLDR (PEGASUS)	41.56	18.63	33.45
AIC	TLDR	41.99	19.52	33.89
Abs+Title	TLDR	42.02	19.44	33.78
Abs+Disci	TLDR	42.01	19.34	33.72
Abs	TLDR/Title	41.85	19.21	33.42
Abs	TLDR/Title/Disci	41.89	19.51	33.73
EXT-LEAD		21.94	7.35	17.36
EXT-HEURISTIC		29.32	12.53	23.99
EXT-ORACLE		44.17	27.22	38.32

Table 5: Performance of different methods on CiteSum. BART-large is used as the base model if not otherwise specified. "/" indicates multi-task learning. R stands for ROUGE (Lin, 2004) in all the tables.

CiteSum.

We believe that CiteSum provides a well-established testbed for future studies on (scientific) extreme summarization. The following future directions may be worth exploring: 1) how to better understand the structure and content of scientific papers with domain knowledge (via relevant papers, terminology, taxonomies, etc); 2) how to better capture the differences in writing styles across various domains; and 3) how to improve the saliency, factual correctness, and explainability of TLDR summaries given their conciseness.

4 Transferring to New Tasks and Domains with CITES

To further verify the quality and usefulness of CiteSum, we adapt models pre-trained on CiteSum to new tasks and domains, some of which are rather different from CiteSum and make model transfer with limited supervision very challenging.

Specifically, we name our pre-trained model as CITES (**C**itation **T**ext-guided **S**ummarizer). CITES uses the simplest form in Sec. 3 with paper abstract as input and TLDR as target output. We evaluate CITES on various downstream tasks with no fine-tuning (zero-shot) or limited training examples (few-shot), including scientific extreme summarization on SciTLDR (Cachola et al., 2020), news extreme summarization on XSum (Narayan et al., 2018), and news headline generation on Gigaword (Rush et al., 2015). Additionally, we evaluate CITES on two more diverse tasks in the scientific domain, namely discipline classification and title generation, in a fully-supervised setting.

4.1 Scientific Extreme Summarization

Setup SciTLDR (Cachola et al., 2020), the human-annotated scientific extreme summarization dataset, is an ideal testbed for further verifying the quality and usefulness of CiteSum since they both target extreme summarization, belong to the scientific domain (though CiteSum involves more disciplines), and share similar input/output formats (though CiteSum has slightly longer inputs). One noticeable difference, however, is the point of view of the summaries – in SciTLDR the reference summaries typically start with "We" or "This paper", while in CiteSum they often begin with "Author-Name et al." (replaced by a special token "REF" during preprocessing).

We propose two simple techniques to tackle such subtle style differences when adapting CITES to SciTLDR in a zero-shot setting without fine-tuning. The first technique is *post-processing*: we replace "REF" with "This paper" if the summary begins with "REF" and remove all other "REF" within the summary. The second technique is *prompting*: we use "This paper" as a prompt in the model decoder such that the summary always starts with "This paper". Similarly, in the few-shot setting, we replace the leading "We" with "This paper REF" in the reference summaries of SciTLDR (on the training set only) to alleviate the style mismatch.

We use BART-large (Lewis et al., 2020) as the base model of CITES since most baselines on SciTLDR, including the state-of-the-art methods, use the same base model.

Zero-shot Results In Table 6, we show the performance comparison of different methods on SciTLDR. In the zero-shot setting, CITES (post-processing) outperforms competitive *fully-supervised* baselines such as BERTSum (Liu and Lapata, 2019). CITES (prompting) performs even better than CITES (post-processing), outperforming the fully-supervised BART model it is based upon. Such results demonstrate the benefits of pre-training on CiteSum. CITES (prompting), without any fine-tuning, is also on par with the state-of-the-art method CATTS (Cachola et al., 2020), while slightly worse than CATTS$_{XSUM}$, which pre-trains on the XSum dataset (Narayan et al., 2018) first.

We additionally test a zero-shot upper bound for CITES by providing our prompting model with the *first 3 tokens* in the reference summary (the most common ones are "We propose a" and "We present a") such that it knows how to start to sum-

Method	R-1	R-2	R-L
EXT-ORACLE	**47.7**	24.7	38.5
Fully-supervised			
PACSUM (Zheng and Lapata, 2019)	19.3	4.0	15.1
BERTSum (Liu and Lapata, 2019)	38.5	16.6	30.5
MatchSum (Zhong et al., 2020)	42.7	20.0	34.0
BART (Lewis et al., 2020)	43.3	20.8	35.0
BART$_{XSUM}$ (Lewis et al., 2020)	42.5	21.1	34.9
CATTS (Cachola et al., 2020)	43.8	20.9	35.5
CATTS$_{XSUM}$ (Cachola et al., 2020)	44.3	21.3	35.9
Zero-shot			
CITES (post-processing)	43.4	19.7	34.8
CITES (prompting)	43.5	20.9	**36.0**
CITES (prompting, gold 3 tokens)	46.3	**27.4**	**41.9**
Few-shot			
CITES (32-shot)	43.8	21.4	36.4
BART (128-shot)	42.1	19.1	33.4
CITES (128-shot)	**44.5**	**21.6**	**36.5**

Table 6: Performance comparison on SciTLDR (Cachola et al., 2020) using its official evaluation script.

marize and (hopefully) which aspect to focus on. CITES (prompting, gold 3 tokens) achieves competitive ROUGE-1 and significantly better ROUGE-2/L than the extractive upper bound EXT-ORACLE that has access to the *entire reference summary*.

Few-shot Results In the few-shot setting, CITES with 32 examples improves over its zero-shot counterpart. Furthermore, 128-shot CITES outperforms all fully-supervised methods and achieves new state-of-the-art results on SciTLDR. In contrast, a 128-shot BART model without first pre-training on CiteSum largely lags behind, performing even worse than our zero-shot CITES. Such results again show the effectiveness of our pre-training strategy and the quality of CiteSum despite being automatically created thanks to our quality control.

Data Overlap To ensure that the superior generalizability of CITES does not merely come from data leakage, we detect the overlap between CiteSum and SciTLDR. We consider two papers (near) identical if their TF-IDF cosine similarity is greater than 0.9 and find that only 9.7% papers in the test set of SciTLDR appear in the training set of CiteSum. Also, note that the training labels in CiteSum are automatically extracted citation sentences and different from SciTLDR.

4.2 Scientific Discipline Classification and Title Generation

We have demonstrated the effectiveness of CITES on the task of scientific extreme summarization.

	Title Generation			Discipline Classification	
Method	R-1	R-2	R-L	Macro-F1	Weighted-F1
BART	52.03	30.15	45.99	0.24	0.77
CITES	**52.50**	**30.42**	**46.26**	**0.30**	**0.78**

Table 7: Comparison of CITES and its base model (BART) on title generation and discipline classification.

Next, we explore the feasibility of transferring CITES to more diverse tasks.

Setup We evaluate CITES with the task of scientific discipline classification and title generation. Similar to the multi-task experiments in Sec. 3, we use the same dataset split and model input, while replacing the generation target from summaries (citation texts) to the discipline or title of the papers. Examples with unavailable discipline or title are removed. We use BART-large as the base model for this experiment and compare BART with CITES in an apple-to-apple comparison.

Results In Table 7, we show the performance comparison on title generation and discipline classification. CITES consistently outperforms BART on both tasks, although the differences are not as significant as in other low-resource transfer experiments. The moderate gains are possibly because there is abundant training data for the two tasks and continuous pre-training thus does not help much. As evidence, the (unweighted) Macro-F1 of CITES is considerably better than BART, which we found is because CITES performs well on those disciplines with fewer examples. Regarding the Weighted-F1, CITES is only slightly better as most papers belong to a single discipline (computer science) that dominates the score.

4.3 News Extreme Summarization

Setup With the success on different tasks in the scientific domain, we next evaluate CITES on a more difficult setting where the domain is significantly different while the task is still extreme summarization. We take the XSum dataset (Narayan et al., 2018) in the news domain for this purpose. We mainly use PEGASUS-large (Zhang et al., 2020) as the base model of CITES as its fully-supervised version holds the state-of-the-art results on XSum. We additionally evaluate CITES_Title in the zero-shot setting, which is the variant used for title generation in Sec. 4.2.

Method	R-1	R-2	R-L
Fully-supervised			
PTGEN (See et al., 2017)	29.70	9.21	23.24
BERTSum (Liu and Lapata, 2019)	38.81	16.50	31.27
BART (Lewis et al., 2020)	45.14	22.27	37.25
PEGASUS (Zhang et al., 2020)	**47.21**	**24.56**	**39.25**
Zero-shot			
BART (Lewis et al., 2020)	15.40	2.63	10.74
PEGASUS (Zhang et al., 2020)	19.27	3.00	12.72
T5-LB (Zhu et al., 2021a)	26.06	6.77	20.47
BART-LB (Zhu et al., 2021a)	26.18	7.60	20.92
WikiTransfer (Fabbri et al., 2021)	**31.85**	**10.44**	**23.75**
CITES	26.43	7.17	19.64
CITES_Title	28.21	8.40	21.81
10-shot			
BART (Lewis et al., 2020)	31.34	9.98	23.44
PEGASUS (Zhang et al., 2020)	19.39	3.45	14.02
WikiTransfer (Fabbri et al., 2021)	35.17	12.76	26.80
CITES	**36.21**	**14.22**	**28.18**
100-shot			
BART (Lewis et al., 2020)	34.16	12.62	26.66
PEGASUS (Zhang et al., 2020)	39.07	16.44	31.27
WikiTransfer (Fabbri et al., 2021)	37.26	14.20	28.85
CITES	**41.45**	**18.74**	**33.29**

Table 8: Performance comparison on the XSum dataset. Our few-shot results are averaged over 3 runs.

Zero-shot Results In Table 8, we show the results on XSum with various training data sizes. In the zero-shot setting, CITES significantly improves over its base model PEGASUS (+7.2 ROUGE-1). In addition, CITES is on par with other pre-trained models such as BART-LB and T5-LB (Zhu et al., 2021a), which are specifically designed to leverage the lead bias in the news domain and require much more resources (32 vs. 1 GPU, 21.4M vs. 83K training examples) for summarization pre-training. CITES_Title further improves over CITES and outperforms most zero-shot baselines (+8.9 ROUGE-1 over PEGASUS). CITES_Title does not outperform WikiTransfer (Fabbri et al., 2021), which is somewhat expected as WikiTransfer carefully prepares its pre-training data to specific downstream tasks given, *e.g.*, summary length and its level of abstraction. Unlike WikiTransfer (Fabbri et al., 2021), CITES does not involve any downstream task-specific data selection or model tuning – we use the same CiteSum corpus in all the experiments.

Few-shot Results When given a few examples for fine-tuning, CITES quickly adapts to the new task despite the domain mismatch during pre-training. We observe that CITES consistently outperforms not only its base model but all other base-

Method	R-1	R-2	R-L
Fully-supervised			
PEGASUS (Zhang et al., 2020)	39.12	19.86	36.24
Unsupervised			
SEQ³ (Baziotis et al., 2019)	25.39	8.21	22.68
Brief (Wang and Lee, 2018)	21.26	5.60	18.89
TED (Yang et al., 2020)	25.58	8.94	22.83
Zero-shot			
T5 (Raffel et al., 2020)	15.67	4.86	14.38
BART (Lewis et al., 2020)	22.07	7.47	20.02
PEGASUS (Zhang et al., 2020)	23.39	7.59	20.20
T5-LB (Zhu et al., 2021a)	24.00	8.19	21.62
BART-LB (Zhu et al., 2021a)	25.14	8.72	22.35
CITES	24.75	8.42	21.84
CITES$_{Title}$	**27.87**	**10.43**	**24.56**

Table 9: Performance comparison on the Gigaword news headline generation dataset (Rush et al., 2015).

line methods, including WikiTransfer, and achieves state-of-the-art few-shot performance on XSum. In particular, CITES performs better than fully-supervised methods such as BERTSum (Liu and Lapata, 2019) with only 100 examples.

4.4 News Headline Generation

Setup To take a step further, we study the transfer performance of CITES to news headline generation. We use the Gigaword headline generation dataset (Rush et al., 2015) for this evaluation. We again consider two variants of CITES, one pre-trained with citation texts as the generation target and the other further pretrained with paper titles as in Sec. 4.2. We use BART-large (Lewis et al., 2020) as the base model in this evaluation.

Results In Table 9, we show the results of various methods on news headline generation. CITES again outperforms its base model (BART) significantly and achieves competitive performance with most unsupervised and zero-shot methods designed for news summarization (Zhang et al., 2020; Zhu et al., 2021a). CITES$_{Title}$ further achieves state-of-the-art zero-shot performance despite pre-training on the scientific domain, demonstrating the generalizability and usefulness of CiteSum.

5 Related Work

Citation Text Generation There have been prior studies utilizing citation texts for different purposes. One popular line of work focuses on the generation of the citation texts for writing assistance or paper comparison (Xing et al., 2020; Luu et al., 2021;

Chen et al., 2021; Ge et al., 2021). However, they typically do not distinguish the citation texts that can serve as summaries of the cited paper from those used for other purposes, *e.g.*, background or result comparison (Cohan et al., 2019). For example, Chen et al. (2021) treat citation text generation as a multi-document summarization task, where the target output is a paragraph with more than two citations and the model input is the abstracts of all cited papers. There is no filtering regarding the citation texts and all the paragraphs with enough citations are included. Besides including citation texts with various intents and the lack of quality control, prior studies differ from CiteSum in that they target longer outputs, *e.g.*, multiple sentences (Xing et al., 2020) or the entire Related Work section (Lu et al., 2020; Chen et al., 2021).

Citation Text for Paper Summarization Another line of work does not generate but extracts the citation texts and either uses them to form a summary directly (Nakov et al., 2004; Abu-Jbara and Radev, 2011; Qazvinian et al., 2013) or treats them as a bridge to the cited paper (Cohan and Goharian, 2015; Yasunaga et al., 2019). Specifically, the citation texts in the latter studies are used to find relevant contexts in the cited paper (called *citation contexts*). Then, a *long* summary is formulated primarily using the cited paper, *e.g.*, by selecting sentences from the citation contexts (Cohan and Goharian, 2015). Unlike CITES, prior citation-based summarization methods require (often multiple) citation texts of a paper as *input*, which are unavailable for new papers. In addition, they do not target ultra-short but abstract-long summaries.

Extreme Summarization Extreme summarization aims to form ultra-short summaries of the documents. Notable benchmarks in this direction include XSum (Narayan et al., 2018) and NewSHead (Gu et al., 2020) in the news domain, Sci-TLDR (Cachola et al., 2020) in the scientific domain, and Webis-TLDR-17 (Völske et al., 2017) for social media summarization. Compared to Sci-TLDR, our CiteSum dataset is significantly larger in scale, from more venues than OpenReview, and composed of various disciplines.

Summarization with Limited Supervision Our work is also related to unsupervised and zero/few-shot summarization that constructs weakly supervised guidance signals using e.g., data characteristics (Chu and Liu, 2019; Mao et al., 2020b), domain

knowledge (Zhu et al., 2021b), or pseudo labeled data (Yang et al., 2020; Zhong et al., 2022). Compared to prior studies, CITES shows great cross-domain capability that has not been well explored.

6 Conclusion

In this paper, we propose a simple yet effective approach to automatically extracting ultra-short paper summaries from citation texts. Based on the proposed approach, we create a large-scale, high-quality benchmark for scientific extreme summarization. We conduct a comprehensive analysis on the created benchmark and further demonstrate that models pre-trained on it exhibit superior generalizability to new tasks and domains such as news extreme summarization and headline generation with limited supervision.

Limitations

Regarding data collection, while we have taken multiple steps to improve data quality, as in all automatically created datasets, there are still low-quality examples. We show some examples of low quality in App. A. Limiting citation texts to the Related Work section improves data quality, but also excludes the majority of available citation sentences and makes CiteSum concentrated in the field of computer science and engineering.

Regarding model performance, our transfer experiments are performed in scientific and news domains. While promising, there is no guarantee that CITES works well in other domains. Also, with abundant in-domain training data, pre-training on CiteSum may not lead to significant improvements.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 500–509, Portland, Oregon, USA. Association for Computational Linguistics.

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. SEQ^3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. 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 673–681, Minneapolis, Minnesota. 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.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6068–6077, Online. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596, 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.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400, Lisbon, Portugal. Association for Computational Linguistics.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online. Association for Computational Linguistics.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. BACO:

A background knowledge- and content-based framework for citing sentence generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1466–1478, Online. Association for Computational Linguistics.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating Representative Headlines for News Stories. In *Proc. of the the Web Conf. 2020*.

Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2016. Overview of the CL-SciSumm 2016 shared task. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 93–102.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. TalkSumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2125–2131, Florence, Italy. 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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. MultiXScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 8068–8074, Online. Association for Computational Linguistics.

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining relationships between scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2130–2144, Online. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Reader-guided passage reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 344–350, Online. Association for Computational Linguistics.

Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han. 2020a. Facet-aware evaluation for extractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4941–4957, Online. Association for Computational Linguistics.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020b. 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.

Preslav I Nakov, Ariel S Schwartz, Marti Hearst, et al. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR*, volume 4, pages 81–88. Citeseer.

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.

Vahed Qazvinian, Dragomir R Radev, Saif M Mohammad, Bonnie Dorr, David Zajic, Michael Whidby, and Taesun Moon. 2013. Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research*, 46:165–201.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015*

Conference on Empirical Methods in Natural Language Processing, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92, Melbourne, Australia. Association for Computational Linguistics.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Yaushian Wang and Hung-Yi Lee. 2018. Learning to encode text as human-readable summaries using generative adversarial networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4187–4195, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6181–6190, Online. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. 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, Yang Liu, Suyu Ge, Yuning Mao, Yizhu Jiao, Xingxing Zhang, Yichong Xu, Chenguang Zhu, Michael Zeng, and Jiawei Han. 2022. Unsupervised summarization with customized granularities. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021a. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1462–1471.

Qi Zhu, Fang Guo, Jingjing Tian, Yuning Mao, and Jiawei Han. 2021b. Sumdocs: Surrounding-aware unsupervised multi-document summarization. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 477–485. SIAM.

A Data Examples in CiteSum

In Tables 10 and 11, we show four data examples in CiteSum corresponding to different ratings in the human evaluation. While some of the examples are still of low quality after quality control, most of the filtered citation texts can serve as high-quality summaries.

B Implementation Details

Official results of the baselines are taken from prior studies when possible. Model checkpoint selection is done on the validation set for every task. The special token "REF" (used to indicate citation span) is removed from model output for all transfer experiments. Paper abstract is used as input for all compared methods on SciTLDR. We experimented with other prompts such as "We" and "In REF" but found "This paper" works the best. FP16 is used in most experiments for training efficiency except for pre-training PEGASUS on CiteSum, with which it failed to learn. We use a batch size of 8. Hyperparameters like min/max generation length are generally set following prior work (Zhang et al., 2020).

All the experiments are conducted with 1 Nvidia RTX A6000 GPU. Pre-training on CiteSum only takes about 6.5h for BART and 10h for PEGASUS. The transfer experiments typically take less than 1h (time mostly spent on evaluation) as we use very few labeled data for training. The codebase is based on Huggingface transformers (Wolf et al., 2020).

<Rating 1>

Paper Title: Congested traffic states in empirical observations and microscopic simulations

Paper Abstract: We present data from several German freeways showing different kinds of congested traffic forming near road inhomogeneities, specifically lane closings, intersections, or uphill gradients. The states are localized or extended, homogeneous or oscillating. Combined states are observed as well, like the coexistence of moving localized clusters and clusters pinned at road inhomogeneities, or regions of oscillating congested traffic upstream of nearly homogeneous congested traffic. The experimental findings are consistent with a recently proposed theoretical phase diagram for traffic near on-ramps [D. Helbing, A. Hennecke, and M. Treiber, Phys. Rev. Lett. 82, 4360 (1999)]. We simulate these situations with a novel continuous microscopic single-lane model, the "intelligent driver model" (IDM), using the empirical boundary conditions. All observations, including the coexistence of states, are qualitatively reproduced by describing inhomogeneities with local variations of one model parameter. We show that the results of the microscopic model can be understood by formulating the theoretical phase diagram for bottlenecks in a more general way. In particular, a local drop of the road capacity induced by parameter variations has practically the same effect as an on-ramp.

Citation Text: In a first approach, we use the well-known "intelligent driver model" (IDM) REF to show that the method works.

<Rating 2>

Paper Title: Probabilistic Model-Agnostic Meta-Learning

Paper Abstract: Meta-learning for few-shot learning entails acquiring a prior over previous tasks and experiences, such that new tasks be learned from small amounts of data. However, a critical challenge in few-shot learning is task ambiguity: even when a powerful prior can be meta-learned from a large number of prior tasks, a small dataset for a new task can simply be too ambiguous to acquire a single model (e.g., a classifier) for that task that is accurate. In this paper, we propose a probabilistic meta-learning algorithm that can sample models for a new task from a model distribution. Our approach extends model-agnostic meta-learning, which adapts to new tasks via gradient descent, to incorporate a parameter distribution that is trained via a variational lower bound. At meta-test time, our algorithm adapts via a simple procedure that injects noise into gradient descent, and at meta-training time, the model is trained such that this stochastic adaptation procedure produces samples from the approximate model posterior. Our experimental results show that our method can sample plausible classifiers and regressors in ambiguous few-shot learning problems.

Citation Text: They extended their approach by incorporating a probabilistic component such that for a new task, the model is sampled from a distribution of models to guarantee a higher model diversification for ambiguous tasks REF.

Table 10: Examples in CiteSum with different quality ratings.

<Rating 3>

Paper Title: A Generic Multi-Projection-Center Model and Calibration Method for Light Field Cameras

Paper Abstract: Light field cameras can capture both spatial and angular information of light rays, enabling 3D reconstruction by a single exposure. The geometry of 3D reconstruction is affected by intrinsic parameters of a light field camera significantly. In the paper, we propose a multi-projection-center (MPC) model with 6 intrinsic parameters to characterize light field cameras based on traditional twoparallel-plane (TPP) representation. The MPC model can generally parameterize light field in different imaging formations, including conventional and focused light field cameras. By the constraints of 4D ray and 3D geometry, a 3D projective transformation is deduced to describe the relationship between geometric structure and the MPC coordinates. Based on the MPC model and projective transformation, we propose a calibration algorithm to verify our light field camera model. Our calibration method includes a close-form solution and a non-linear optimization by minimizing re-projection errors. Experimental results on both simulated and real scene data have verified the performance of our algorithm.

Citation Text: Zhang et al REF proposed a multi-projection-center (MPC) model with six intrinsic parameters to characterize both conventional and focused LF cameras.

<Rating 4>

Paper Title: Advancing Research Infrastructure Using OpenStack

Paper Abstract: Abstract-Cloud computing, which evolved from grid computing, virtualisation and automation, has a potential to deliver a variety of services to the end user via the Internet. Using the Web to deliver Infrastructure, Software and Platform as a Service (SaaS/PaaS) has benefits of reducing the cost of investment in internal resources of an organisation. It also provides greater flexibility and scalability in the utilisation of the resources. There are different cloud deployment models -public, private, community and hybrid clouds. This paper presents the results of research and development work in deploying a private cloud using OpenStack at the University of Huddersfield, UK, integrated into the University campus Grid QGG. The aim of our research is to use a private cloud to improve the High Performance Computing (HPC) research infrastructure. This will lead to a flexible and scalable resource for research, teaching and assessment. As a result of our work we have deployed private QGG-cloud and devised a decision matrix and mechanisms required to expand HPC clusters into the cloud maximising the resource utilisation efficiency of the cloud. As part of teaching and assessment of computing courses an Automated Formative Assessment (AFA) system was implemented in the QGG-Cloud. The system utilises the cloud's flexibility and scalability to assign and reconfigure required resources for different tasks in the AFA. Furthermore, the throughput characteristics of assessment workflows were investigated and analysed so that the requirements for cloud-based provisioning can be adequately made.

Citation Text: In REF , the authors focus on the use of a private cloud environment in order to improve the High Performance Computing (HPC) research infrastructure.

Table 11: Examples in CiteSum with different quality ratings.