applied sciences

MDPI

Article

PaperNet: A Dataset and Benchmark for Fine-Grained Paper Classification

Tan Yue, Yong Li, Xuzhao Shi, Jiedong Qin, Zijiao Fan and Zonghai Hu *

School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; yuetan@bupt.edu.cn (T.Y.); yli@bupt.edu.cn (Y.L.); sxzs@bupt.edu.cn (X.S.); qjd2020@bupt.edu.cn (J.Q.); fzj6@bupt.edu.cn (Z.F.)

* Correspondence: zhhu@bupt.edu.cn; Tel.: +86-010-6228-3467

Abstract: Document classification is an important area in Natural Language Processing (NLP). Because a huge amount of scientific papers have been published at an accelerating rate, it is beneficial to carry out intelligent paper classifications, especially fine-grained classification for researchers. However, a public scientific paper dataset for fine-grained classification is still lacking, so the existing document classification methods have not been put to the test. To fill this vacancy, we designed and collected the PaperNet-Dataset that consists of multi-modal data (texts and figures). PaperNet 1.0 version contains hierarchical categories of papers in the fields of computer vision (CV) and NLP, 2 coarse-grained and 20 fine-grained (7 in CV and 13 in NLP). We ran current mainstream models on the PaperNet-Dataset, along with a multi-modal method that we propose. Interestingly, none of these methods reaches an accuracy of 80% in fine-grained classification, showing plenty of room for improvement. We hope that PaperNet-Dataset will inspire more work in this challenging area.

Keywords: artificial intelligence application; dataset; multi-modal information processing; machine learning; paper classification

Citation: Yue, T.; Li, Y.; Shi, X.; Qin, J.; Fan, Z.; Hu, Z. PaperNet: A Dataset and Benchmark for Fine-Grained Paper Classification. *Appl. Sci.* **2022**, *12*, 4554. https://doi.org/10.3390/app12094554

Academic Editor: Evgeny Nikulchev

Received: 10 April 2022
Accepted: 25 April 2022
Published: 30 April 2022

1. Introduction

The number of scientific papers has been increasing ever more rapidly [1–5]. Researchers have to spend a lot of time classifying papers relevant to their studies, especially into fine-grained sub-fields. When the number of papers in a field reaches the order of 10^4, it becomes really hard to track them manually. Therefore, intelligent fine-grained paper-classification is highly desirable.

However, a lot of existing paper classification models are coarse-grained [6,7]; i.e., the classification model simply divides papers into few large fields such as "math", "chemistry", "physics", and "biology". In sub-fields such as "machine translation" or "text generation", the cost of data labeling could be quite high because scientific expertise is usually needed. As a result, fine-grained paper classification datasets are lacking. Additionally, papers usually contain text and figure information. This may not be sufficient to classify paper documents based only on texts [8].

To address this issue, we introduce the PaperNet-Dataset, which includes 12 datasets and multi-modal data, and ran experiments on it using current mainstream classification models. PaperNet contains 2 coarse-grained (CV and NLP) and 20 fine-grained (7 in CV and 13 in NLP) classes.

The main contributions of this paper are summarized as follows:

- We introduce PaperNet-Dataset, which contains multi-modal data (text and figure) for fine-grained paper classification. To the best of our knowledge, this is the first multi-modal fine-grained paper dataset. In addition, it was pre-processed for convenience of use.

- Extensive experiments using current mainstream models were conducted to evaluate PaperNet. None of them reached an accuracy of 80%. This shows that fine-grained paper classification is a challenging task and PaperNet could be used as a worthy benchmark.
- Additionally, we propose a multi-modal paper classification method as a potential direction for better performance. The proposed method combines the strengths of MobileNetV3 and Albert for multi-modal representation fusion and shows promising results.

2. Background and Related Works

2.1. Related Datasets

Reuters dataset [9] contains lots of short news and corresponding topics. There are 10,789 samples and 90 classes in the dataset. Cifar-10 is an image classification dataset for identifying universal objects. The dataset contains 10 categories of RGB color images: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. The sizes of images are 32 × 32, and there are 50,000 training images and 10,000 test images in the dataset. CUB-200 is a multi-modal fine-grained classification dataset and contains 11,788 images of birds, including 200 classes of birds, among which 5994 images are in the training dataset and 5794 images are in the test dataset. Each sample provides information about image tags, bird attributes, and image description.

In terms of paper classification dataset, AAPD (Arxiv Academic Paper Dataset) [10] is a large dataset in the field of computer science for multi-label text classification. There are 55,840 papers, including abstracts and corresponding topics, with a total of 54 classes. The DocFigure dataset [11] consists of 33K annotated figures of 28 different categories present in scientific articles published in the CVPR, ECCV, ICCV, etc., conferences in the last several years. The details of related datasets including the average word count in each sample are shown in Table 1.

Due to a large number of categories of scientific papers, researchers more frequently need to locate papers in subdivisions of their research fields. Compared with coarse-grained classification, the sample similarity in the sub-fields is higher, making classification more difficult. Another aspect of scientific papers is that they contain multi-modal information. Figures are usually indispensable besides texts. It is insufficient to classify paper based only on texts. Therefore, multi-modal document classification in the paper field is a potential way to increase accuracy in fine-grained paper classification.

However, because scientific expertise is usually needed, the cost of sample labeling can be quite high. As a result, fine-grained paper classification datasets are lacking. Here, we introduce PaperNet-Dataset, a multi-modal paper classification dataset.

Table 1. The details of related datasets.

Modality	Dataset	Class	Detail Sample	Word
Text	Reuters	90	10,789	144.3
	AAPD	54	55,840	167.3
	IMDB	10	135,669	393.8
Figure	Cifar-10	10	60,000	-
	DocFigure	28	33,000	-
	Deepchart	5	5000	-
Multi-modal	CUB	200	11788	-
	Food-101	101	90,704	-
	PaperNet v1.0	20	38,608	150.43

2.2. Multi-Modal Learning

Modalities are different ways in which people receive information. Researchers have achieved remarkable research results in the field of multi-modal learning [12,13].

Multi-modal learning aims to integrate multiple modal information to obtain a consistent and common model output. The fusion of multi-modal information can obtain more comprehensive features, improve the robustness of the model, and ensure that the model can work efficiently even when some modes are missing.

Ref. [14] proposed a new multi-modal event topic model to model the social media documents. Ref. [15] proposed a new hashing algorithm, which integrates the multi-modal features extracted by weak supervision into binary code, thus using the kernel function and SVM for classification. Ref. [16] built their fusion layer by the outer product instead of simple concatenation in order to obtain more features.

2.3. Paper Document Classification

Paper classification belongs to document-level text classification. Compared with other document-level text classification tasks, paper classification contains lots of levels of categories, which are from coarse-grained to fine-grained. Paper documents consist of image information and text information. As for text information, we could get richer semantic information from title to abstract to full document content. Ref. [17] proposed XMLCNN which is based on the popular model [18]. Another popular model, Hierarchical Attention Network [19], models Hierarchical information of documents to extract meaningful features and classify documents in combination with word level and sentence level encoders. Nguyen et al. [20] proposed an improved feature weighting technique for document representation. SGM [10] is a generative method for classifying multi-label documents, which uses encoder-decoder sequence generation model to generate labels for each document. Ref. [21] proposed a simple and properly-regularized single-layer BiLSTM.

Recently, a large number of studies have shown that the "pre-trained model" based on large corpus can learn general language representations, which is beneficial to downstream text classification tasks and can avoid training the model from scratch. Some pre-trained model focuses on learning "context-sensitive word embedding", such as Elmo [22], OpenAI GPT [23], and Bert [24]. These learned encoders are also used to represent words in downstream tasks. In addition, various pre-trained tasks have been proposed to learn pre-trained model for different purposes.

DocBERT [7] presents the first application of BERT to document classification. As for multi-modal classification, some popular visual question answering (VQA) and Image caption methods such as VisualBERT and UNITER cannot be directly applied. The method used in [25] first studied the relation between the textual and visual aspects in multi-modal posts from major social media platforms.

3. Method

Inspired by [25] and DocBERT [7], we propose a multi-modal paper classification method. More modal information is extracted from scientific paper documents. We also combine the complementary strengths of MobileNetV3 and Albert for better multi-modal information joint representation.

3.1. Figure Feature Representation

The figures are put into the pre-trained MobileNetV3 model for feature extraction. Here we fine-tune the MobileNetV3 to achieve better results for this task. The extracted figure vector is subsequently fused with the text vector of the paper.

$$\mathcal{V}_i^m = MobileNetV3(\mathcal{I}_i) \tag{1}$$

The figure features $\mathcal{F}_i^m = [I_{i1}^m, I_{i2}^m, I_{i3}^m, \ldots, I_{in}^m]$, $\mathcal{F}_i^m \in \mathcal{V}_i^m$. Where I_{ik}^m means the figure feature, $\mathcal{V}_i^m$ means the feature vector of encoded figure.

3.2. Text Feature Representation

For the text of the paper, we first vectorize the text information. The Albert pre-processing model is used for embedding. Then, the Albert pre-trained model is used for encoding and feature extraction.

$$\mathcal{V}_i^t = Albert(\mathcal{T}_i) \tag{2}$$

The text features $\mathcal{F}_i^t = [T_{i1}^t, T_{i2}^t, T_{i3}^t, ..., T_{in}^t]$, $\mathcal{F}_i^t \in \mathcal{V}_i^t$. Where T_{ik}^t means the text feature, $\mathcal{V}_i^t$ means the feature vector of text.

3.3. Multi-Modal Feature Fusion

After Albert processing, the text vectors are concatenated with the weighted figure vectors.

$$\mathcal{V}_i^f = \mathcal{V}_i^t \oplus (\mathcal{V}_i^m \times W) \tag{3}$$

where $\oplus$ is the concatenation operator, W is the weighting matrix. After the concatenation, the fully connected layer is used for the multi-modal information joint representation. Then, we combine the text vector, the figure vector, and the fusion vector for classification.

$$\mathcal{V}_i^f = ReLU(w \cdot \mathcal{V}_i^f + b) \tag{4}$$

$$\mathcal{V} = w_1 \cdot \mathcal{V}_i^f + w_2 \cdot \mathcal{V}_i^t + w_3 \cdot \mathcal{V}_i^m \tag{5}$$

where w is the weighting coefficient of each vector.

3.4. Classification

We use a two-layer fully-connected neural network as our classification layer. The activation function of the hidden layer and the output layer are element-wise ReLu and softmax functions, respectively. The loss function is cross-entropy.

We show that this modification significantly helps the optimization of this model and outperforms other baselines significantly in most cases.

4. Dataset

4.1. Papernet-Dataset

More than 38,000 samples were collected from "Google Scholar" and well-known NLP and CV conferences such as ACL, EMNLP, CVPR, etc. Because of the academic nature of the papers, we invite scholars or experts who have researched for some years in the related field to label the samples. In our experiments, the dataset is divided with 80% of samples as the training set and validation set. These samples are shuffled and the 10-fold cross validation technique is used. The remaining 20% of the samples are used as the test set. PaperNet contains following subsets.

- **PaperNet_2.** PaperNet_2 dataset is a coarse-grained paper classification dataset and contains 2 classes, CV and NLP.
- **PaperNet_20.** PaperNet_20 dataset is for fine-grained paper classification. It contains 20 classes, 7 in CV and 13 in NLP.
- **PaperNet_CV.** PaperNet_CV dataset is a subset of the PaperNet_20 dataset. The dataset includes 7 classes.
- **PaperNet_NLP.** PaperNet_NLP dataset is another subset of the PaperNet_20 dataset which includes 13 classes.

Each of the above 4 datasets contains text, figure, and multi-modal subsets. So, there are 12 datasets in PaperNet 1.0 version. The details are shown in Table 2.

Table 2. Details of PaperNet datasets (Multi. means the pre-processed multimodal data).

Dataset	Subset	Class	Text	Figure	Multi.
PaperNet_2		CV	1863	12,774	25,548
		NLP	2538	6609	13,060
Average		Coarse-grained class	2200.5	9691.5	19,304
PaperNet_20	PaperNet_CV	CV_attention	201	1157	2314
		CV_classification	387	1951	3902
		CV_detection	621	4280	8560
		CV_GAN	228	1932	3864
		CV_recognition	284	1432	2864
		CV_retrieval	82	270	540
		CV_segmentation	260	1752	3504
	PaperNet_NLP	NLP_bert	372	1215	2430
		NLP_conversation	262	698	1396
		NLP_cross	277	576	1152
		NLP_extraction	340	681	1362
		NLP_Few_shot	119	309	614
		NLP_knowledge_graph	82	244	488
		NLP_machine_reading	59	104	208
		NLP_machine_translation	490	832	1664
		NLP_multilingual	253	557	1114
		NLP_multimodal	145	829	1656
		NLP_named_entity_recognition	105	189	378
		NLP_sentiment_analysis	172	179	358
		NLP_text_generation	100	121	240
Average		Fine-grained class	241.95	965.4	1930.4

4.2. Data Pre-Processing and Feature Engineering

Because of the scientific expertise of the paper, the cost of sample labeling is extremely high when the paper is classified in a fine-grained way. To solve the problem of a limited number of samples, we extract the figure information and the abstract information for information expansion.

Due to the need for data transmission and storage, most scientific papers exist in PDF format including text content and figures. We used the PDFplumber framework to extract the text content of the abstract part of the paper in PDF, and use the PIL framework to extract the figures. Figure extraction is a challenge because of the size and format of figures in PDF documents. Additionally, some figures contain useless features such as the logo or biography figures which need to be removed in the figure extraction step. Therefore, we modify the PIL framework for better extraction performance. We resize the images and use ResNet for feature extraction. Next, for the text abstract of the paper, we vectorize the text information. TF-IDF algorithm and word2vector algorithm are used for encoding and feature extraction. The details of the paper dataset are shown in Table 2.

5. Experiment

In this section, we evaluate our PaperNet-Dataset with a series of experimental tasks.

5.1. Algorithms

We compare our proposed multi-modal fusion method with multiple well-known classification and embedding methods such as:

- **ResNet50:** Residual Network [26] is widely used in the image classification field and as part of the backbone of neural networks in computer vision tasks.
- **DenseNet121:** The DenseNet model [27] alleviates the problem of gradient disappearance, strengthens feature propagation, and reduces the number of parameters.

- **MobileNetV3:** MobileNetV3 [28] is a light-weight model. Combined with network design and NAS technology, a new generation of MobileNets is proposed.
- **ULMFiT:** This is a general language model based on fine tuning (ULMFiT), which can be applied to a variety of tasks in NLP [29].
- **Albert:** Albert model from paper [30]. Compared to BERT, Albert achieves better results with fewer parameters. We use the Albert model as part of our proposed model for text encoding.
- **Concat:** Previous work [25] concatenates different feature vectors of different modalities as the input of the classification layer. We implement this concatenation model with our feature vectors of different modalities and apply it for classification.

5.2. Settings

For baseline models, we used default parameter settings as in their original papers or implementations and add the dropout layer to the pre-trained models with a dropout rate of 0.3. The learning rate is 1×10^{-4}. We selected 10% of the training set as the validation set and the 10-fold cross validation technique was used. Following [31], we trained the model using Adam [32] and stopped training if the validation loss did not decrease for 10 consecutive epochs.

5.3. Main Results

We aimed to use popular models to evaluate the PaperNet-Dataset and set up benchmarks for paper classification. The results are shown in Figures 1 and 2.

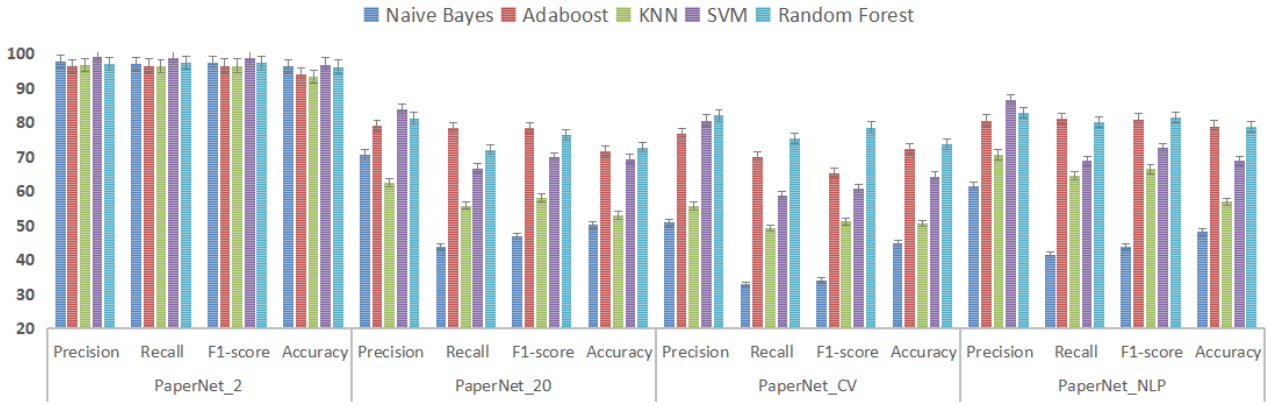


Figure 1. The machine learning algorithm experiments.

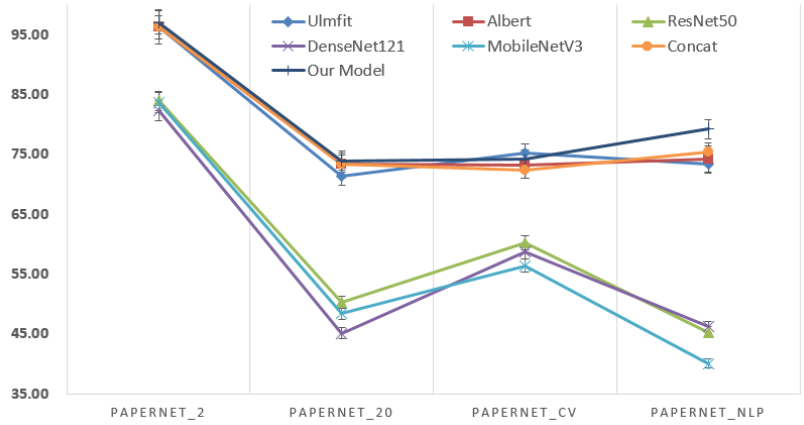


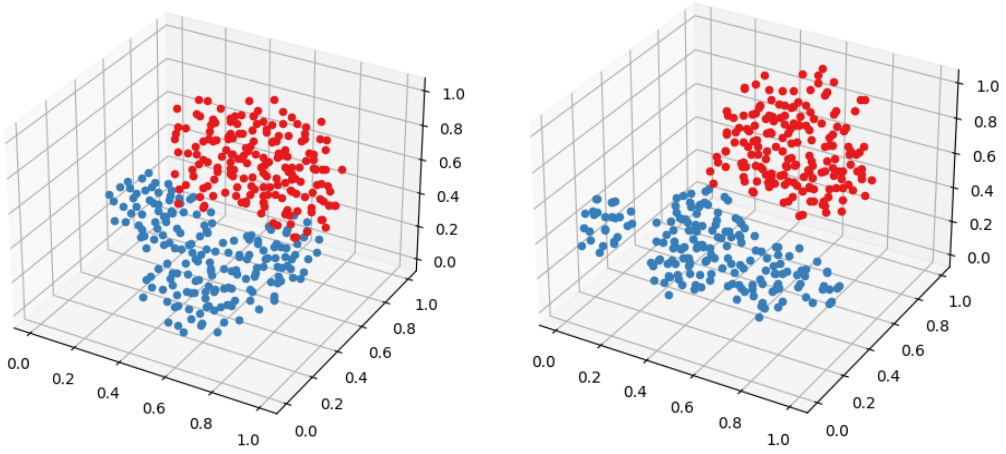
Figure 2. Popular pre-trained model accuracy experiments.

5.3.1. Text Classification

In the experiments, we first evaluated text classification models. The Ulmfit and Albert model perform well on PaperNet_2 dataset, which is coarse-grained. As for fine-grained classification, the accuracy of the two models dropped significantly.

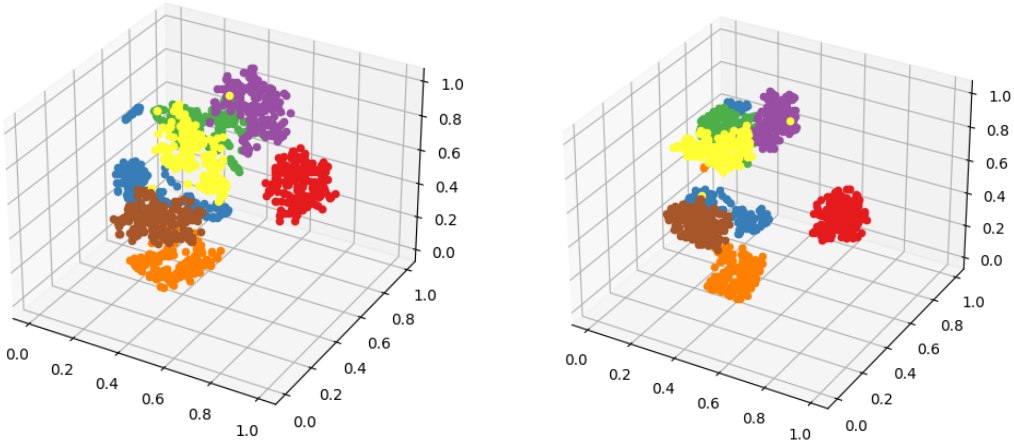
5.3.2. Image Classification

Next, we used popular image classification models to conduct experiments on four datasets. To our surprise, all the models did not perform well on fine-grained datasets. It is more difficult to carry out fine-grained classification because of the high similarity in the subdivided fields. Details of our further analysis can be found in Figures 3 and 4. The typical figures in datasets are shown in Figures 5 and 6.



(**a**) MobileNetV2 (**b**) Our Model

Figure 3. Visualization of the classification results on the PaperNet_2 dataset. (Red: NLP; Blue: CV.)



(**a**) Albert (**b**) Our Model

Figure 4. Visualization of the classification results on the PaperNet_CV dataset. (Red: CV_attention; Yellow: CV_classification; Green: CV_segmentation; Brown: CV_retrieval; Purple: CV_recognition; Blue: CV_detection; Orange: CV_GAN.)

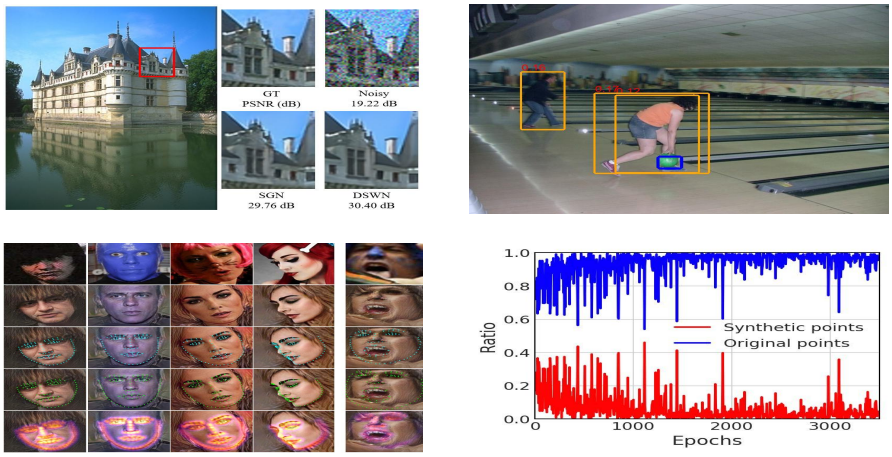


Figure 5. Examples of the PaperNet_CV dataset figures.

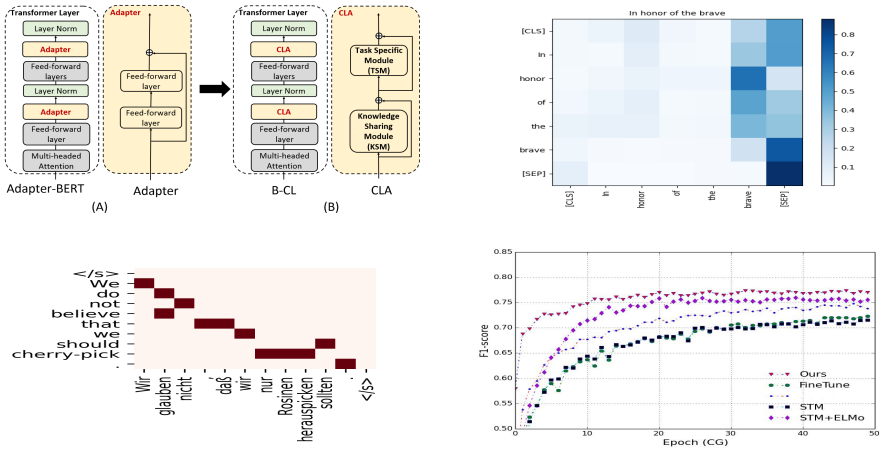


Figure 6. Examples of the PaperNet_NLP dataset figures.

5.3.3. Multi-Modal Classification

Since the single-modal classification methods can not achieve satisfactory results, we consider using multi-modal classification methods. Some popular VQA and Image caption methods such as VisualBERT or UNITER are not suitable for the multi-modal fine-grained paper classification task. We used the model proposed by [25] for multi-modal paper classification. Additionally, in order to make the model more suitable for paper classification tasks, we improved the model and propose our method introduced in Section 3. As shown in Table 3, the proposed method achieves the highest accuracy in three datasets.

Table 3. Comparison of accuracy.

Modality	Algorithm	PaperNet_2	PaperNet_20	PaperNet_CV	PaperNet_NLP
Image	ResNet50	83.94 ± 0.55	50.30 ± 0.66	60.23 ± 0.61	45.33 ± 0.50
	DenseNet121	82.34 ± 0.96	45.16 ± 0.46	58.80 ± 0.40	46.21 ± 0.46
	MobileNetV3	83.66 ± 0.86	48.38 ± 0.56	56.45 ± 0.40	40.08 ± 0.54
Text	Ulmfit	96.26 ± 0.44	71.30 ± 1.12	**75.30 ± 1.06**	73.33 ± 0.18
	Albert	96.31 ± 0.09	73.32 ± 0.24	73.18 ± 0.12	74.23 ± 0.36
Multi-modal	Concat	96.27 ± 0.11	73.45 ± 0.69	72.43 ± 0.27	75.36 ± 0.51
	Our method	**97.05 ± 0.05**	**73.85 ± 0.39**	74.26 ± 0.32	**79.27 ± 0.37**

5.4. Other Machine Learning Algorithms

As shown in Table 4, compared with popular pre-trained models, the Random Forest algorithm, the SVM algorithm, and the Naive Bayes algorithm surprisingly achieve higher accuracy on coarse-grained PaperNet_2 dataset. The algorithms consume very little time and computing resources. However, in fine-grained classification, the performance also drops significantly. More details of the results are shown in Appendix A.

Table 4. Performance comparison of machine learning algorithms.

Dataset	Metrics	Naive Bayes	Adaboost	KNN	SVM	Random Forest
PaperNet_2	Precision	97.91 ± 0.65	96.51 ± 0.56	96.87 ± 0.63	99.12 ± 0.62	97.23 ± 0.12
	Recall	97.20 ± 0.46	96.67 ± 0.32	96.42 ± 0.26	98.94 ± 0.33	97.58 ± 0.26
	F1-score	97.56 ± 0.23	96.66 ± 0.25	96.64 ± 0.21	98.96 ± 0.16	97.42 ± 0.22
	Accuracy	96.44 ± 0.36	94.08 ± 0.23	93.56 ± 0.43	96.98 ± 0.26	96.32 ± 0.16
PaperNet_20	Precision	70.75 ± 1.31	79.15 ± 0.39	62.64 ± 1.31	83.79 ± 1.21	81.32 ± 0.17
	Recall	43.98 ± 0.63	78.36 ± 0.32	55.89 ± 0.75	66.72 ± 0.69	72.15 ± 0.08
	F1-score	46.92 ± 0.39	78.46 ± 0.16	58.01 ± 0.87	69.97 ± 0.52	76.46 ± 0.09
	Accuracy	50.22 ± 0.59	71.76 ± 0.26	53.01 ± 0.68	69.43 ± 0.31	72.89 ± 0.21
PaperNet_CV	Precision	50.89 ± 1.29	76.81 ± 0.89	55.85 ± 0.85	80.67 ± 1.12	82.12 ± 0.06
	Recall	32.85 ± 0.64	70.02 ± 0.49	49.33 ± 0.64	58.67 ± 0.62	75.43 ± 0.15
	F1-score	34.11 ± 0.67	65.33 ± 0.28	51.22 ± 0.47	60.98 ± 0.21	78.64 ± 0.08
	Accuracy	44.96 ± 0.91	72.40 ± 0.63	50.67 ± 0.56	64.34 ± 0.49	73.86 ± 0.36
PaperNet_NLP	Precision	61.58 ± 1.16	80.68 ± 0.67	70.65 ± 0.83	86.56 ± 1.24	82.89 ± 0.18
	Recall	41.43 ± 0.87	81.29 ± 0.56	64.55 ± 0.65	68.85 ± 0.72	80.25 ± 0.16
	F1-score	43.81 ± 0.64	81.04 ± 0.51	66.51 ± 0.39	72.64 ± 0.51	81.55 ± 0.11
	Accuracy	48.16 ± 0.89	79.01 ± 0.32	56.97 ± 0.41	68.92 ± 0.55	78.89 ± 0.26

6. Conclusions and Future Work

To facilitate research on fine-grained paper classification, we introduce PaperNet-Dataset Version 1.0, which consists of multi-modal data (texts and figures). It contains hierarchical categories, 2 coarse-grained and 20 fine-grained (7 in CV and 13 in NLP). We ran multiple well known mainstream models on the PaperNet-Dataset. They performed poorly in the fine-grained tasks, never reaching an accuracy of 80%. In addition, we propose a multi-modal fusion method that increases the accuracy but still not satisfactory. The results show that there is plenty of room for improvement in fine-grained classification and PaperNet could be used as a benchmark dataset. We plan to expand PaperNet in future versions and hope that it will inspire more work in this challenging area of fine-grained paper classification.

Author Contributions: Conceptualization, T.Y.; Data curation, X.S. and J.Q.; Formal analysis, T.Y.; Funding acquisition, Z.H.; Investigation, T.Y. and Z.F.; Methodology, T.Y.; Project administration, T.Y.; Resources, Y.L. and Z.H.; Software, T.Y.; Supervision, Y.L. and Z.H.; Validation, T.Y.; Visualization, T.Y.; Writing—original draft preparation, T.Y.; Writing—review and editing, Y.L. and Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: The work described in this paper was supported by the BUPT innovation and entrepreneurship support program (022-YC-S002) and the Beijing Key Laboratory of Work Safety and Intelligent Monitoring.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CV	Computer Vision
NLP	Natural Language Processing
KNN	K-Nearest Neighbor
SVM	Support Vector Machine

Appendix A

Some detailed experiment results.

Table A1. Detailed results of the Naive Bayes algorithm.

Dataset	Class	Precision	Recall	F1-Score
PaperNet_2	CV	0.99	0.94	0.97
	NLP	0.96	0.99	0.98
PaperNet_20	CV_attention	0.60	0.15	0.55
	CV_classification	0.52	0.47	0.50
	CV_detection	0.42	0.93	0.58
	CV_GAN	0.68	0.47	0.55
	CV_recognition	0.82	0.38	0.52
	CV_retrieval	0.99	0.05	0.10
	CV_segmentation	0.61	0.21	0.31
	NLP_bert	0.46	0.66	0.54
	NLP_conversation	0.67	0.81	0.73
	NLP_cross	0.43	0.45	0.44
	NLP_extraction	0.49	0.62	0.55
	NLP_Few_shot	0.83	0.21	0.33
	NLP_knowledge_graph	0.99	0.40	0.57
	NLP_machine_reading	0.99	0.05	0.10
	NLP_machine_translation	0.67	0.88	0.76
	NLP_multilingual	0.65	0.52	0.58
	NLP_multimodal	0.67	0.53	0.59
	NLP_named_entity_recognition	0.99	0.20	0.33
	NLP_sentiment_analysis	0.70	0.40	0.51
	NLP_text_generation	0.89	0.40	0.55
PaperNet_CV	CV_attention	0.67	0.15	0.24
	CV_classification	0.52	0.41	0.46
	CV_detection	0.40	0.95	0.56
	CV_GAN	0.41	0.20	0.27
	CV_recognition	0.70	0.23	0.35
	CV_retrieval	0.41	0.21	0.27
	CV_segmentation	0.50	0.15	0.24
PaperNet_NLP	NLP_bert	0.42	0.65	0.51
	NLP_conversation	0.64	0.79	0.71
	NLP_cross	0.44	0.41	0.43
	NLP_extraction	0.46	0.62	0.52
	NLP_Few_shot	0.75	0.12	0.21
	NLP_knowledge_graph	0.99	0.25	0.40
	NLP_machine_reading	0.51	0.12	0.20
	NLP_machine_translation	0.57	0.86	0.69
	NLP_multilingual	0.58	0.42	0.49
	NLP_multimodal	0.71	0.50	0.59
	NLP_named_entity_recognition	0.80	0.20	0.32
	NLP_sentiment_analysis	0.69	0.31	0.43
	NLP_text_generation	0.50	0.13	0.20

Table A2. Detailed results of the Adaboost algorithm.

Dataset	Class	Precision	Recall	F1-Score
PaperNet_2	CV	0.96	0.96	0.96
	NLP	0.97	0.97	0.97
PaperNet_20	CV_attention	0.83	0.85	0.84
	CV_classification	0.93	0.84	0.88
	CV_detection	0.93	0.97	0.95
	CV_GAN	0.73	0.60	0.66
	CV_recognition	0.95	0.90	0.92
	CV_retrieval	0.95	0.90	0.92
	CV_segmentation	0.92	0.87	0.89
	NLP_bert	0.82	0.80	0.81
	NLP_conversation	0.81	0.85	0.83
	NLP_cross	0.76	0.79	0.77
	NLP_extraction	0.78	0.82	0.80
	NLP_Few_shot	0.59	0.67	0.63
	NLP_knowledge_graph	0.54	0.70	0.61
	NLP_machine_reading	0.91	0.50	0.65
	NLP_machine_translation	0.86	0.86	0.86
	NLP_multilingual	0.80	0.82	0.81
	NLP_multimodal	0.75	0.70	0.72
	NLP_named_entity_recognition	0.74	0.85	0.79
	NLP_sentiment_analysis	0.59	0.74	0.66
	NLP_text_generation	0.68	0.65	0.67
PaperNet_CV	CV_attention	0.52	0.99	0.68
	CV_classification	0.99	0.81	0.90
	CV_detection	0.92	0.98	0.95
	CV_GAN	0.50	0.49	0.50
	CV_recognition	0.99	0.05	0.10
	CV_retrieval	0.51	0.50	0.50
	CV_segmentation	0.88	0.99	0.94
PaperNet_NLP	NLP_bert	0.87	0.85	0.86
	NLP_conversation	0.96	0.83	0.89
	NLP_cross	0.87	0.93	0.90
	NLP_extraction	0.83	0.85	0.84
	NLP_Few_shot	0.69	0.75	0.72
	NLP_knowledge_graph	0.68	0.75	0.71
	NLP_machine_reading	0.56	0.45	0.50
	NLP_machine_translation	0.92	0.90	0.91
	NLP_multilingual	0.84	0.84	0.84
	NLP_multimodal	0.93	0.93	0.93
	NLP_named_entity_recognition	0.75	0.90	0.82
	NLP_sentiment_analysis	0.76	0.89	0.82
	NLP_text_generation	0.93	0.70	0.80

Table A3. Detailed results of the KNN algorithm.

Dataset	Class	Precision	Recall	F1-Score
PaperNet_2	CV	0.98	0.94	0.96
	NLP	0.96	0.98	0.97
PaperNet_20	CV_attention	0.24	0.23	0.23
	CV_classification	0.44	0.44	0.44
	CV_detection	0.57	0.75	0.65
	CV_GAN	0.44	0.60	0.50
	CV_recognition	0.62	0.55	0.58
	CV_retrieval	0.38	0.15	0.21
	CV_segmentation	0.54	0.40	0.46
	NLP_bert	0.62	0.64	0.63
	NLP_conversation	0.71	0.79	0.75
	NLP_cross	0.49	0.52	0.50
	NLP_extraction	0.68	0.60	0.64
	NLP_Few_shot	0.48	0.46	0.47
	NLP_knowledge_graph	0.88	0.75	0.81
	NLP_machine_reading	0.99	0.50	0.67
	NLP_machine_translation	0.73	0.85	0.79
	NLP_multilingual	0.66	0.54	0.59
	NLP_multimodal	0.77	0.67	0.71
	NLP_named_entity_recognition	0.80	0.60	0.69
	NLP_sentiment_analysis	0.57	0.60	0.58
	NLP_text_generation	0.92	0.55	0.69
PaperNet_CV	CV_attention	0.45	0.38	0.41
	CV_classification	0.53	0.47	0.50
	CV_detection	0.56	0.75	0.64
	CV_GAN	0.48	0.58	0.53
	CV_recognition	0.64	0.48	0.55
	CV_retrieval	0.70	0.35	0.47
	CV_segmentation	0.55	0.44	0.49
PaperNet_NLP	NLP_bert	0.60	0.62	0.61
	NLP_conversation	0.67	0.73	0.70
	NLP_cross	0.53	0.55	0.54
	NLP_extraction	0.71	0.66	0.69
	NLP_Few_shot	0.45	0.58	0.51
	NLP_knowledge_graph	0.89	0.80	0.84
	NLP_machine_reading	0.99	0.50	0.67
	NLP_machine_translation	0.72	0.83	0.77
	NLP_multilingual	0.57	0.50	0.53
	NLP_multimodal	0.82	0.90	0.86
	NLP_named_entity_recognition	0.80	0.60	0.69
	NLP_sentiment_analysis	0.64	0.66	0.65
	NLP_text_generation	0.90	0.45	0.60

Table A4. Detailed results of the SVM algorithm.

Dataset	Class	Precision	Recall	F1-Score
PaperNet_2	CV	0.97	0.99	0.98
	NLP	0.98	0.99	0.98
PaperNet_20	CV_attention	0.63	0.30	0.41
	CV_classification	0.63	0.80	0.70
	CV_detection	0.68	0.96	0.79
	CV_GAN	0.82	0.62	0.71
	CV_recognition	0.86	0.73	0.79
	CV_retrieval	0.99	0.05	0.10
	CV_segmentation	0.87	0.87	0.87
	NLP_bert	0.62	0.93	0.74
	NLP_conversation	0.93	0.79	0.85
	NLP_cross	0.75	0.77	0.76
	NLP_extraction	0.75	0.81	0.78
	NLP_Few_shot	0.88	0.29	0.44
	NLP_knowledge_graph	0.99	0.50	0.67
	NLP_machine_reading	0.99	0.30	0.46
	NLP_machine_translation	0.83	0.95	0.88
	NLP_multilingual	0.91	0.78	0.84
	NLP_multimodal	0.88	0.77	0.82
	NLP_named_entity_recognition	0.94	0.75	0.83
	NLP_sentiment_analysis	0.77	0.69	0.73
	NLP_text_generation	0.99	0.70	0.82
PaperNet_CV	CV_attention	0.75	0.30	0.43
	CV_classification	0.68	0.79	0.73
	CV_detection	0.62	0.96	0.75
	CV_GAN	0.87	0.58	0.69
	CV_recognition	0.84	0.68	0.75
	CV_retrieval	0.99	0.05	0.10
	CV_segmentation	0.89	0.75	0.81
PaperNet_NLP	NLP_bert	0.57	0.94	0.71
	NLP_conversation	0.93	0.79	0.85
	NLP_cross	0.76	0.73	0.75
	NLP_extraction	0.70	0.82	0.76
	NLP_Few_shot	0.90	0.38	0.53
	NLP_knowledge_graph	0.99	0.50	0.67
	NLP_machine_reading	0.99	0.10	0.18
	NLP_machine_translation	0.84	0.97	0.90
	NLP_multilingual	0.93	0.76	0.84
	NLP_multimodal	0.89	0.80	0.84
	NLP_named_entity_recognition	0.94	0.85	0.89
	NLP_sentiment_analysis	0.86	0.71	0.78
	NLP_text_generation	0.99	0.60	0.75

References

1. Zyuzin, V.; Ronkin, M.; Porshnev, S.; Kalmykov, A. Automatic Asbestos Control Using Deep Learning Based Computer Vision System. *Appl. Sci.* **2021**, *11*, 532. [CrossRef]
2. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [CrossRef]
3. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
4. Dhaliwal, S.S.; Nahid, A.A.; Abbas, R. Effective Intrusion Detection System Using XGBoost. *Information* **2018**, *9*, 149. [CrossRef]
5. Mukhamediev, R.I.; Symagulov, A.; Kuchin, Y.; Yakunin, K.; Yelis, M. From Classical Machine Learning to Deep Neural Networks: A Simplified Scientometric Review. *Appl. Sci.* **2021**, *11*, 5541. [CrossRef]
6. Ma, X.; Wang, R. Personalized Scientific Paper Recommendation Based on Heterogeneous Graph Representation. *IEEE Access* **2019**, *7*, 79887–79894. [CrossRef]

7. Adhikari, A.; Ram, A.; Tang, R.; Lin, J. DocBERT: BERT for Document Classification. *arXiv* **2019**, arXiv:cs.CL/1904.08398.
8. Quan, J.; Li, Q.; Li, M. Computer Science Paper Classification for CSAR. In *New Horizons in Web Based Learning*; Cao, Y., Väljataga, T., Tang, J.K., Leung, H., Laanpere, M., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 34–43.
9. Apté, C.; Damerau, F.; Weiss, S.M. Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst. (TOIS)* **1994**, *12*, 233–251. [CrossRef]
10. Yang, P.; Sun, X.; Li, W.; Ma, S.; Wu, W.; Wang, H. SGM: Sequence generation model for multi-label classification. *arXiv* **2018**, arXiv:1806.04822.
11. Jobin, K.; Mondal, A.; Jawahar, C. DocFigure: A dataset for scientific document figure classification. In Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, NSW, Australia, 22–25 September 2019; Volume 1, pp. 74–79.
12. Cadene, R.; Ben-younes, H.; Cord, M.; Thome, N. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
13. Zhu, J.; Zhou, Y.; Zhang, J.; Li, H.; Zong, C.; Li, C. Multimodal Summarization with Guidance of Multimodal Reference. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 9749–9756. [CrossRef]
14. Qian, S.; Zhang, T.; Xu, C.; Shao, J. Multi-Modal Event Topic Model for Social Event Analysis. *IEEE Trans. Multimed.* **2016**, *18*, 233–246. [CrossRef]
15. Xia, Y.; Zhang, L.; Liu, Z.; Nie, L.; Li, X. Weakly Supervised Multimodal Kernel for Categorizing Aerial Photographs. *IEEE Trans. Image Process.* **2017**, *26*, 3748–3758. [CrossRef]
16. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the Empirical Methods in Natural Language Processing, EMNLP, Copenhagen, Denmark, 7–11 September 2017.
17. Liu, J.; Chang, W.C.; Wu, Y.; Yang, Y. Deep learning for extreme multi-label text classification. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, 7–11 August 2017; pp. 115–124.
18. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:cs.CL/1408.5882.
19. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
20. Nguyen, D.B.; Shenify, M.; Al-Mubaid, H. Biomedical Text Classification with Improved Feature Weighting Method. In Proceedings of the International Conference on Bioinformatics and Computational Biology, Las Vegas, NV, USA, 4–6 April 2016.
21. Adhikari, A.; Ram, A.; Tang, R.; Lin, J. Rethinking complex neural network architectures for document classification. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4046–4051.
22. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
23. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:cs.CL/2005.14165.
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
25. Schifanella, R.; de Juan, P.; Tetreault, J.; Cao, L. Detecting Sarcasm in Multimodal Social Platforms. In *Proceedings of the 24th ACM International Conference on Multimedia (MM '16)*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1136–1145. [CrossRef]
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Chengdu, China, 15–17 December 2016; pp. 770–778.
27. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
28. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1314–1324.
29. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. *arXiv* **2018**, arXiv:1801.06146.
30. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2020**, arXiv:1909.11942.
31. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
32. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:cs.LG/1412.6980.