applied sciences

MDPI

Article

Generating Image Descriptions of Rice Diseases and Pests Based on DeiT Feature Encoder

Chunxin Ma [1,2,3,†], Yanrong Hu [1,2,3,*,†], Hongjiu Liu [1,2,3,*], Ping Huang [1,2,3,*], Yikun Zhu [1,2,3] and Dan Dai [1,2,3]

[1] College of Mathematics and Computer Science, Zhejiang A & F University, Hangzhou 311300, China; colorbeetlek@stu.zafu.edu.cn (C.M.); zhuyikun@stu.zafu.edu.cn (Y.Z.); d_dan1978@163.com (D.D.)
[2] Key Laboratory of Forestry Intelligent Monitoring and Information Technology Research of Zhejiang Province, Zhejiang A & F University, Hangzhou 311300, China
[3] Key Laboratory of Forestry Sensing Technology and Intelligent Equipment, State Forestry and Grassland, Zhejiang A & F University, Hangzhou 311300, China
* Correspondence: yanrong_hu@zafu.edu.cn (Y.H.); joe_hunter@zafu.edu.cn (H.L.); huangping@zafu.edu.cn (P.H.)
† These authors contributed equally to this work.

Abstract: We propose a DeiT (Data-Efficient Image Transformer) feature encoder-based algorithm for identifying disease types and generating relevant descriptions of diseased crops. It solves the scarcity problem of the image description algorithm applied in agriculture. We divided the original image into a sequence of image patches to fit the input form of the DeiT encoder, which was distilled by RegNet. Then, we used the Transformer decoder to generate descriptions. Compared to "CNN + LSTM" models, our proposed model is entirely convolution-free and has high training efficiency. On the Rice2k dataset created by us, the model achieved a 47.3 BLEU-4 score, 65.0 ROUGE_L score, and 177.1 CIDEr score. The extensive experiments demonstrate the effectiveness and the strong robustness of our model. It can be better applied to automatically generate descriptions of similar crop disease characteristics.

Keywords: agriculture; image captioning; rice pests and diseases; DeiT

Citation: Ma, C.; Hu, Y.; Liu, H.; Huang, P.; Zhu, Y.; Dai, D. Generating Image Descriptions of Rice Diseases and Pests Based on DeiT Feature Encoder. *Appl. Sci.* **2023**, *13*, 10005. https://doi.org/10.3390/app131810005

Academic Editors: Salik Khanal and Vitor Filipe

Received: 19 July 2023
Revised: 29 August 2023
Accepted: 2 September 2023
Published: 5 September 2023

1. Introduction

Rice is an indispensable food crop, with its total yield ranking third in the world. Various diseases and pests [1–4] seriously affect the yield of rice. The image caption generation technology (the goal of this technology is to enable computers to understand and describe the content of images in natural language) has the ability to automatically generate feature descriptions by capturing essential information in images. The application of this technology in the diagnosis of rice diseases and pests can not only help farmers identify the types of diseases and pests but also deepen their understanding of the characteristics of diseases and pests through the generated descriptions. However, there are still some difficulties in image caption generation technology for rice pests and diseases. First, due to the lack of rice disease and pest data samples, it is not suitable to use big data learning methods for model training. Second, the small proportion of the incidence site within the overall image presents a challenge. Convolution and pooling operations in Convolutional Neural Networks (CNNs) lead to a reduction in spatial resolution as the network deepens. The receptive field [5] becomes larger, which is beneficial for capturing larger structures but can make it harder to detect smaller details. Deep CNNs with numerous convolutional and pooling layers may fail to capture crucial information from small targets within the incidence site. Third, it is necessary to correctly handle the relationship between Computer Vision (CV) modal data and Natural Language Processing (NLP) modal data and build a cross-modal model for the high-level understanding of vision.

In previous studies, due to limitations in predefined sentence structures and phrases, template filling methods [6–8] have been unable to handle complex scenes, resulting in a

single description format and poor quality. The descriptions generated by retrieval [9,10] methods lack diversity and novelty because they rely on text library description datasets. In contrast, the descriptions generated by the approach based on the encoder–decoder structure [11] have higher accuracy and flexibility. The training process of this method is simpler and more efficient for researchers, so it is more popular. The encoder takes in the input data and processes it to extract important image features, and the decoder takes the encoded representation from the encoder and generates the descriptions. This method generally uses a CNN [12–16] as an encoder to automatically extract image features in the model, and the decoder relies on the advantages of an RNN (Recurrent Neural Network) [17] in processing sequential data to generate the predicted descriptions [11,18,19]. Long Short-Term Memory (LSTM), proposed by Vinyals et al. [20], solves the gradient vanishing problem of RNN. Thereafter, decoders mainly used LSTM and its variants [21–24]. However, there is a problem with insufficient detail and consistency in the description generation guided by global CNN features. Therefore, Xu et al. [25] introduced an attention mechanism between the encoder and decoder so that the model can selectively focus on certain regions of the image when generating each word. Lu et al. [26] further proposed an adaptive attention mechanism, which allows the model to choose whether to focus more on visual or semantic information in the process of generating descriptions. Anderson et al. [27] proposed an attention mechanism combining bottom-up and top-down approaches based on biological vision theory. Guo et al. [28] proposed using a GCN (Graph Convolutional Network) to obtain visual relationships to generate descriptions that fit the image more closely through the target semantic and spatial relations. It further narrowed the semantic gap between image and text. Xie et al. [29] used ResNet18 as the image feature encoder and a LSTM decoder with an attention mechanism to construct a model. They employed a CNN and attention-based LSTM architecture to generate image descriptions in the field of rice pests and diseases. Although its effect is good, there is still great room for improvement.

Recently, Transformer [30] has achieved great success in the field of NLP. Transformer model architecture has been considered for applications in the field of CV: Li et al. [31] used two independent Transformer encoders as the visual and semantic encoders of the model, respectively, to obtain more accurate visual and semantic relationships. Huang et al. [32] added the mechanism of AoA (Attention on Attention) to the Transformer to avoid misleading irrelevant information. The Meshed-Memory Transformer proposed by Cornia et al. [33] encodes image regions in a multi-level manner and uses the learned prior knowledge by persistent memory vector modeling, allowing the Transformer architecture to perform better on image description generation. However, all the above approaches require an additional CNN to extract image features. The ViT (Vision Transformer) [34] divides the image into image patches for input into the model, and it does not use any convolution operations but only the architectural model of Transformer for the image classification task. It demonstrates the remarkable ability of the Transformer in visual tasks without the need for a CNN encoder. Liu et al. [35] designed the first completely convolution-free architecture for image captioning using pre-trained ViT as an encoder and a standard Transformer decoder for caption generation. The DeiT [36] architecture proposed and used a teacher–student distillation training strategy for ViT, which solves the problem of poor results of ViT models due to insufficient data volume and enables Transformer to obtain superior results than the traditional convolutional methods in the vision domain. Overall, the training cost of these image captioning models for the Transformer architecture is higher than the CNN models.

Most of image captioning tasks use large public datasets and benchmarks such as Flickr8k [37], Flickr30k [38], and Microsoft COCO [39] to evaluate model performance. However, all of these datasets describe daily life scenarios. Therefore, there is a lack of mature datasets oriented to the image descriptions of rice pests and diseases. Although the accuracy of image description models trained on public datasets has been high, they cannot be directly applied to the image captioning task of rice pests and diseases.

In summary, we propose an image description generation method for rice pests and diseases based on the DeiT feature encoder, focusing on the image captioning task of 10 common rice pests and diseases. First, the traditional CNN + Attention + LSTM architecture is replaced by the architecture of a completely convolution-free Transformer. Second, the images are encoded using a distillation pre-trained DeiT feature encoder with stronger global information extraction ability than a traditional CNN. Then, a decoder of Transformer, with better parallelism and the ability to capture inter-sentence features, is used to generate descriptions. Finally, Rice2k, an image description dataset for rice diseases and pests, is constructed. The performance of the model is evaluated by comparative experiments.

2. Materials and Methods

2.1. Research Framework

With the depth and development of deep learning research, the research content is more complex for computer vision in agriculture. It has been extended from pest and disease diagnosis [40,41] to image description tasks. The research framework of the image captioning is shown in Figure 1. In the data acquisition and preprocessing module, images of 10 types of rice pests and diseases were obtained from Baidu's library (https://image.baidu.com/ (accessed on 15 March 2022)) and field shots. The model uses a sequence-to-sequence approach to segment the images into sequences of sub-image patches for input to the encoder, and then the encoding results are input to the decoder for decoding and prediction.

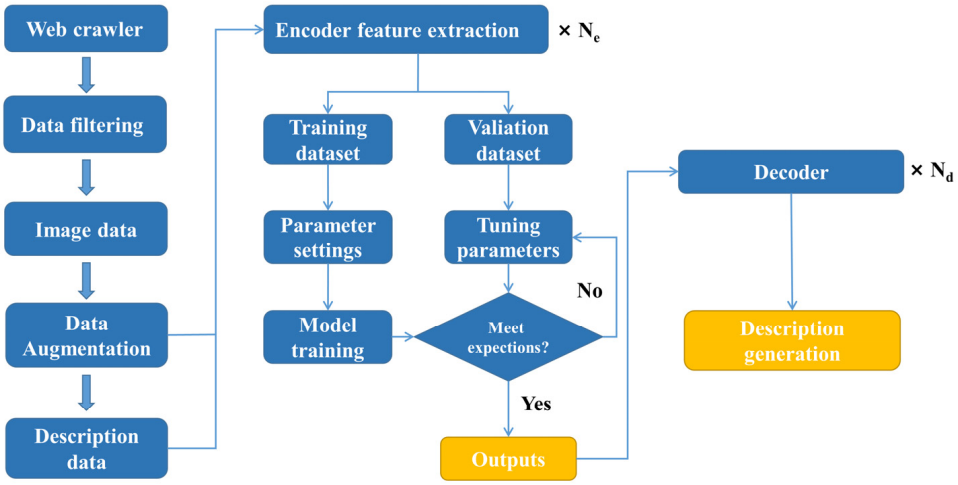

Figure 1. Research framework.

2.2. Dataset Construction

2.2.1. Data Collection

There are many types of rice pests and diseases in the growth of rice, and the distribution of the onset time is irregular. Thus, the 10 types of rice pests and diseases images required for the experiment could not be obtained in a short time. Consequently, we produced training data based on Baidu's library. In addition, most of the web images were taken at the typical time of disease onset, so the images of rice diseases and pests at the typical time were finally selected for the study. Compared with manual work, Python web crawler technology can collect more images in a short period of time and effectively reduce the pressure of data acquisition. Overall, 10 types of rice disease and pest samples were obtained from Baidu's image library to construct Rice2k. Rice2k includes 7 diseases and 3 insect pests in typical periods.

2.2.2. Data Preprocessing

In total, 4160 original samples of 10 types of rice pests and diseases were obtained after image acquisition. The images unrelated to pests and diseases were screened out, and the image set was normalized as follows:

1. The images were named by serial numbers and saved as .jpg format images.
2. The Python code implemented a perceptual hashing algorithm [42,43] (it generates compact digital signatures from multimedia data by emphasizing perceptually significant features to enable efficient content-based similarity comparisons between media files) to remove similar images without human intervention and reduce the workload of image selection.
3. The noisy data that could not meet the requirements were manually eliminated.

Only 373 original samples remained after normalization. In order to prevent the model from overfitting due to an insufficient sample size, we used data augmentation techniques to make the training data more suitable for the actual rice field growth environment. These techniques included brightness adjustment, flipping, and noise addition. After the data augmentation transformations, the sample images were effectively expanded, and the data diversity was enhanced. We obtained 2283 image samples, and the number of sample images for 10 types of rice pests and diseases was distributed as shown in Table 1. Rice2k was divided into an 8/10 training set, a 1/10 validation set, and a 1/10 testing set. The training set consisted of 1832 samples, the validation set consisted of 229 samples, and the testing set contained 222 samples. For evaluating the generalization ability of the model, we used 342 images of rice pests and diseases in actual field environments for testing.

Table 1. Image quantity distribution of typical rice pests and diseases.

Category	Latin Name	Number of Sample Images	Number of Actual Field Images
Rice bacterial blight	*Xanthomonas oryzae* pv. *Oryzae*	179	34
Rice bacterial streak disease	*Xanthomonas oryzicoia* pv. *Oryzicola*	372	45
Rice bakanae disease	*Fusarium moniliforme* Sheld	150	19
Rice three chemical borers	*Tryporyza incertulas* (Walker)	144	19
Rice brown spot	*Cochliobolus miyabeanus*	264	48
Rice planthopper	*Laodelphax striatellus* Fallén	252	34
Rice blast	*Pyricularia oryzae* Cavara	276	46
Rice false smut	*Ustilaginoidea oryzae*	240	41
Rice sheath blight	*Rhizoctonia solani*	298	39
Rice thrip	*Stenchaetothrips biformis*	108	17

2.2.3. Data Preprocessing

The image captioning dataset consists of two parts: the sample image and the five corresponding descriptions. In this paper, we annotated 5 descriptions for each image and saved them in JSON format. An example of the description part is shown in Table 2, where each sentence includes the diagnosis of the pests and diseases and a brief description of their features.

2.3. Experimental Platform

The experiments were conducted using Python 3.9.12 and PyTorch 1.10.0, running on 8 GB of RAM, with an Intel(R) Core(TM) i5-7300HQ CPU@ 2.50 GHz processor (The name of the manufacturer is Intel, and the equipment was sourced from Hangzhou, China), with an NVIDIA GeForce GTX1050 GPU (The name of the manufacturer is NVIDIA, and the equipment was sourced from Hangzhou, China) and an operating system of Windows 10, CUDA version 11.7. Both were optimized using the Adam (Adaptive momentum) algorithm. Dropout was added to prevent overfitting of the model. In the training parameters, the batch size was set to 64, the learning rate was 1×10^{-3}, and the gradient threshold

was set to 5. The model was trained through CPU for 20 rounds, which took 6.9 h, and the model test took 12 min.

Table 2. Example of rice pest and disease image descriptions.

Image	Manually Annotated Reference Sentences
 Bacterial blight	1. Bacterial blight of rice in which the leaves have yellow stripes 2. Rice bacterial leaf blight, rice leaf with yellow stripes 3. Bacterial blight of rice in which the leaves have yellow streaked lesions 4. Rice bacterial leaf blight, rice leaf lesions with yellow stripes. 5. Bacterial blight disease of rice in which the leaves show yellow stripe lesions.

3. Image Caption Model

3.1. Model Structure

The model framework of this paper is shown in Figure 2. We used the architecture of CPTR [34]. Initially, the original image is divided into image patches and flattened prior to being fed into the encoder. Subsequently, the visual features are extracted by the encoder and are input into the decoder. Lastly, the decoder is used to learn the mapping relationship between image features and syntactic features to generate the description text related to the content of pest and disease pictures. The model considers the image description as a sequence-to-sequence prediction task.

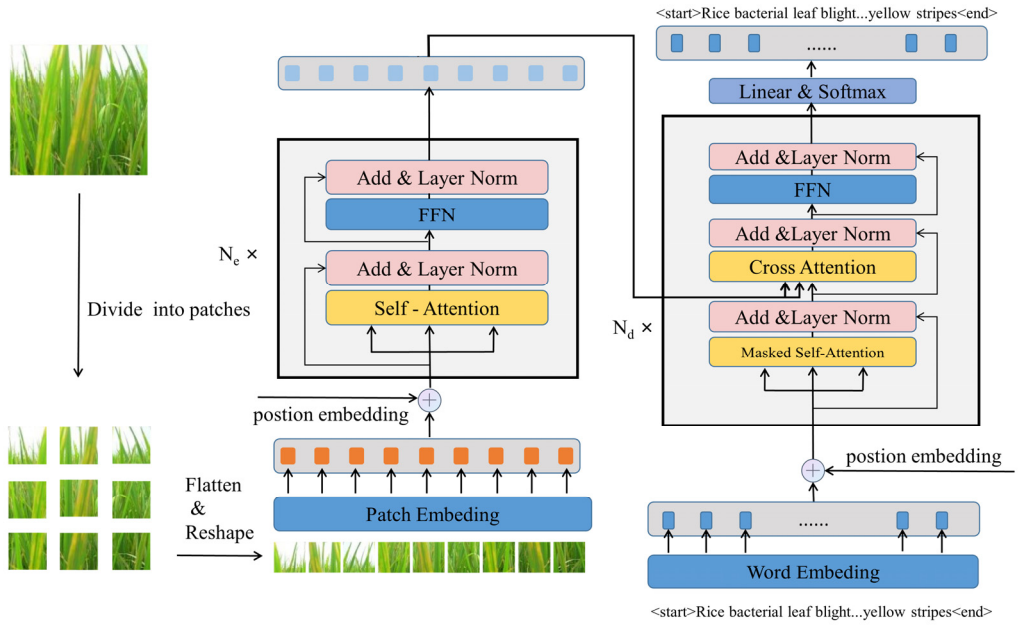

Figure 2. Rice pest and disease image description model.

3.1.1. DeiT

To compensate for the lack of data in the image description dataset of rice diseases and pests, we used the distillation pre-trained DeiT [36] as the encoder of this model. The DeiT model is small and effective. Using it as an encoder in the model can significantly reduce the training time of the model on the rice pest image description generation task and achieve great training results.

The DeiT model was used for the image classification task. It was trained by the teacher–student training strategy. The distillation training process is shown in Figure 3. The teacher model is typically selected as a high-performing image classification model, often a CNN-based model like RegNet. Meanwhile, the student model is based on the fundamental structure of ViT (Vision Transformer). Compared with the ViT model, the DeiT model has an additional distillation token through which the student model continuously extracts from the teacher model to learn the generalization ability of the teacher model.

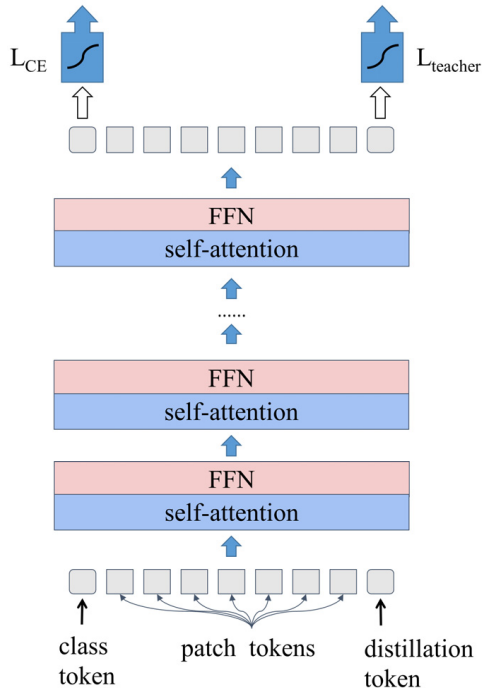

Figure 3. DeiT Model Distillation Training Process.

There are two types of distillation methods, soft distillation and hard-label distillation, of which soft distillation is:

$$\mathcal{L}_{global} = (1-\lambda)\,\mathcal{L}_{CE}(\varphi(Z_s), \mathcal{Y}) + \lambda\tau^2 KL(\varphi(Z_s/\tau), \varphi(Z_t/\tau)) \quad (1)$$

The loss is calculated from the output of the student model and the true label. The KL (Kullback–Leibler) divergence loss is calculated from the output of the student model and the output of the teacher model. λ is the coefficient that balances the KL divergence loss. The cross-entropy ($\mathcal{L}_{\mathrm{CE}}$) is on ground-truth label $\mathcal{Y}$, and φ is the softmax function. $\mathcal{L}_{\mathrm{global}}$ guides the whole distillation training process.

The hard-label distillation is:

$$\mathcal{L}_{global}^{hardDistill} = \frac{1}{2}\,\mathcal{L}_{CE}(\varphi(Z_s), \mathcal{Y}) + \frac{1}{2}\,\mathcal{L}_{CE}(\varphi(Z_s), \mathcal{Y}_t) \quad (2)$$

The first loss is calculated from the output of the student model and the true label. The second loss is calculated from the output of the student model and the teacher model. $\mathcal{L}_{global}^{hardDistill}$ is obtained by adding the two.

3.1.2. Image Feature Encoder

The encoder is DeiT, a hard distillation pre-trained image feature encoder whose distillation training teacher model is RegNet.

The input image is initially adjusted to a fixed resolution $X \in R^{H\times W\times 3}$, and then the adjusted image is divided into N patches, where $N = \frac{H}{P} \times \frac{W}{P}$. P represents the patch size, and we set P to 16. Subsequently, each patch is flattened in one dimension and reshaped

into a 1-dimensional patch sequence $X_p \in R^{N\times(P^2\cdot 3)}$. The flattened patch sequences are mapped to the latent space using a linear embedding layer, and a learnable 1-dimensional position embedding is added to the patch features, which gives the final input to the DeiT encoder: $P_a = [p_1, \cdots, p_N]$.

The image feature encoder DeiT consists of a stack of Ne identical layers (N_e = 4), each consisting of a multi-headed self-attentive (*MSA*) sublayer and a position feedforward (*FFN*) sublayer. *MSA* contains *H* heads, each head h_i corresponds to an independent scaled dot product attention function, enabling the model to jointly focus on different subspaces. The attention results of the different heads are then aggregated by a linear transformation W^O, the process of which can be expressed as:

$$MSA(Q,\ K,\ V) = ConcaI\ (h_1, \ldots, h_H)W^O \tag{3}$$

The scaled dot product attention is a special kind of attention proposed in Transformer model, which is calculated as:

$$Attention\ (Q, K, V) = Softmax\ (\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

where $Q \in R^{N_q\times d_k}$, $K \in R^{N_k\times d_k}$, and $V \in R^{N_v\times d_k}$ are the query matrix, the key matrix, and the value matrix, respectively.

The next positional feed-forward sublayer is implemented as two linear layers with *GELU* activation function and dropout between them to further transform the features. It can be expressed as:

$$FFN\ (x) = FC_2(Dropout(GELU(FC_1(x)))) \tag{5}$$

In each sublayer, there exists a sublayer connection consisting of residual connections, which are then layer-normalized.

$$x^{out}\ = LayerNorm\left(x^{in}\ + Sublayer\left(x^{in}\right)\right) \tag{6}$$

3.1.3. Decoder

At the decoder, a sinusoidal positional embedding of the word embedding features is performed. The result of the addition and the encoder output features are simultaneously used as inputs. The decoder consists of a stack of Nd layers (N_d = 4), and each layer contains a masked multi-headed self-attentive sublayer, a multi-headed cross-attentive sublayer, and a position-forecasting sublayer. Using the output features of the previous decoder layer, the next word is predicted by a linear layer with an output dimension equal to the size of the vocabulary. Given a basic truth sentence $y^*_{1:t-1}$ and the prediction y^*_t of a headline model with parameter θ and minimizing the following cross-entropy loss:

$$L_{XE}\ (\theta) = -\sum_{t=1}^{T} log(\ p_\theta(y^*_t \mid y^*_{1:t-1})) \tag{7}$$

3.2. Evaluation Metrics

The Cross-Entropy Loss function defines the effect of the neural network model and the goal of optimization. The convergence of the model is judged according to the decline of the loss function.

The image captioning task eventually generates the description text, and the evaluation metrics for natural language processing-related tasks are also applicable to it. Therefore, we used the following evaluation metrics for common natural language processing-related tasks to measure the performance of the model: BLEU (Bilingual Evaluation Understudy) [44], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [45], ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) [46], and CIDEr (Consensus-based Image Description Evaluation) [47]. The BLEU, including BLEU-1, BLEU-

2, BLEU-3, and BLEU-4, measures the similarity between machine-generated descriptions and human descriptions based on the number of consecutive words matched. The percentage of matched words is calculated to evaluate the effectiveness of the model's predictions. BLEU-1 assesses the accuracy of generated descriptions at the word level, while higher-order BLEU measures the coherence at the sentence level. METEOR evaluates the accuracy and recall of words in a sentence. It combines accuracy and recall metrics to provide an overall assessment. ROUGE-L calculates accuracy and recall based on the longest common subsequence, examining the adequacy of the generated descriptions. CIDEr focuses on importance and calculates the TF-IDF (Term Frequency-Inverse Document Frequency) weights of each n-gram. It measures the cosine similarity between the evaluated statements and manually described statements. The higher the scores of the above indexes are, the better the model prediction effect is.

4. Results and Analysis

4.1. Model Training

We trained the proposed model using the dataset Rice2k, and the loss function curve of the training is shown in Figure 4. When the training reached about 20 rounds, the loss value of the used model stabilized at 0.6 and did not decrease anymore.

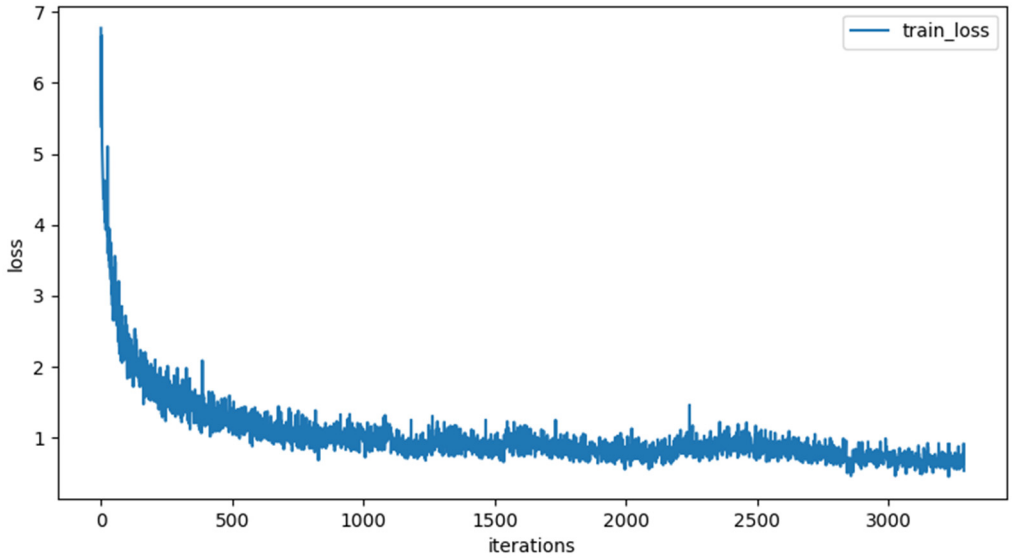

Figure 4. Training loss curve based on the DeiT feature encoder.

4.2. Ablation Experiments

The efficacy of the image description generation model was assessed using evaluation metrics such as BLEU, CIDEr, ROUGE-L, and METEOR. We conducted a comparative experiment using the proposed model and different CNN encoder models on Rice2k and obtained the results shown in Table 3. Analysis of Table 3 reveals that the overall performance of our model was superior on the Rice2k dataset, achieving the highest scores from BLEU-1 to BLEU-4. Notably, the performance of our model exhibited a significant enhancement of the CIDEr metric, attaining a score of 177.1, which is the highest score. This indicates its rich semantic representation capabilities. The METEOR value indicates a moderate level of performance, denoting a strong correspondence between the model-generated description and the reference description. Furthermore, the model proposed in this paper outperformed other models based on the ROUGE-L, reaching a score of 65.0, which suggests that it can autonomously generate more comprehensive descriptive statements, which is particularly well suited for automated image description generation tasks involving limited datasets.

Table 3. Evaluation results of the ablation experiments.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L	METEOR
AlexNet	72.1	59.1	50.9	44.8	162.8	63.8	37.3
VGG16	71.1	58.3	50.5	43.8	161.6	62.5	36.0
ResNet101	70.4	57.8	50.0	44.4	173.1	64.5	36.9
ResNet18	71.8	60.1	52.3	46.4	174.9	64.8	38.2
InceptionV3	70.9	58.7	50.7	45.0	172.4	64.8	38.2
MobileNetV2	71.7	59.5	51.5	45.3	174.1	64.3	37.5
DeiT	76.4	63.2	53.9	47.3	177.1	65.0	37.0

The encoders shown in the above table are pre-trained, and the decoders of the models with CNN encoders are all Attention + LSTM structures.

4.3. Image Description Results

The image caption generation model proposed in this paper demonstrated a strong performance, as evident from the selected prediction results shown in Table 4. Compared to the method proposed by Xie [29], our approach can capture more intricate details of plant diseases and pests. In image 1, our model not only accurately predicted the presence of bacterial stripe disease but also described that the entire rice leaf exhibits a withered yellow color, and the disease is serious. In image 2, our method precisely described the yellow stripes indicating the occurrence of sheath blight in the middle section of the leaf. Additionally, in image 3, our model accurately portrayed the grayish-white spots located at the center of the leaf, which are symptomatic of blast disease. Some of the predicted image descriptions align with the descriptions in the dataset, as depicted in image 4, where both exhibited the description "The body of the planthopper is semicircular".

To assess the generalization ability of the proposed model, we conducted simultaneous tests on a total of 342 authentic field images. The corresponding evaluation metrics and results are presented in Table 5. Inspection of the table reveals a slight decline in the performance of all models across the seven indicators measuring image description quality when applied to the new sample. Moreover, the values of the BLEU-1 to BLEU-4 indicators exhibited a gradual decrease. Notably, our model attained the highest scores on all the indicators. BLEU scores were still high. The results indicate a good grammatical and lexical match between the generated texts and the reference texts. Our model achieved 112.0 CIDEr and 54.1 ROUGE-L, which means that the key information between the generated texts of our model and the reference texts was better matched, and the generated descriptions were more diverse. In addition, the highest METEOR score indicates the overall high quality and fluency of the generated descriptions. The highest scores of these various indicators demonstrated superior performance when applied to the real-world image dataset captured in the field.

The prediction accuracy of 10 types of pests and diseases in this model is shown in Table 6, with an average accuracy of 87.67%. Notably, the accuracy for rice bakanae disease was 52.63%, while the remaining categories pertaining to rice pests and diseases exhibited remarkably high prediction accuracy. Specifically, five categories achieved a perfect accuracy of 100%, which demonstrates the model's proficient diagnostic capability in identifying rice pests and diseases. The test outcomes are communicated to users in a textual format and hold potential value as a point of reference for non-experts.

Table 4. Prediction results of the image description models based on the DeiT feature encoder.

Image	Predicted Results
 1. Rice bacterial streak	1. Rice is affected by bacterial streak disease. The leaves of rice show yellow stripes. 2. Rice is affected by bacterial streak disease, the disease is serious most of rice leaf surface whole leaf is withered and yellow. 3. Rice is affected by bacterial stripe disease.
 2. Rice bacterial blight	1. Rice is infected with bacterial blight and the leaves of rice show yellow pinstripes. 2. Rice is affected by bacterial blight, which has obvious yellow stripe in the middle of the leaves. 3. Rice is infected with bacterial blight and has yellow pinstripes on its leaves.
 3. Rice blast	1. Rice suffers from rice blast, irregular gray and white spot. 2. Rice suffers from rice blast, irregular gray and white spot. 3. Rice is afflicted with blast, the plant has a pale yellow gray-white spots.
 4. Planthopper	1. The body of the planthopper is semicircular. 2. The body of the planthopper is semicircular. 3. The planthopper has a pair of antennae.

1 is the annotated text of the dataset, 2 is the prediction result of the DeiT encoder model, and 3 is the prediction result of the ResNet18 encoder model.

Table 5. Evaluation results on the actual field images.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L	METEOR
AlexNet	63.6	47.0	37.4	30.8	103.9	51.2	29.6
VGG16	63.3	47.5	38.4	32.1	107.5	51.9	29.9
ResNet101	62.9	45.8	36.4	29.9	104.3	51.4	29.5
ResNet18	62.6	47.2	38.0	31.7	110.7	52.7	31.0
InceptionV3	63.8	47.1	37.8	31.5	109.2	52.0	30.2
MobileNetV2	65.0	47.9	38.0	31.0	107.9	52.2	30.1
DeiT	68.8	50.8	40.0	32.2	112.0	54.1	31.0

Table 6. Accuracy of category prediction on the actual field images.

Category	Accuracy %
Rice bacterial blight	79.49
Rice bacterial streak disease	82.22
Rice bakanae disease	52.63
Rice three chemical borers	89.47

Table 6. *Cont.*

Category	Accuracy %
Rice brown spot	72.92
Rice planthopper	100.00
Rice blast	100.00
Rice false smut	100.00
Rice sheath blight	100.00
Rice thrip	100.00
Average	87.67

4.4. Discussion

Insufficient data pose a challenge in the study of various rice pests and diseases, resulting in the impractical application of large-scale data learning methods. Utilizing a Convolutional Neural Network (CNN) as the feature encoder for image feature extraction presents limitations in terms of the perceptual field, primarily capturing localized information. Addressing this limitation requires multi-layer stacking to obtain global information. However, the adoption of deep networks may lead to the loss of smaller targets within disease sites, ultimately yielding inaccurate predictions. In comparison, the Transformer decoder exhibits enhanced semantic and long-distance feature extraction capabilities when compared to the traditional LSTM decoder, allowing for the generation of more accurate descriptions. Accordingly, we employed the convolution-free Transformer model architecture and leveraged the DeiT encoder to capture comprehensive information from the images. The proposed approach enables the generation of image descriptions for rice diseases and pests by transforming sequences of image patches into word sequences. This methodology facilitates the progression from categorizing rice diseases and pests to comprehending their intricate characteristics, which not only achieves the diagnosis of the category but also deepens people's understanding of the disease symptoms.

The image description dataset of rice pests and diseases established in this paper included three pests and seven diseases, fully considering various situations such as poor lighting, shooting angles, noise interference, etc. To enhance the dataset's fidelity, data augmentation techniques were employed to simulate complex field environments, ensuring a more realistic representation of the actual conditions in rice fields. Subsequently, the model was trained using this dataset, and comprehensive evaluation metrics for image description tasks were utilized to assess its performance in a comprehensive manner. Field samples were also employed to verify the method's feasibility. The experimental results demonstrate that the model exhibited a remarkable capability to capture crucial visual features, including shape, color, and texture, owing to the DeiT encoder's robust capacity to perceive and extract global information from images. Moreover, the model utilized the Transformer's decoder to effectively capture and learn syntactic features among words within a sentence, generating corresponding description texts by leveraging the interplay between image features and semantic attributes of words. With its powerful image perception, understanding ability, and syntactic feature capture and learning capabilities, this model outperformed the traditional CNN encoder model, achieving the highest scores across all evaluation metrics when tested on the actual scene dataset obtained from rice fields.

The research model presented in this paper holds considerable potential for diverse applications, particularly in the realm of rice production. It can effectively assist individuals with limited knowledge about diseases and pests of rice in accurately identifying the types of diseases and pests and further deepening their understanding of symptoms through descriptive text. Additionally, this model exhibits high training efficiency and demonstrates promising outcomes even when trained with a small-size dataset. Consequently, its application to other pests, diseases, or image description generation tasks offers the advantage of reduced data collection requirements and training costs.

However, it is important to acknowledge certain limitations within the agricultural image description experiments conducted in this paper. Specifically, the scope of the examined rice pest categories was relatively narrow, encompassing only 10 categories when compared to other studies. Consequently, the direct application of this model for the identification and description of other pests and diseases is not feasible. Moreover, existing public datasets contain a vast number of samples accompanied by detailed and diverse descriptions. Models trained on these extensive datasets generally exhibit strong generalization capabilities for describing objects and scenes beyond the dataset. In contrast, this model relies on a small dataset with a limited number of description texts, resulting in a relatively simple and less diverse representation of descriptions. Furthermore, despite the superior performance of the network model on the small-size dataset employed in this study, a risk of overfitting remains. The average prediction accuracy of the model on the actual field images was 87.67%, with substantial variations in accuracy observed across different categories. This disparity in accuracy could be attributed to potential shortcomings in the maturity of manually labeled description statements within the dataset. Additionally, it is worth noting that the image descriptions generated in this study are exclusively available in English and cannot be generated in other languages.

The subsequent research direction should be to consider expanding the dataset to include not only rice pest species but also other crop pest species to reduce the limitations of the model's prediction capabilities. Augmenting the image data volume is essential, along with expanding the descriptive dataset to enhance the diversity of sentence representations. Additionally, incorporating multiple languages can contribute to improved generalization and prediction accuracy, rendering the study more relevant and applicable in a broader context. Moreover, considering advancements to the basic structure of the Transformer model is recommended. Enhancements may involve refining the model's architecture to reduce computational complexity while maintaining its effectiveness. And the exploring reinforcement learning strategies can be beneficial in improving the description performance of the model, particularly in practical application scenarios.

5. Conclusions

In this paper, we proposed a DeiT feature encoder-based image description method for rice pests and diseases, constructed an image description dataset of 10 common rice pests and diseases, and conducted experiments on it. The results of the experiments show that the model has the ability to discriminate and describe the rice pests and diseases with the small training dataset. The generated text and image content were consistent. The description performance was stronger than that of the traditional convolutional method. This outcome highlights the efficacy of incorporating a distillation pre-trained DeiT encoder, which effectively resolves the challenge of the slow training of Transformer-structured models on extensive datasets. It enables the application of the model to the task of generating rice pest descriptions with limited data samples. Moving forward, future research endeavors aim to expand the agricultural pest and disease dataset, further improving the model's generalization ability. Additionally, it is necessary to extend the model's applicability to a broader range of agricultural pest and disease image descriptions, thereby facilitating the control of diverse pests and diseases and fostering agricultural development.

Author Contributions: Conceptualization, C.M. and Y.H.; methodology, C.M.; software, C.M.; validation, C.M., Y.Z. and H.L.; formal analysis, P.H.; investigation, D.D.; resources, D.D.; data curation, C.M.; writing—original draft preparation, C.M.; writing—review and editing, Y.H.; visualization, H.L.; supervision, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Humanity and Social Science Foundation of the Ministry of Education of China, grant numbers 18YJA630037 and 21YJA630054.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

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

References

1. Asibi, A.E.; Chai, Q.; Coulter, J.A. Rice Blast: A Disease with Implications for Global Food Security. *Agronomy* **2019**, *9*, 451. [CrossRef]
2. Huang, S.W.; Wang, L.; Liu, L.M.; Fu, Q.; Zhu, D.F. Nonchemical pest control in China rice: A review. *Agron. Sustain. Dev.* **2014**, *34*, 275–291. [CrossRef]
3. Singh, P.; Mazumdar, P.; Harikrishna, J.A.; Babu, S. Sheath blight of rice: A review and identification of priorities for future research. *Planta* **2019**, *250*, 1387–1407. [CrossRef] [PubMed]
4. Wang, P.; Liu, J.; Lyu, Y.; Huang, Z.; Zhang, X.; Sun, B.; Li, P.; Jing, X.; Li, H.; Zhang, C. A Review of Vector-Borne Rice Viruses. *Viruses* **2022**, *14*, 2258. [CrossRef] [PubMed]
5. Liu, Y.; Yu, J.; Han, Y. Understanding the effective receptive field in semantic image segmentation. *Multimed. Tools Appl.* **2018**, *77*, 22159–22171. [CrossRef]
6. Girish, K.; Visruth, P.; Vicente, O.; Sagnik, D.; Siming, L.; Yejin, C.; Berg, A.C.; Berg, T.L. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1601–1608.
7. Kuznetsova, P.; Ordonez, V.; Berg, T.L.; Choi, Y. TREETALK: Composition and Compression of Trees for Image Descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 351–362. [CrossRef]
8. Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Daumé, I. Midge: Generating Image Descriptions From Computer Vision Detections. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–27 April 2012.
9. Karpathy, A.; Joulin, A.; Li, F.F. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; Volume 3.
10. Kuznetsova, P.; Ordonez, V.; Berg, A.C.; Berg, T.L.; Choi, Y. Collective generation of natural image descriptions. In Proceedings of the Meeting of the Association for Computational Linguistics: Long Papers, Jeju Island, Republic of Korea, 8–14 July 2012.
11. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. *arXiv* **2014**, arXiv:1411.4555.
12. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2012; Volume 25.
13. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Comput. Sci.* **2014**. *peer reviewed*.
14. 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, Las Vegas, NV, USA, 27–30 June 2016.
15. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
16. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
17. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent Neural Network Regularization. *arXiv* **2014**, arXiv:1409.2329.
18. Cho, K.; Merrienboer, B.V.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
19. Mao, J.; Wei, X. Explain Images with Multimodal Recurrent Neural Networks. *arXiv* **2014**, arXiv:1410.1090.
20. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
21. Donahue, J.; Hendricks, L.A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. *Long-Term Recurrent Convolutional Networks for Visual Recognition and Description*; Elsevier: Amsterdam, The Netherlands, 2015.
22. Zhang, W.; He, X.Y.; Lu, W.Z. Exploring Discriminative Representations for Image Emotion Recognition With CNNs. *IEEE Trans. Multimed.* **2020**, *22*, 515–523. [CrossRef]
23. Huang, L.; Wang, W.; Xia, Y.; Chen, J. Adaptively Aligned Image Captioning via Adaptive Attention Time. *arXiv* **2019**, arXiv:1909.09060.
24. Ke, L.; Pei, W.; Li, R.; Shen, X.; Tai, Y.W. Reflective Decoding Network for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
25. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:1502.03044.
26. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

27. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *arXiv* **2017**, arXiv:1707.07998.
28. Guo, L.; Liu, J.; Tang, J.; Li, J.; Luo, W.; Lu, H. Aligning Linguistic Words and Visual Semantic Units for Image Captioning. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019.
29. Xie, Z.; Feng, Y.; Hu, Y.; Liu, H. Generating image description of rice pests and diseases using a ResNet18 feature encoder. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 197–206.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
31. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled Transformer for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
32. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on Attention for Image Captioning. *arXiv* **2019**, arXiv:1908.06954.
33. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
35. Liu, W.; Chen, S.; Guo, L.; Zhu, X.; Liu, J. CPTR: Full Transformer Network for Image Captioning. *arXiv* **2021**, arXiv:2101.10804.
36. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv* **2020**, arXiv:2012.12877.
37. Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. In Proceedings of the International Conference on Artificial Intelligence, Phuket, Thailand, 26–27 July 2015; pp. 853–899.
38. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
39. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Zitnick, C.L. *Microsoft COCO: Common Objects in Context*; Springer International Publishing: Berlin/Heidelberg, Germany, 2014.
40. Lu, X.Y.; Yang, R.; Zhou, J.; Jiao, J.; Liu, F.; Liu, Y.F.; Su, B.F.; Gu, P.W. A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1755–1767. [CrossRef]
41. Nazari, K.; Ebadi, M.J.; Berahmand, K. Diagnosis of Alternaria disease and leafminer pest on tomato leaves using image processing techniques. *J. Sci. Food Agric.* **2022**, *102*, 6907–6920. [CrossRef]
42. Chao, D.W.; Jun, S.S.; Bin, S.W. An algorithm of image hashing based on image dictionary of CBIR. *Microcomput. Its Appl.* **2010**. [CrossRef]
43. Yumei, Y.; Yi, P.; Junhui, Q. Research on the Image Similarity Retrieval Algorithm Based on Double Hash. *Inf. Commun. Technol.* **2019**.
44. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002.
45. Satanjeev, B. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the Second Workshop on Statistical Machine Translation, Ann Arbor, MI, USA, 25–30 June 2005; pp. 228–231.
46. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 25–26 July 2004.
47. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.