symmetry

MDPI

Article

Framework of Specific Description Generation for Aluminum Alloy Metallographic Image Based on Visual and Language Information Fusion

Dali Chen [1,*], Yang Liu [1], Shixin Liu [1], Fang Liu [2] and Yangquan Chen [3]

[1] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; ly1562945324@163.com (Y.L.); liushixin@ise.neu.edu.cn (S.L.)
[2] College of Materials Science and Engineering, Northeastern University, Shenyang 110819, China; liufang@smm.neu.edu.cn
[3] School of Engineering, University of California, Merced, CA 95343, USA; ychen53@ucmerced.edu
* Correspondence: chendali@ise.neu.edu.cn

Received: 2 April 2020; Accepted: 1 May 2020; Published: 6 May 2020

Abstract: The automatic generation of language description is an important task in the intelligent analysis of aluminum alloy metallographic images, and is crucial for the high-quality development of the non-ferrous metals manufacturing industry. In this paper, we propose a methodological framework to generate the language description for aluminum alloy metallographic images. The framework consists of two parts: feature extraction and classification. In the process of feature extraction, we used ResNet (residual network) and CNN (convolutional neural network) to extract visual features from metallographic images. Meanwhile, we used LSTM (long short term memory), FastText, and TextCNN to extract language text features from questions. Then, we implemented a fusion strategy to integrate these two features. Finally, we used the fused features as the input of the classification network. This framework turns the description generation problem into a classification task, which greatly simplifies the generation process of language description and provides a new idea for the description of metallographic images. Based on this basic framework, we implemented seven different methods to generate the language description of aluminum alloy metallographic images, and their performance comparisons are given. To verify the effectiveness of this framework, we built the aluminum alloy metallographic image dataset. A large number of experimental results show that this framework can effectively accomplish the given tasks.

Keywords: aluminum alloy; feature fusion; image description generation; metallographic image

1. Introduction

Aluminum alloy is one of the most widely-used non-ferrous metal materials in industry. Because of its good performance, it is widely used in aviation, aerospace, navigation, railways, highways, and other fields [1–6]. The properties of aluminum alloy mainly depend on its microstructures, and metallographic analysis is the main method to evaluate its microstructures [7–10]. In practice, material science experts evaluate the properties of aluminum alloys by observing and analyzing the given metallographic images. The analysis of complex metallographic images generally requires a lot of time and energy from experts, and suffers from poor repeatability due to the different experience of participants [11].

In order to solve these problems, more and more scholars have begun to pay attention to research into intelligent metallographic image processing and analysis methods [12–15]. In recent years, many automatic metallographic image-processing and analysis methods have been proposed, which can greatly improve the efficiency of metallographic analysis tasks. According to different functions, these

methods can be divided into four categories: microstructural classification, segmentation, quantitative calculation, and grain boundary extraction.

The aim of the microstructural classification method is to classify different microstructures in a given metallographic image. For example, Decost and Holm proposed a computer vision approach for automatic analysis and classification of microstructural image data. This approach was able to classify microstructures into one of seven groups with greater than 80% accuracy [16]. In Gola et al.'s paper [17], a data-mining process is presented based on a support vector machine (SVM), which was able to distinguish between different microstructures of the two-phase steels.

Microstructural segmentation methods aim to segment the different microstructures in a given metallographic image. For example, Jiang et al. applied an improved SLIC (simple linear iterative clustering) algorithm and region-merging technique to automatically segment grain regions [18]. Albuquerque et al. applied multilayer perceptron and self-organizing map neural network topologies to segment microstructures from metallographic images [19]. In Bulgarevich et al.'s paper [20], a fast random forest-based method is proposed for reliable and automatic segmentation of typical steel microstructures. In Albuquerque et al.'s paper [21], a neuronal network-based method is proposed for automatic segmentation of nickel alloy secondary phases from SEM (scanning electron microscope) images. In Papa et al.'s paper [22], the automatic segmentation of graphite particles in metallographic images is achieved by using Otsu, SVM, Bayesian, and optimum-path forest methods. Deep learning methods have dramatically improved conventional machine learning techniques due to their strong ability to learn the hierarchical latent features of high-dimensional data [23,24]. These methods have been successfully applied in metallographic image segmentation. Azimi et al. [25], proposed a fully-convolutional neural network (FCNN) accompanied by max-voting scheme to segment some given microstructures of low carbon steel. In Ma et al.'s paper [26], the DeepLab network was used for Al–La alloy metallographic images segmentation. These deep-learning-based methods achieved satisfactory results. However, they always needed a large number of hand-labeled data to achieve accurate microstructural segmentation. In order to solve this problem, a fast automatic labeling method is proposed to label metallographic images quickly [27].

Microstructural quantitative calculation aims to obtain the quantitative information from the given metallographic image, such as the size, shape, and distribution of the different microstructures. For example, in references [28] and [29], conventional digital image processing methods are used for automatic quantification of microstructural features.

Grain boundary extraction aims to extract the grain boundaries. For example, in Xu et al.'s paper [30], an improved mean shift method is presented for automatically extracting grain boundaries, solving the problem of grain boundary blurring or disconnection. In Journaux et al.'s paper [31], the directional wavelets and mathematical morphology are used for grain boundary extraction.

Differently from the existing metallographic image processing methods, in this paper, we focus our attention on the automatic generation of language description from metallographic images. The purpose is to automatically generate a description of the content of a given metallographic image similar to one obtained from material science experts. It is an important part of the intelligent metallographic image analysis system. The automatic generation of a language description of metallographic images is a very challenging task, because it requires the combination of image processing and natural language processing. Recently, many methods have been proposed for natural scene images [32–34]. These methods obtain multiple objects and their spatial relationships, and then generate the language description to fit these constituent parts. These language descriptions are often general.

In contrast to natural scene images, aluminum alloy metallographic images need more specific language description, which is useful for subsequent analysis. To address this requirement, we propose a novel method to automatically generate specific language descriptions from aluminum alloy metallographic images. Inspired by Antol, Wu and Kazemi et al.'s papers [35–37], we considered the aluminum alloy metallographic image description task as a classification problem. This method consists of two parts. The first part, feature extraction, extracts and fuses the best visual and language

features for use in the generation of the language description of aluminum alloy metallographic images. The second part, classification, predicts classification to generate a natural language description based on the extracted features.

We summarize the contributions of this paper as follows.

(1) We achieved automatic generation of the language description for given aluminum alloy metallographic images. In this framework, the aluminum alloy metallographic image description task can be considered as a classification problem, which greatly simplifies the generation process of language description and provides a new idea for the description of metallographic images.

(2) We used ResNet [38–40] and convolutional neural network (CNN) to extract visual features from metallographic images. Meanwhile, we used LSTM [41], FastText [42], and TextCNN [43] to extract language text features from given questions. Moreover, we present the comparative analysis among these seven combination strategies applied to generate natural language description of aluminum alloy metallographic images.

(3) The proposed method can not only obtain the language description, but also obtain the attention map. This attention map can correctly reflect the high attention area of the given aluminum alloy metallographic image. This is helpful for the professionals to analyze the aluminum alloy metallographic images.

This paper is organized as follows: Section 1 introduces prior work and our contributions. In Section 2, we introduce the proposed method, including the feature extraction scheme and classification method. Section 3 presents the performance comparisons, attention map analysis, and convergence analysis. The paper is concluded in Section 4.

2. The Proposed Methods

The automatic generation of the language description is an important part of the automated analysis system of aluminum alloy metallographic images. The basic framework of our proposed method is shown in Figure 1. It consists of two parts: feature extraction and classification. Differently from the typical automatic generation method of language description, the input of our method includes not only image, but also one text question associated with this image. This question is often the most important issue, as it must generate a specific language description and be helpful for analysis of the metallographic image. In the feature extraction scheme, we extract a metallographic image feature and a text question feature at the same time, and then merge them into one latent feature. The classification method is used to classify the obtained latent feature and get the specific language description of the given aluminum alloy metallographic image. The output includes not only the language description, but also an attention map of the given metal micrograph. This attention map is learned automatically by the proposed deep neural network. From the attention map, we can find the key visual features that affect the generation of language description. This will provide more valuable information for us to analyze the aluminum alloy metallographic images. In the next section, we will introduce the proposed method in detail.

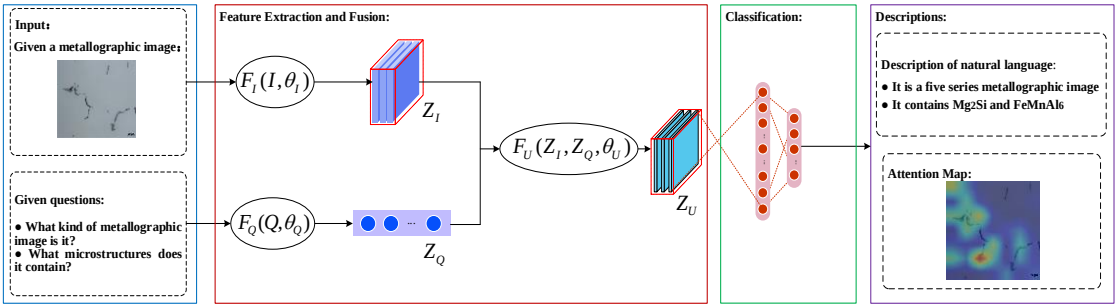


Figure 1. The flowchart of the proposed method.

2.1. Feature Extraction and Fusion Scheme

The aim of the feature extraction scheme is to transform the given aluminum alloy metallographic images and corresponding questions from image and text data space to latent feature space. For this purpose, the deep neural network is used due to its strong ability to learn the hierarchical latent features of given metallographic images and questions. This scheme consists of three parts: metallographic image feature extraction, question text feature extraction, and features fusion.

For the metallographic image feature extraction, the latent visual feature z_I can be computed by

$$z_I = F_I(I, \theta_I), \tag{1}$$

where I is the given metallographic image and F_I represents a certain convolutional neural network (CNN) method. θ_I represents the parameter in the given convolutional neural network F_I. In this paper, we used CNN and improved ResNet. The size of input image was 224×224 pixels and the size of image features extracted by ResNet was 14×14. In order to keep the 14×14 size of the feature map, the convolution neural network including four convolutional layers and four pooling layers was used to extract image features. In the training process, the parameter θ_I is adjusted to fit the given metallographic image training dataset.

Similarly, in the process of question feature extraction, the latent question text feature z_Q can be computed by

$$z_Q = F_Q(Q, \theta_Q), \tag{2}$$

where Q is the given question, F_Q represents a certain deep neural network method, and θ_Q represents the parameter in the given deep neural network F_Q. In this paper, we use improved LSTM, FastText and TextCNN.

The latent visual feature z_I and text feature z_Q have different dimensions. Therefore, we needed to design a fusion method to integrate the two features. Let θ_1 be the 1×1 dimensional convolution layer of depth 512, we have

$$z_I^* = CONV(z_I, \theta_1), \tag{3}$$

and

$$z_Q^* = CONV(z_Q, \theta_1), \tag{4}$$

where $CONV$ is the convolution operator, and z_I^* and z_Q^* have the same dimension. Therefore, we can compute

$$z^* = z_I^* + z_Q^*, \tag{5}$$

Let θ_2 be the 1×1 dimensional convolution layer of depth 2, we have

$$w = softmax(CONV(ReLU(CONV(z^*, \theta_1)), \theta_2)), \tag{6}$$

where $softmax$ is the $softmax$ function and $ReLU$ is the rectified linear unit activation function. The $softmax$ classifier is the most popular classifier and many experiments have shown that the $softmax$ classifier can get satisfactory results. Therefore, we used the $softmax$ function as the classifier in our framework. Then we could compute

$$z_A = F_w(w, z_I), \tag{7}$$

where F_w is the weighted average operator. The fused feature z_U can be obtained by

$$z_U = F_C(z_A, z_Q), \tag{8}$$

where F_C is the concatenate operator. For easy description, we define

$$z_U = F_U(z_I, z_Q, \theta_U), \tag{9}$$

where F_U represents the fusion operator and θ_U is the parameter of fusion network.

In the process of feature fusion, our purpose was to fuse the visual feature with the text feature. However, the image feature dimension was different from the text feature dimension. The dimension of the image feature was $14 \times 14 \times 2048$, the dimension of the question feature was 1×1024. We expanded the dimension of text so that we could make the image feature and question feature have the same dimension. Finally, we added them up to get the fusion feature.

To summarize, the fusion strategy in a form of a pseudo-code is shown in Algorithm 1, as follows:

Algorithm 1: The fusion strategy of latent visual and text features

Inputs: Latent visual feature Z_I, text feature Z_Q and network parameter θ_U.
Output: Fused feature z_U.
Step 1: Compute z^* by Equations (3)–(5).
Step 2: Compute w by Equation (6).
Step 3: Compute z_A by Equation (7).
Step 4: Compute z_U by Equation (8).

In the training step, the parameters θ_I, θ_Q, and θ_U will be adjusted to fit the given metallographic image training dataset.

2.2. Classification Method

The aluminum alloy metallographic image description task aims to generate the specific and accurate language description of aluminum alloy metallographic images. In fact, in the metallographic analysis of aluminum alloys, people often want a limited number of questions. We can list these questions and give the corresponding answers. Therefore, we can consider the aluminum alloy metallographic image description task as a classification problem, which can reasonably simplify the generation process of language description. The input of classifier is the fusion feature, and the output is the language description, as shown in Figure 1.

Let c_k represent the k-th description and $p(c_k|z)$ be the probability that the feature z generates the k-th description, which is given by the $softmax$ transformation of linear functions of the feature variable,

$$p(c_k|z;\theta_A) = \frac{\exp(z^T\theta_A^{(k)})}{\sum_{i=1:K}\exp(z^T\theta_A^{(i)})}, \tag{10}$$

and θ_A is the parameters of classifier. The best description can be obtained by

$$y = \underset{k=1:K}{\operatorname{argmax}} p(c_k|z;\theta_A) = \underset{k=1:K}{\operatorname{argmax}} p(c_k|I,Q;\theta), \tag{11}$$

where K is the number of descriptions and the description set $c = \{c_1, c_2, \ldots c_K\}$. The network parameter θ is defined by $\theta = \left\{\theta_I, \theta_Q, \theta_U, \theta_A\right\}$, which is very important for generating the accurate language description.

In order to deal with the problem of parameter estimation, we use the maximum likelihood estimation (MLE) to compute the network model parameter θ. Suppose that we are given a training dataset $D = \{I_n, Q_n, t_n; n = 1:N\}$. Here, I_n is the n-th metallographic image, Q_n is the n-th question, and t_n is a binary class label vector $t_n \in \{0,1\}^K$, where $\sum_{k=1:K} t_n^{(k)} = 1$,$k = 1, \ldots, K$. The 1 of K coding scheme is used in label vector t_n if the input $\{I_n, Q_n\}$ belongs to class c_k,$t_n^{(k)} = 1$. For easy description, we set $I = \{I_1, I_2, \ldots, I_N\}$, $Q = \{Q_1, Q_2, \ldots, Q_N\}$ and $T = \{t_1, t_2, \ldots, t_N\}$. Assume that the class labels are independent, then the likelihood function is given by

$$p(I,Q,T|\theta) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(c_k|I_n,Q_n;\theta)^{t_n^{(k)}}, \tag{12}$$

Using the maximum likelihood estimation, we can compute the network model parameters θ by solving the following optimization problem

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\{-\ln(p(I,Q,T|\theta))\} = \underset{\theta}{\operatorname{argmin}}\sum_{n=1}^{N}\sum_{k=1}^{K} t_n^{(k)} \ln p(c_k|I_n,Q_n;\theta), \quad (13)$$

which is known as the cross-entropy loss function for the multiclass classification problem. The stochastic gradient descent (SGD) algorithm is used to solve this optimization problem.

To summarize, our metallographic image description method in a form of a pseudo-code is as shown in Algorithm 2, as follows:

Algorithm 2: Generation method of the language description for given aluminum alloy metallographic image

Inputs: Training dataset $D = \{I_n, Q_n, t_n\}$, new aluminum alloy metallographic image and question $\{I', Q'\}$.
Output: The description c^*.
Step 1: Initialization:

- Learning method: epochs, batch size, initial learning rate, weight decay.
- Network parameter θ

Step 2: Optimize θ by using D:

- **while** not converge **do**
- Compute network parameter θ^* by solving Equation (13) using SGD algorithm.
- Update $\theta^* \rightarrow \theta$.
- **end while**

Step 3: Generate description by using $\{I', Q'\}$ and θ:

- Compute the latent feature z_U by using Algorithm 1.
- Compute $\boldsymbol{y}$ by solving Equation (11).
- Generate the language description $c^* = c_y$

3. Experimental Results

3.1. Experimental Dataset

In order to verify the proposed method, we built the experimental dataset, which contained 180 aluminum alloy metallographic images and 180 natural scene images. The natural scene images were obtained by randomly sampling from the COCO (common objects in context) public dataset. The aluminum alloy metallographic images were taken by metallographic microscope, and consisted of 100 5-series metallographic images and 80 6-series metallographic images. These metallographic images included six different types of phases, such as Mg_2Si, $FeMnAl_6$, $FeAl_3$, $MnAl_6$, Si, and $FeMnSiAl_6$. Two typical aluminum alloy metallographic images are shown in Figure 2.

(**a**) (**b**)

Figure 2. Aluminum alloy metallographic images. (**a**) 5-series and (**b**) 6-series.

In addition, in this dataset, we designed four questions and eleven language descriptions for each image according to the practical requirements of aluminum alloy metallographic image analysis, as shown in Table 1. In our experiment, we have eleven classes or descriptions in total, as shown in the second row, and each class corresponded to one combination of given metallographic image and question. In order to clarify the description of the relationship, we marked the questions and descriptions in Table 1. For example, the first question is labeled 1 and the corresponding descriptions is also labeled 1.

Table 1. Four questions and eleven language descriptions.

Questions (Q)	• Is it a metallographic image? (1) • What type of metallographic image is this? (2) • How many types of microstructures in this image? (3) • What type of microstructure does this image contain? (4)
Descriptions (C)	• It is a natural scene image and does not contain any microstructure. (1) • It is a metallographic image. (1) • It is a 5-series metallographic image. (2) • It is a 6-series metallographic image. (2) • There are two types of microstructures in this metallographic image. (3) • There are three types of microstructures in this metallographic image. (3) • There are four types of microstructures in this metallographic image. (3) • This metallographic image contains Mg2Si and FeMnAl6. (4) • This metallographic image contains Mg2Si, FeMnAl6, and FeAl3. (4) • This metallographic image contains Mg2Si, MnAl6, and Si. (4) • This metallographic image contains Mg2Si, MnAl6, Si, and FeMnSiAl6. (4)

3.2. Performance Comparison

The aim of this section is to analyze the performance of the proposed description method. Let $D_v = \{I_n, Q_n, C_n\}$ represent the test dataset and f_θ represent our proposed method. We used the accuracy (ACC) to evaluate our proposed method f_θ, which has been widely used in many literatures. It can be calculated by the following formula

$$ACC(f_\theta; D_v) = \frac{1}{N}\sum_{n=1}^{N}\prod(c_n^* = C_n), \tag{14}$$

where $\prod(c_n^* = C_n)$ is the indicator function defined as

$$\prod(c_n^* = C_n) = \begin{cases} 1, if c_n^* = C_n \\ 0, else \end{cases}, \tag{15}$$

where c_n^* is the estimated result and C_n is the ground truth.

In addition, we have implemented seven different methods on the basis of the basic proposed framework. For easy description, we set f_i, $i = 1:7$, denotes the i-th method. The networks used in these seven methods is shown in Table 2. We set F_I denotes the visual feature extraction network and F_Q denotes the text feature extraction network. In our framework, these two networks are critical. As shown in Table 2, the first method, f_1 consists of two networks, ResNet152 and LSTM1024. Similarly, f_2 consists of ResNet152 and LSTM256, f_3 consists of ResNet34 and LSTM1024, f_4 consists of ResNet34 and LSTM256, f_5 consists of CNN and LSTM256, f_6 consists of CNN and TextCNN, and f_7 consists of CNN and FastText.

Table 2. The networks used in the seven methods.

Methods		f_1	f_2	f_3	f_4	f_5	f_6	f_7
F_I	ResNet152	✓	✓					
	ResNet34			✓	✓			
	CNN					✓	✓	✓
F_Q	LSTM1024	✓		✓				
	LSTM256		✓		✓	✓		
	TextCNN						✓	
	FastText							✓

The detailed network structure and parameter settings are shown below:

- ResNet152: 50 residual blocks (each residual block consists of three convolutional layers), two convolutional layers, and five pooling layers.
- ResNet34: 16 residual blocks (each residual block consists of two convolutional layers), two convolutional layers, and five pooling layers.
- CNN: four convolutional layers and four pooling layers.
- LSTM1024: Each LSTM unit consists of three gate control systems and one cell, output dimension is 1024.
- LSTM256: Similar to LSTM1024, output dimension is 256.
- TextCNN: 100-dimensional word embedding, three convolutional layers and three pooling layers; they are individuals.
- FastText: 100-dimensional word embedding and two linear layers.

In the process of training, we use a fixed epochs of 500 and initial learning rate of 0.001. In ResNet, the number of residual blocks was 50 or 16. In CNN, the number of convolution layers was four and the number of pooling layers was four. In LSTM (long short term memory), the output dimension was 1024 or 256. In TextCNN, the word embedding was 100 dimensions, there were three convolutional layers and the three pooling layers. In FastText, the word embedding was 100 and the linear layers was two. Moreover, we used the SGD optimizer with L2 regularization. The weight decay was 0.02. It could accelerate the training process of the model.

In experiments, we used the cross-validation method to ensure the accuracy of the evaluation results. We divide dataset D into six mutually exclusive subsets with the same size, $D = D_1 \cup D_2 \cup D_3 \cup D_4 \cup D_5 \cup D_6$ and $D_i \cap D_j = \phi(i \neq j)$. We used five subsets as the training set and the remaining subset as the test set, and then got six experimental results, as shown in the second to seventh columns in Table 3.

Table 3. The ACC comparison among the seven different methods.

ACC	1	2	3	4	5	6	Average
f_1	0.87461	0.99168	0.99179	0.99140	0.95819	0.96650	0.96236
f_2	0.87538	0.99167	0.99179	0.99173	0.95808	0.96647	0.96252
f_3	0.87216	0.99123	0.99152	0.99917	0.95824	0.96680	0.96319
f_4	0.87527	0.99146	0.99140	0.99157	0.95841	0.96630	0.96241
f_5	0.86434	0.99191	0.97439	0.98305	0.95832	0.95028	0.95371
f_6	0.87059	0.99180	0.99030	0.98471	0.95825	0.94962	0.95754
f_7	0.86858	0.99168	0.98622	0.98766	0.95796	0.95076	0.95714

The last column in the Table 3 shows the average of six experiments. The first row denotes the experiment number. From these experimental results, we can see that all the seven methods had more than 90% accuracy, and the third method f_3 (ResNet34 and LSTM1024) had the best average accuracy. Therefore, we can conclude that the proposed method can accurately generate the language description of the aluminum alloy metallographic image. In addition, the box plots of experiment results obtained by the seven methods are shown in Figure 3.

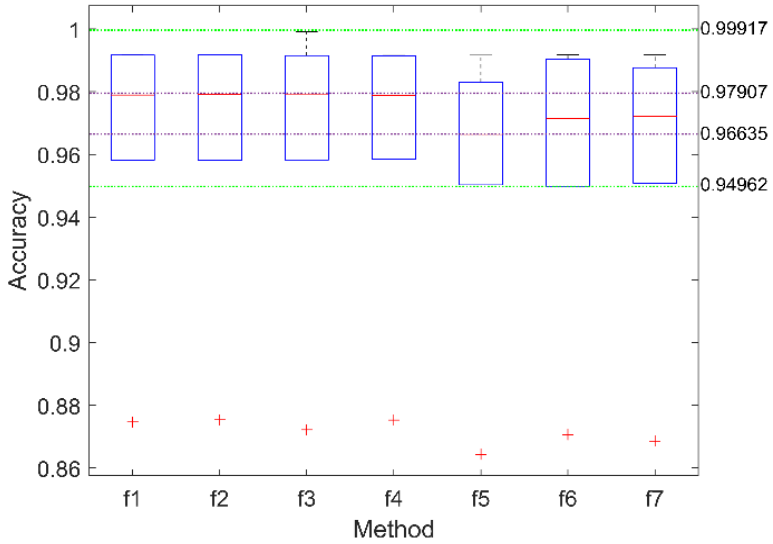


Figure 3. Box plot of experimental results obtained by the seven methods.

In Figure 3, the red plus sign denotes outlier and red line denotes median. From Figure 3, we can observe: (1) the outliers are caused by the first experiment, so we compute the median without using outliers, (2) all median values are concentrated between 0.96635 and 0.97907, and the length of the interval is less than 0.013, and (3) the experimental results are mainly distributed between 0.94962 and 0.99917, and the length of the interval is less than 0.050. Therefore, we can conclude that the proposed method has good robustness.

In addition, we randomly divided the dataset into six mutually-exclusive subsets with the same size for the experiment, this is 6-fold cross validation. In this way, we could improve the stability of the model. However, the dataset was randomly divided, which may have led to incomplete classes of some datasets. Therefore, this cross validation method leads to some outliers. For example, from Table 3 and Figure 3, we can see that dataset 1 does not contain all classes.

The results of training time of the seven different methods are shown in Table 4. We can see that the training time for the third and fourth methods were both less than 20 minutes.

Table 4. Training time comparison among seven different methods.

Training Time	1	2	3	4	5	6	Average
f_1	27 min 25.9 s	35 min 0.6 s	34 min 4.4 s	34 min 5.6 s	35 min 17.3 s	34 min 35.7 s	33 min 4.8 s
f_2	24 min 12.4 s	33 min 51.1 s	26 min 28.1 s	32 min 3.6 s	32 min 46.8 s	33 min 2.2 s	30 min 24.0 s
f_3	16 min 12.4 s	17 min 14.6 s	17 min 49.1 s	16 min 47.8 s	16 min 52.6 s	16 min 12.5 s	16 min 51.5 s
f_4	14 min 32.5 s	14 min 26.8 s	15 min 58.6 s	14 min 40.2 s	14 min 14.6 s	14 min 30.6 s	14 min 43.9 s
f_5	48 min 56.6 s	49 min 4.2 s	49 min 17.5 s	44 min 23.3 s	48 min 4.1 s	47 min 40.5 s	47 min 54.6 s
f_6	35 min 17.4 s	38 min 44.2 s	40 min 2.1 s	50 min 12.1 s	45 min 37.7 s	45 min 4.3 s	40 min 57.1 s
f_7	46 min 9.6 s	40 min 31.5 s	41 min 56.0 s	41 min 56.9 s	41 min 16.4 s	38 min 50.9 s	41 min 46.9 s

3.3. Attention Map Analysis

Our method generates not only the language description, but also an attention map of a given metal micrograph. In this section, we will analyze the importance of attention images. Figure 4 shows an example of an attention map experimental result.

The attention map is learnt automatically by the proposed deep neural network and can correctly reflect the high attention area of the given aluminum alloy metallographic image. It is a probability map which is extracted from a convolutional neural network. In our network, we sent the fusion feature to a convolutional neural network and we could get the attention map. The process was as follows: (1) we used a two-layer convolution network to process the fusion feature and get two initial attention maps, (2) we averaged these two attention maps in pixels to get the final attention map and transform the pixel value into probability value, and (3) the attention map was processed by pseudo-color.

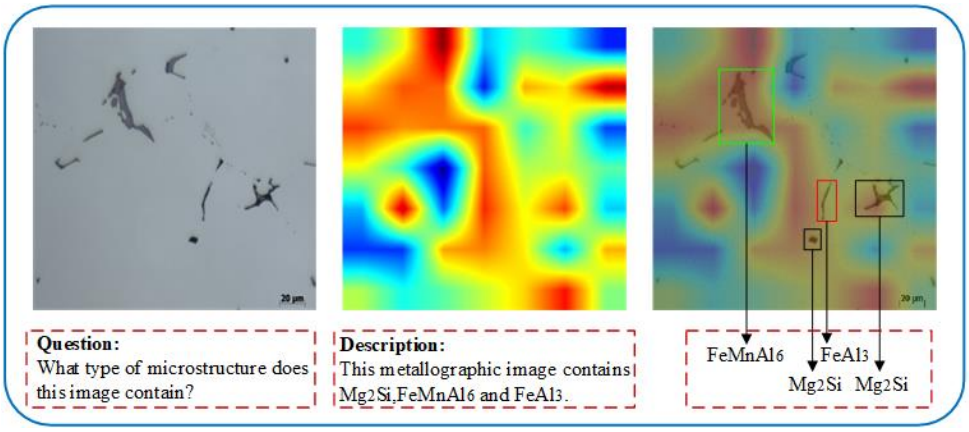
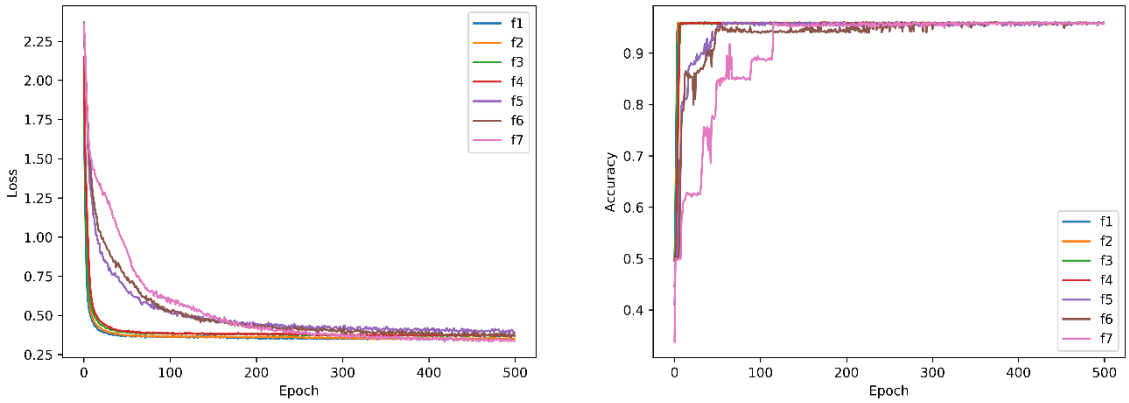


Figure 4. An example of an attention map experimental result.

The left figure shows the system input, which included the given aluminum alloy metallographic image and question of interest. The output includes the attention map and language description, as shown in the middle figure. In the attention map, the red color denotes the regions with high attention. For convenient analysis, we overlie the attention map on the original image, as shown in the right one. From Figure 4, we can observe that the main microstructures are distributed in the regions with high attention. This verifies the effectiveness of attention maps. Therefore, we can conclude that attention maps are helpful for the analysis of aluminum alloy metallographic images.

3.4. Convergence Analysis

The aim of this experiment is to analyze the convergence of the proposed framework. The loss and accuracy curves with times for seven different methods are shown in Figure 5. From Figure 5, we can observe that these seven methods can converge, and that methods f_1,f_2,f_3, and f_4 have better convergence speed than the other three methods. This verifies the convergence of the proposed methods.



(**a**) (**b**)

Figure 5. (**a**) The loss curves and (**b**) The accuracy curves.

4. Conclusions

In this paper, we propose a basic framework to generate the language description for aluminum alloy metallographic image. This framework is considered as a classification problem and includes feature extraction and classification. Using this basic framework, we implemented seven different methods to generate the language description of aluminum alloy metallographic images. A large number of experimental results show that this framework can effectively accomplish the given tasks and has good convergence. In the future, we plan to investigate the use of semantic segmentation of metallographic image for further improvement. In addition, we are also interested in the use of high level semantic priors of microstructures.

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

Funding: This research was funded by the National Key R&D Program of China under Grant 2017YFB0306400 and the National Natural Science Foundation of China under Grant 61773104.

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

References

1. Hirsch, J.; Al-Samman, T. Superior light metals by texture engineering: Optimized aluminum and magnesium alloys for automotive applications. *Acta Mater.* **2013**, *61*, 818–843. [CrossRef]
2. Zhang, J.L.; Song, B.; Wei, Q.S.; Bourell, D.; Shi, Y.S. A review of selective laser melting of aluminum alloys: Processing, microstructure, property and developing trends. *J. Mater. Sci. Technol.* **2019**, *35*, 270–284. [CrossRef]
3. Kadleckova, M.; Minarik, A.; Smolka, P.; Mracek, A.; Wrzecionko, E.; Novak, L.; Musilova, L.; Gajdosik, R. Preparation of Textured Surfaces on Aluminum-Alloy Substrates. *Materials* **2019**, *12*, 109. [CrossRef]
4. Heinz, A.; Haszler, A.; Keidel, C. Recent development in aluminium alloys for aerospace applications. *Mater. Sci. Eng. A* **2000**, *280*, 102–107. [CrossRef]
5. Du, Y.J.; Damron, M.; Tang, G. Inorganic/organic hybrid coatings for aircraft aluminum alloy substrates. *Progress Org. Coat.* **2001**, *41*, 226–232.
6. Martin, J.H.; Yahata, B.D.; Hundley, J.M.; Mayer, J.A.; Schaedler, T.A.; Pollock, T.M. 3D printing of high-strength aluminium alloys. *Nature* **2017**, *549*, 365–369. [CrossRef]
7. Girault, E.; Jacques, P.; Harlet, P. Metallographic Methods for Revealing the Multiphase Microstructure of TRIP-Assisted Steels. *Mater. Charact.* **1998**, *40*, 111–118. [CrossRef]
8. Roy, N.; Samuel, A.M.; Samuel, F.H. Porosity formation in AI9 Wt Pct Si3 Wt Pct Cu alloy systems: Metallographic observations. *Metall. Mater. Trans. A* **1996**, *27*, 415–429. [CrossRef]

9. Rohatgi, A.; Vecchio, K.S.; Gray, G.T. A metallographic and quantitative analysis of the influence of stacking fault energy on shock-hardening in Cu and Cu–Al alloys. *Acta Mater.* **2001**, *49*, 427–438. [CrossRef]
10. Rajasekhar, K.; Harendranath, C.S.; Raman, R. Microstructural evolution during solidification of austenitic stainless steel weld metals: A color metallographic and electron microprobe analysis study. *Mater. Charact.* **1997**, *38*, 53–65. [CrossRef]
11. Tamadon, A.; Pons, D.J.; Sued, K.; Clucas, D. Development of Metallographic Etchants for the Microstructure Evolution of A6082-T6 BFSW Welds. *Metals* **2017**, *7*, 423. [CrossRef]
12. Moreira, F.D.; Xavier, F.G.; Gomes, S.L.; Santos, J.C.; Freitas, F.N.; Freitas, R.G. New Analysis Method Application in Metallographic Images through the Construction of Mosaics via Speeded up Robust Features and Scale Invariant Feature Transform. *Materials* **2015**, *8*, 3864–3882.
13. Paulic, M.; Mocnik, D.; Ficko, M. Intelligent system for prediction of mechanical properties of material based on metallographic images. *Teh. Vjesn.—Tech. Gaz.* **2015**, *22*, 1419–1424.
14. Povstyanoi, O.Y.; Sychuk, V.A.; Mcmillan, A. Metallographic Analysis and Microstructural Image Processing of Sandblasting Nozzles Produced by Powder Metallurgy Methods. *Powder Metall. Metal Ceram.* **2015**, *54*, 234–240. [CrossRef]
15. Chowdhury, A.; Kautz, E.; Yener, B.; Lewis, D. Image driven machine learning methods for microstructure recognition. *Comput. Mater. Sci.* **2016**, *123*, 176–187. [CrossRef]
16. DeCost, B.L.; Holm, E.A. A computer vision approach for automated analysis and classification of microstructural image data. *Comput. Mater. Sci.* **2015**, *110*, 126–133. [CrossRef]
17. Gola, J. Advanced microstructure classification by data mining methods. *Comput. Mater. Sci.* **2018**, *148*, 324–335. [CrossRef]
18. Jiang, F.; Gu, Q.; Hao, H. A method for automatic grain segmentation of multi-angle cross-polarized microscopic images of sandstone. *Comput. Geosci.* **2018**, *115*, 143–153. [CrossRef]
19. De Albuquerque, V.H.; de Alexandria, A.R.; Cortez, P.C.; Tavares, J.M. Evaluation of multilayer perceptron and self-organizing map neural network topologies applied on microstructure segmentation from metallographic images. *NDT E Int.* **2009**, *42*, 644–651. [CrossRef]
20. Bulgarevich, S. Pattern recognition with machine learning on optical microscopy images of typical metallurgical microstructures. *Sci. Rep.* **2018**, *8*, 2078. [CrossRef]
21. De Albuquerque, V.H.; Silva, C.C.; Menezes, T.I.; Farias, J.P.; Tavares, J.M. Automatic evaluation of nickel alloy secondary phases from sem images. *Microsc. Res. Tech.* **2011**, *74*, 36–46. [CrossRef] [PubMed]
22. Papa, J.P.; Nakamura, R.Y.; De Albuquerque, V.H.; Falcão, A.X.; Tavares, J.M. Computer techniques towards the automatic characterization of graphite particles in metallographic images of industrial materials. *Expert Syst. Appl.* **2013**, *40*, 590–597. [CrossRef]
23. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
24. Bengio, Y.; Goodfellow, J.; Courville, A. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2016.
25. Azimi, S.M.; Britz, D.; Engstler, M.; Fritz, M.; Mücklich, F. Advanced steel microstructure classification by deep learning methods. *Sci. Rep.* **2018**, *8*, 2128. [CrossRef] [PubMed]
26. Ma, B.; Ban, X.; Huang, H.; Chen, Y.; Liu, W.; Zhi, Y. Deep learning-based image segmentation for al-la alloy microscopic images. *Symmetry* **2018**, *10*, 107. [CrossRef]
27. Zhang, S.; Chen, D.; Liu, S.; Zhang, P.; Zhao, W.C. Aluminum alloy microstructural segmentation method based on simple noniterative clustering and adaptive density-based spatial clustering of applications with noise. *J. Electron. Imaging* **2019**, *28*, 33035. [CrossRef]
28. Campbell, A.; Murray, P.; Yakushina, E.; Marshall, S.; Ion, W. New methods for automatic quantification of microstructural features using digital image processing. *Mater. Design* **2018**, *141*, 395–406. [CrossRef]
29. Campbell, A.; Murray, P.; Yakushina, E.; Marshall, S.; Ion, W. Automated microstructural analysis of titanium alloys using digital image processing. *IOP Conf. Ser. Mater. Sci. Eng.* **2017**, *179*, 012011. [CrossRef]
30. Zhenying, X.; Jiandong, Z.; Qi, Z.; Yamba, P. Algorithm Based on Regional Separation for Automatic Grain Boundary Extraction Using Improved Mean Shift Method. *Surf. Topogr. Metrol. Prop.* **2018**, *6*, 25001. [CrossRef]
31. Journaux, S.; Pierre, G.; Thauvin, G. Evaluating creep in metals by grain boundary extraction using directional wavelets and mathematical morphology. *J. Mater. Process. Technol.* **2001**, *117*, 132–145. [CrossRef]
32. Karpathy, A.; Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *39*, 664–676.

33. Zhao, G.; Ahonen, T.; Matas, J.; Pietikainen, M. Rotation-Invariant Image and Video Description with Local Binary Pattern Features. *IEEE Trans. Image Process.* **2011**, *21*, 1465–1477. [CrossRef]
34. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. BabyTalk: Understanding and Generating Simple Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *35*, 2891–2903. [CrossRef]
35. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. *Int. J. Comput. Vis.* **2015**. [CrossRef]
36. Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; van den Hengel, A. Visual Question Answering: A Survey of Methods and Datasets. *Comput. Vis. Image Underst.* **2017**, *163*, 21–40. [CrossRef]
37. Kazemi, V.; Elqursh, A. Show, Ask, Attend, and Answer: A Strong Baseline for Visual Question Answering. *arXiv* **2017**, arXiv:1704.03162.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Washington DC, USA, 2016.
39. Kwan, C.; Chou, B.; Yang, J.; Rangamani, A.; Tran, T.; Zhang, J.; Etienne-Cummings, R. Deep Learning-Based Target Tracking and Classification for Low Quality Videos Using Coded Aperture Cameras. *Sensors* **2019**, *19*, 3702. [CrossRef]
40. Kwan, C.; Chou, B.; Yang, J.; Tran, T. Deep Learning Based Target Tracking and Classification for Infrared Videos Using Compressive. *J. Signal Inf. Process.* **2019**, *10*, 167. [CrossRef]
41. Gers, F. Long Short-Term Memory in Recurrent Neural Networks. Ph.D. Thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland, 2001. [CrossRef]
42. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. In *European Association of Computational Linguistics (EACL)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017. [CrossRef]
43. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:1408.5882.