Learning to Read Academic Papers

Yining Hong, Jialu Wang, Xueheng Zhang, Zheng Wu
Shanghai JiaoTong University
{evelinehong, faldict, zhangxueheng, 14wuzheng}@sjtu.edu.cn

Abstract

We present *PaperQA*, a challenging dataset of over 6000 human-generated question-answer pairs concerning academic knowledge. Crowdworkers supply questions and answers based on a set of over 1,000 abstracts from deep learning papers, with answers consisting of spans of text from the corresponding abstracts. This dataset is aimed at helping machines learn to read academic papers. We collect this dataset through a four-stage process designed to solicit exploratory questions that require reasoning. Then we propose a semantic segmentation model to solve this task and evaluate it on our dataset. Finally, we build a website which can interactively display the results of our model on newest papers and improve our model.

Introduction

Teaching machine to read is a non-negligible part of 'True AI', people are making progress since the renaissance of deep learning, however, were not even close, the state-of-the-art models still hard to beat a human kid. Teaching machine to read paper is an even more untouchable dream. To challenge this task, we can start with training machines to do reading comprehension questions, like a child, and use the accuracies of question answers to indirectly represent how machines read and comprehend, which is smart because we need some metrics to evaluate.

Nowadays, there are several medias in China which provide latest news about machine learning papers, such as PaperWeekly and so on. In order to extract the most important information from these papers, paper reading groups are formed. However, this requires a large amount of human resource. We intend to replace human resource with machine in this process, and use machine to present some important information for us based on machine reading comprehension. To do so, we first need a machine reading comprehension dataset based on papers.

In this paper, we present a novel dataset for machine reading comprehension on academic abstracts: *PaperQA*. *PaperQA* consists of over 6,000 question-answer pairs based on a set of over 1,000 abstracts from machine learning papers, including papers accepted by top machine learning conferences(such as AAAI, NIPS, CVPR, ICCV, ICML, ACL, ECCV and EMNLP), and papers submitted on arXiv.org. The questions are fixed in each abstract, concerning objectives, methods, models, experiments and others. Answers to these questions consist of spans of the corresponding abstract that are highlighted by students of machine learning background.

The purpose of releasing *PaperQA* is twofold. First, by releasing this dataset, we propose a novel machine reading comprehension task on papers. Second, from an application perspective, it provides researchers with a tool to efficiently identify the most important information in a paper and decide whether to continue with the paper or not.

Some characteristics of *PaperQA* that make it challenging and distinguish it from prior machine reading datasets are listed as follows:

- It is a machine reading comprehension based on academic papers, which requires machines to learn prior knowledge.
- Some of the questions require reasoning beyond simple sentence-level or word-level analysis.
- The answer to each question is a span (i.e., sequence of words) of arbitrary length.
- Some questions have no answer in the corresponding article (the null span).

In this paper, we describe the dataset collection process. To assess the difficulty of *PaperQA*, We propose a baseline model based on sentence-level classification and word-level classification to, and evaluate its performance on our dataset. To set an example of how our dataset can assist in academic research, we build a website called *AceNews* which extracts important information of newest machine reading papers, and presents comprehension of these papers. Moreover, we recommend papers for different users based on the information extracted and user behaviours. All of the above is accomplished by machine.

Existing Datasets

We start with a survey of existing machine reading comprehension datasets and datasets of abstracts, which vary in sources, size, difficulty, collection methodology and format of answers. We discuss about various sources of arti-

[h!] Dataset	Sources	Formulation
PaperQA	machine learning abstracts	spans in abstract
MCTest (Richardson, Burges, and Renshaw 2013)	stories	MRC multiple choice
CBT (Hill et al. 2015)	stories from children's book	MRC cloze
CNN/Daily Mail (Hermann et al. 2015)	CNN news	MRC cloze
SQuAD (Rajpurkar et al. 2016)	Wikipedia articles	MRC spans in passage
PubMed (Dernoncourt and Lee 2017)	medical abstracts	sentence classification

Table 1: A survey of several reading comprehension datasets and datasets of abstracts. *PaperQA* is the only MRC dataset consisting of academic abstracts.

cles and task formulation in these datasets (see Table1 for an overview).

Machine Reading Comprehension Datasets

Machine Comprehension Test (MCTest) This is a dataset from MSR, which contains 660 stories, each story has 4 human asked questions (Natural Language Question), and for each question, therere 4 candidate answers. This is pretty much like reading comprehension questions for pupils. Most of the stories are short and sentences are fairly short as well, and the size of vocabulary is small.

Childrens Book Test (CBT) A dataset from FAIR, which contains stories from childrens books. Each story in this dataset is a 20 consecutive sentences from childrens books, and remove a word from the consecutive 21st sentence, as the question, or query. Therere 4 splits of this dataset which are classified by the distinct types of word removed in queries: Named Entities, Common Nouns, Verbs, Prepositions. This type of fill in the blank query is called Cloze type question. For each question, therere 10 candidate answers which taken from the story, and all have same POS with the correct answer word.

CNN/Daily Mail *CNN/Daily Mail* QA dataset is released by Google DeepMind, which the largest (AFAIK) QA dataset. *CNN* dataset contains over 90K of of CNN news, and averagely has 4 queries per story, which gives 380K of story-question pairs; *Daily Mail* has about 200K new stories, and also, each story has 4 queries, which totally gives 880K story-question pairs.

The Stanford Question Answering Dataset (SQuAD) This dataset is recently released by Stanford University, which contains about 100K of question-answer pairs from 536 articles, the story for each question is a paragraph from these articles. Questions in *SQuAD* dataset are generated by crowdworkers so theyre NLQ. The formulation of answers is spans of the passage, which is similar to *PaperQA*.

Datasets of Abstracts

PubMed 200k RCT It is a dataset based on for sequential sentence classification. The dataset consists of approximately 200,000 abstracts of randomized controlled trials, totaling 2.3 million sentences. Each sentence of each abstract is labeled with their role in the abstract using one of the following classes: background, objective, method, result, or conclusion. *PubMed 200k* is similar to *PaperQA* in that they are both datasets of abstracts, and the categories of questions *PaperQA* are much like the classes in *PubMed 200k*. However, *PaperQA* searches for more specific answers, which are spans rather than sentences, thus increasing the difficulty. Moreover, *PaperQA* is based on machine learning papers while *PubMed 200k* is based on medical papers.

Dataset Construction

We collected *PaperQA* through a four-stage process: paper curation, question posing, answer sourcing, and dataset cleanup. These steps are detailed as follows.

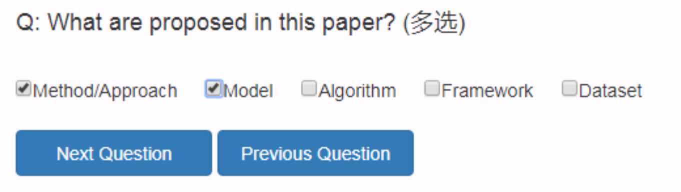
Paper Curation

To retrieve high-quality paper abstracts, we use most cited papers after the year of 2012 in top conferences, including AAAI, ICCV, ECCV, EMNLP, NIPS, CVPR, and ACL. The reason why we do this is that papers of different fields and different time vary a lot from each other. Due to the otherness of contributions in these papers, it is hard for us to raise some general questions concerning important information. Also, the distinctions in abstract structures make it hard for machine to learn the different patterns. The recent five years have witnessed a boost of paper of Artificial Intelligence, especially those on deep learning. These papers share a lot in common in structure, purpose and others. We assume that results on this single area can be good enough to provide for our machine learning researchers some assitance in reading paper or some insights into machine reading comprehension on papers. If so, we'll then move on to other fields.

Question Posing

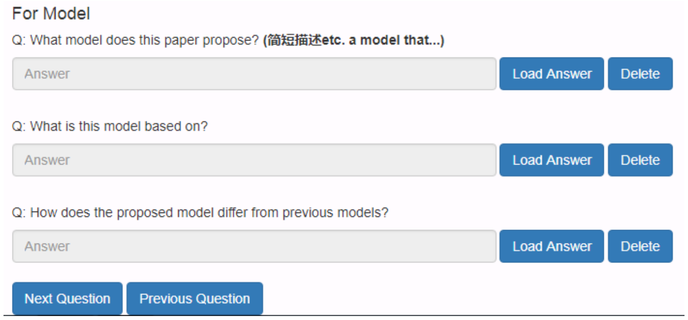
We intend to extract the most important information in papers. The information we want is much alike in every abstract, mainly concerning objective, problem addressed, experiment and its result, and what the paper proposed. Therefore, we decide to fix our questions on all abstracts.

Different papers propose different items, including methods, models, algorithms, frameworks, datasets, and others. We set a checkbox to help people select the items proposed in papers. According to items selected, questions concerning each item is presented in the answer sourcing website, as is shown in Figure1 and Figure2.


Q: What are proposed in this paper? (多选)

☑ Method/Approach ☑ Model ☐ Algorithm ☐ Framework ☐ Dataset

Figure 1: Checkbox of what are proposed in papers on the crowdsourcing website.


For Model

Q: What model does this paper propose? **(简短描述etc. a model that...)**

Answer

Q: What is this model based on?

Answer

Q: How does the proposed model differ from previous models?

Answer

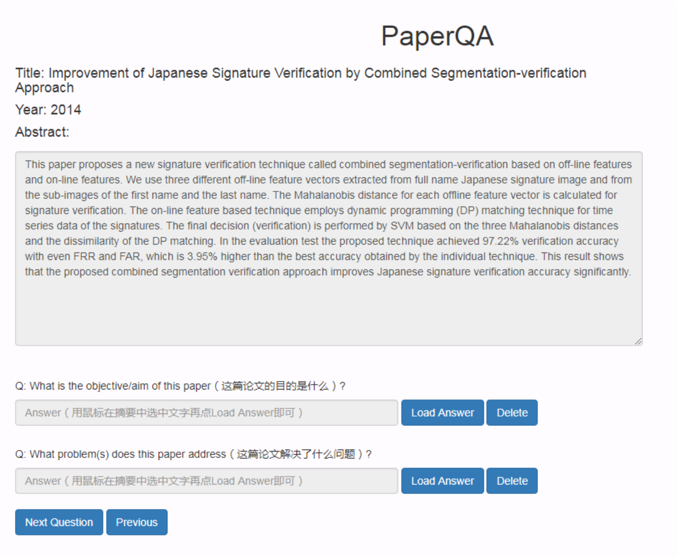
Figure 2: Questions according to what proposed in papers on the crowdsourcing website.

Answer Sourcing

We create an interactive crowdsourcing website, which randomly presents a paper in our database. Users answer questions in the provided paper, and the answers can be stored in our database. Our crowdworkers are students in Shanghai Jiaotong University who have taken machine learning classes before. The students are required to answer 8 questions in each abstract. They may also reject the question as nonsensical, or select the null answer if the abstract contains insufficient information. Answers are submitted by clicking on and highlighting words in the article, while instructions encourage the set of answer words to consist of a single continuous span (again, we give an example prompt in the Appendix). The crowdsourcing website is shown in Figure3.

Dataset Cleanup

After collecting more than 1000 abstracts, we filter the answers too short and check them manually. To our surprise, the filled answers show strong professional skills with a high quality. To obtain a dataset of the highest possible quality we use a validation process that mitigates issues. We examine the dataset by ourselves to leave out some obviously wrong answers. Then we put every abstracts along with all the non-empty question-answer pairs in a json file. Finally our dataset contains 1,030 abstracts and 8,374 question-answer pairs.


PaperQA

Title: Improvement of Japanese Signature Verification by Combined Segmentation-verification Approach

Year: 2014

Abstract:

This paper proposes a new signature verification technique called combined segmentation-verification based on off-line features and on-line features. We use three different off-line feature vectors extracted from full name Japanese signature image and from the sub-images of the first name and the last name. The Mahalanobis distance for each offline feature vector is calculated for signature verification. The on-line feature based technique employs dynamic programming (DP) matching technique for time series data of the signatures. The final decision (verification) is performed by SVM based on the three Mahalanobis distances and the dissimilarity of the DP matching. In the evaluation test the proposed technique achieved 97.22% verification accuracy with even FRR and FAR, which is 3.95% higher than the best accuracy obtained by the individual technique. This result shows that the proposed combined segmentation verification approach improves Japanese signature verification accuracy significantly.

Q: What is the objective/aim of this paper（这篇论文的目的是什么）?

Answer（用鼠标在摘要中选中文字再点Load Answer即可）

Q: What problem(s) does this paper address（这篇论文解决了什么问题）?

Answer（用鼠标在摘要中选中文字再点Load Answer即可）

Figure 3: Crowdsourcing website sample.

Dataset Analysis

Table 2 counts the number of answers per question and shows their category: the least proposed methods are framework and dataset, and we think the common proposed methods, models and algorithms are similar in meaning. Several experiment results are allowed for one paper, so there are 1171 answers for the question ”What experiment does this paper carry out to evaluate the result?”, exceeds the total number of abstracts. This table indicates that our dataset is not excessively unbalanced.

Methods and Performance

We propose a semantic segmentation model and evaluate its performance on our dataset. Our approach can be divided into two parts: sentence-level text classification and word-level sequence tagging.

Sentence-level text classification

We summarize all kinds of the questions and divide them into three categories: *purpose*, *methods* and *experiments*. Then every abstracts in our dataset are separated into sentences. We check each sentence whether a certain answer in our dataset is constituent and label it with corresponding question's category. If the sentence doesn't contain any answers, we label it as *others*. In total, we clean out 6,383 sentences with four-category labels.

Category	Question	Numbers
Purpose	What is the objective/aim of this paper?	963
Purpose	What problem(s) does this paper address?	857
Methods	What method/approach does this paper propose?	594
Methods	What is this method based on?	395
Methods	How does the proposed method differ from previous methods/approaches?	338
Methods	What model does this paper propose?	198
Methods	What is this model based on?	133
Methods	How does the proposed model differ from previous models?	122
Methods	What algorithm does this paper propose?	222
Methods	What is this algorithm based on?	143
Methods	How does the proposed algorithm differ from previous algorithms?	135
Methods	What framework does this paper propose?	120
Methods	What is this framework based on?	70
Methods	How does the proposed framework differ from previous frameworks?	61
Methods	What dataset does this paper propose?	61
Experiments	What experiment does this paper carry out to evaluate the result?	654
Experiments	What does the result of this paper show?	1171
Experiments	How does this result outperform existing work?	542

Table 2: Dataset Analysis

By this way, we take the task as a text classification problem. When we meet an abstract, we separate it into sentences and classify every sentences into a category of questions' candidate answer. After generating the candidate answers, we select continuous words as the final answer as discussed in the next section.

We use the fastText (Joulin et al. 2017) model to deal with sentence classification, which uses a bag of n-grams as features and the hierarchical softmax as the linear classifiers. We use a learning rate of 0.1 to train models. We set size of word vectors to 100 and found that model performance is not sensitive to the size of word vectors. The training with 12 threads requires less than 100 epochs to converge and it in general takes only less than 1 minute. Table 3 shows some sentences' predicted labels.

Word-level sequence tagging

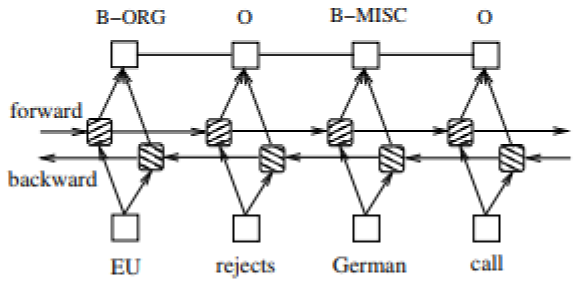


Figure 4: biLSTM-CRF model

Based on the sentence classification, we tag each word a code corresponding to a detailed question. In such sequence tagging task, we have access to both past and future input features for a given time, we can thus utilize a bidirectional LSTM network. In doing so, we can efficiently make use of past features (via forward states) and future features (via backward states) for a specific time frame. We train bidirectional LSTM networks using backpropagation through time (BPTT). The forward and backward passes over the unfolded network over time are carried out in a similar way to regular network forward and backward passes, except that we need to unfold the hidden states for all time steps. We also need a special treatment at the beginning and the end of the data points. Then we use the CRF networks to make uses of neighbor tag information in predicting current tags. Though this biLSTM-CRF model (Huang, Xu, and Yu 2015) reaches 86.7% F1 score on our dataset, it still remains space for improvements due to the lack of context information and prior knowledge.

Website

In order to show some real applications of our dataset and model, we build a website on Acemap called AceNews. AceNews extracts important information of newest arXiv machine reading papers, and presents comprehension(both in English and in Chinese) of these papers. Moreover, we recommend papers for different users based on the information extracted and user behaviours. All of the above is accomplished automatically by machine every day. The construction of our website is introduced below, following four steps.

LatestPpapers from arXiv.org

To further validate and demonstrate our trained model, we want to get more recent papers about machine learning from arXiv.org, a website with many newest papers and test our model on these papers.

we use Python crawler packages, mainly requests and lxml, to crawl and parse arXiv pages which show the latest papers about computer science. Requests is a useful Python HTTP client library that can get server response data, and lxml is a library used for parsing and generating xml files.

Sentence	Prediction	Probability
While deep reinforcement learning has successfully solved many challenging control tasks, its real-world applicability has been limited by the inability to ensure the safety of learned policies.	others	0.931695
We propose an approach to verifiable reinforcement learning by training decision tree policies, which can represent complex policies (since they are nonparametric), yet can be efficiently verified using existing techniques (since they are highly structured).	purpose	0.787732
The challenge is that decision tree policies are difficult to train.	others	1.000000
We propose VIPER, an algorithm that combines ideas from model compression and imitation learning to learn decision tree policies guided by a DNN policy (called the oracle) and its Q-function, and show that it substantially outperforms two baselines.	methods	0.973163
We use VIPER to (i) learn a provably robust decision tree policy for a variant of Atari Pong with a symbolic state space, (ii) learn a decision tree policy for a toy game based on Pong that provably never loses, and (iii) learn a provably stable decision tree policy for cart-pole.	others	0.568271
In each case, the decision tree policy achieves performance equal to that of the original DNN policy.	experiments	0.961477

Table 3: Sample sentences and their predicted results

First, we use requests to request an arXiv web page to get the content, and then use lxml to parse them and build a xml tree. Next we use the xpath method to find specific content of each paper. Finally we store the title, author, subject, release date, and link of each paper, and then select papers related to machine learning, whose subject is cs.[CV,AI,CL,LG,NE], as the data for next validation and display.

Running Models on Newest Papers

After retrieving arXiv papers, we do some preprocessing to these papers, including sentence-level and word-level segmentation. We then run our pretrained models on these papers to generate the answers to questions in each paper. We save these answers in the database.

Autotranslate

We write a script to help us autotranslate answers generated by our models on newest papers. We input English answers, and Chinese answers are automatically saved to our database.

AceNews

The construction of AceNews is completely an automatic process. We use PHP to construct this website. Every day, machine automatically retrieves the newest papers on arXiv, and then there is an auto-testing process based on our pretrained models to generate answers, after autotranslating, both English and Chinese answers in the newest paper is presented online. Moreover, we provide a pdf link of original arXiv paper, so users can jump to the papers they are interested in.

Conclusion

In this paper, we provide *PaperQA*, a QA dataset on academic paper abstracts, which contains more than 1,000 abstracts and 8,000 question-answer pairs. Then we propose a two-level framework to tackle this machine reading comprehension problem, and our model's performance is closed to human performance. We have made our dataset available freely to encourage more expressive models. Finally we build a website to interactively display the newest crawled arxiv paper abstracts and the important information answered by our model. The test result on the newest arxiv papers shows our model's generalization and robustness. Since the release of our dataset, we have already seen considerable interest in building models on this dataset, and the gap between our logistic regression model and human performance has more than halved. We expect that the remaining gap will be harder to close, but that such efforts will result in significant advances in reading comprehension.

References

[1] Dernoncourt, F., and Lee, J. Y. 2017. Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts. *CoRR* abs/1710.06071.

[2] Hermann, K. M.; Kociský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. *CoRR* abs/1506.03340.

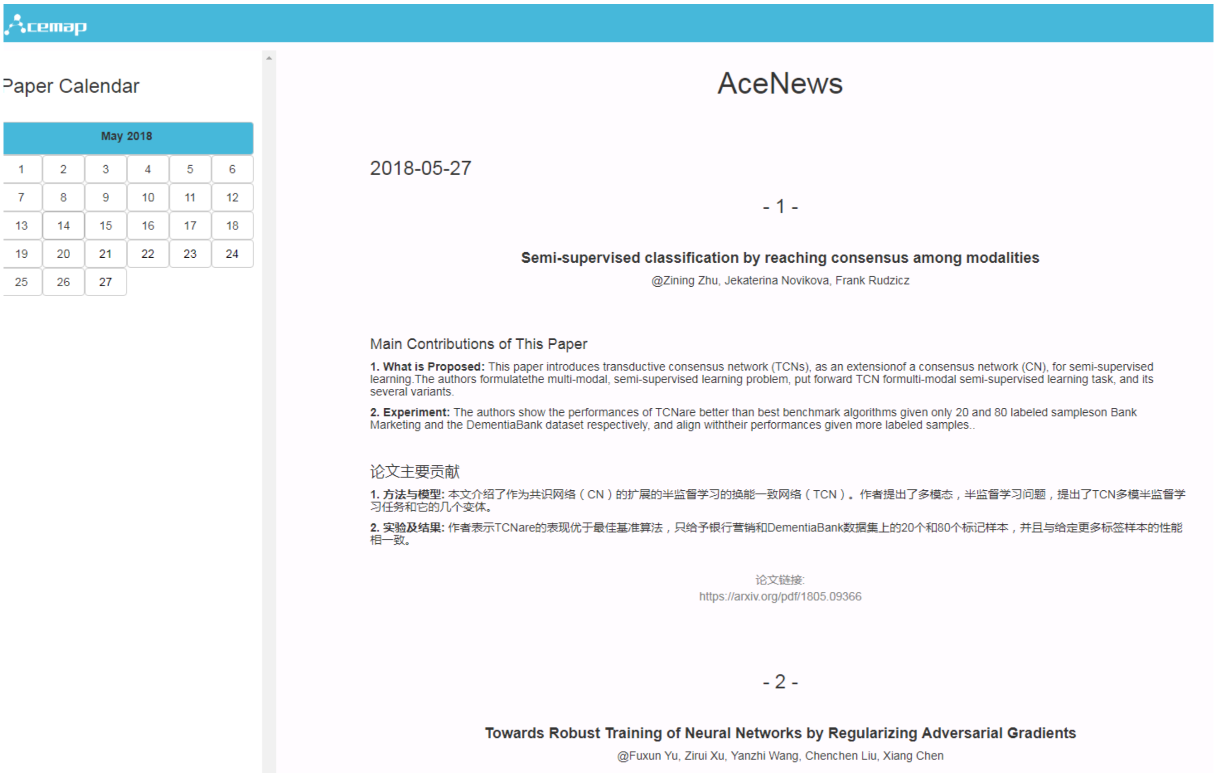
[3] Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *CoRR* abs/1511.02301.

[4] Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR* abs/1508.01991.

[5] Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Association for Computational Linguistics.

[6] Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR* abs/1606.05250.

[7] Richardson, M.; Burges, C. J. C.; and Renshaw, E. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text.



Paper Calendar

May 2018

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27			

AceNews

2018-05-27

- 1 -

Semi-supervised classification by reaching consensus among modalities

@Zining Zhu, Jekaterina Novikova, Frank Rudzicz

Main Contributions of This Paper

1. What is Proposed: This paper introduces transductive consensus network (TCNs), as an extensionof a consensus network (CN), for semi-supervised learning.The authors formulatethe multi-modal, semi-supervised learning problem, put forward TCN formulti-modal semi-supervised learning task, and its several variants.

2. Experiment: The authors show the performances of TCNare better than best benchmark algorithms given only 20 and 80 labeled sampleson Bank Marketing and the DementiaBank dataset respectively, and align withtheir performances given more labeled samples..

论文主要贡献

1. 方法与模型: 本文介绍了作为共识网络（CN）的扩展的半监督学习的换能一致网络（TCN）。作者提出了多模态，半监督学习问题，提出了TCN多模半监督学习任务和它的几个变体。

2. 实验及结果: 作者表示TCNare的表现优于最佳基准算法，只给予银行营销和DementiaBank数据集上的20个和80个标记样本，并且与给定更多标签样本的性能相一致。

论文链接:
https://arxiv.org/pdf/1805.09366

- 2 -

Towards Robust Training of Neural Networks by Regularizing Adversarial Gradients

@Fuxun Yu, Zirui Xu, Yanzhi Wang, Chenchen Liu, Xiang Chen

Figure 5: Acenews.