VQA Model: Based on Different Text Encoding Models and L2 Regularization

Zhengshu Zhang
College of Food Science
Sichuan Agricultural University
Yaan, Sichuan, China
Forestzhang@seu.edu.mk

***Abstract*—Visual Question Answering (VQA) , which is an important branch of artificial intelligence in deep learning has received the attention and investment of many researchers. The core of VQA model is the fusion of text encoding model, image encoding model, and multimodal vector. Currently, many advanced models have been proposed to perform VQA tasks on different datasets to continuously refresh higher model performance. This experiment uses VQA-V2, which is commonly used in VQA tasks, as the dataset to propose a model with Resnet-101 and different text encoding models to extract the features of image and text respectively. Recurrent neural networks (RNN, LSTM, GRU) and their attention attached models, as well as Transformer Encoder models, are used to encode text for VQA tasks. Based on the preliminary experimental results, select the models with better performance and add L2 regularization during its training process to solve overfitting problems to further improving model accuracy.**

Keywords—VQA, text encoding models, L2 regularization, model accuracy

I. INTRODUCTION

VQA, the abbreviation of the visual question answering, as a highly challenging and noteworthy interdisciplinary project in recent years, for it encompasses technologies such as computer vision, natural language processing, and deep learning[1]. The VQA task requires the model to generate corresponding answers based on the provided questions and images. Therefore, the VQA task is generally multimodal task, which requires the model to not only handle text problems with varying length sequences, but also handle rich and complex types of images. Finally, the performance of models can be determined by comparing the generated answers with standard answers based on the model.

The commonly used dataset for VQA project, such as COCO, CLVER, Meteor, VQA-V2, etc., can satisfied the needs of multi-modal tasks, which contain thousands of images and matching text questions and standard answers. This experiment uses the VQA-V2 dataset to provide images, questions, and standard answers. Some examples are shown in Figure 1. Due to the different types of problems, the difficulty of VQA tasks varies greatly, which greatly affects the accuracy of the model. Therefore, this experiment decided to focus on the performance of the model on lower difficulty numerical and counting problems. Therefore, in this experiment, the VQA task is divided into Number (number related problem), Count (count related problem), and All (all kinds of problems including number and count) based on the type of problem. The model calculates accuracy based on each of the three types of problems and conducts more in-depth analysis. For examples of different types of problems, we can refer to Fig.1. Where (a) belongs to Number, (b) belongs to Count and Number, both (a) (b) (c) and (d) belong to All.

(a)

Q: What number is the car? A: 7220

(b)

Q: How many horses are depicted? A: 2

(c)

Q: Is the big bear protecting a little bear? A: yes

(d)

Q: What kind of bread is that? A: muffin

Fig. 1. Examples of VQA task on VQA-V2

In the VQA model, the accuracy of the information contained in the obtained data encoding vector and the close dependency relationship with the image are key factors that affect the performance of the VQA model. As of now, some advanced problems or image coding models have been proposed and tested. The attention based convolutional neural network ABC-CNN model has achieved performance surpassing the latest models on many datasets [2].

This experiment selects three basic recurrent neural network models (traditional RNN, LSTM, GRU) and the most popular Transformer model to encode and generate vectors for VQA-V2 problems, and then obtain the accuracy of the model. It explores the impact of different models on VQA tasks and expects to achieve excellent performance. Overfitting is a common problem in deep learning during model training, mainly manifested as the high accuracy of the model on the training set and its low performance on the validation and testing sets due to poor generalization ability. At present, a more advanced method is to introduce the Specialized Dropout method into the model and effectively alleviate the overfitting problem in the model[3]. Therefore, this experiment added the most commonly used L2 regularization method based on overfitting during model training, striving to achieve higher model performance on the validation set without excessively increasing model complexity.

II. Related Work

Firstly, for the multi-modal task, Kelvin Xu and Jimmy Lei Ba proposed a model that introduces attention mechanism in order to enable machine learning models in image subtitle tasks to automatically learn and describe the content of images. The experiment showed that the hard attention model achieved the best BLEU values on all the three different kinds of datasets, while the soft attention model only achieved the best meter values on the Flickr30k and MS COCO datasets [4]. Steven J. Rennie and Etienne Margaret introduced Reinforcement learning to improve the performance of image caption system and named it self critical sequence training (SCST). The new model proposed by the team successfully increased the CIDEr value from 104.9 to 114.7 on the MS COCO evaluation server [5]. Marcella Cornia and Matteo Stefanini are considering improving the applicability of the Transformer model in multimodal works like the image captioning. The team developed an M2-a Meshed Transformer model with memory function. The results showed that the overall CIDEr and SPICE values of the model in COCO evaluation reached 75.0 and 11.4, respectively, ranking first on the online ranking [6]. In order to achieve higher-order intra modal and inter modal interactions between features, the Yingwei Pan and Ting Yao teams constructed an X-ray attention network based on X-ray attention blocks. The X-LAN model achieved the best CIDEr performance of 132.0% in COCO Karphy test splitting, and when further assigning linear attention blocks to the model Transformer, the CIDEr further increased to 132.8% [7]. Yang Xu, Yiheng Xu's team developed a LayoutLMV2 architecture that better captures cross modal interactions in order to model the interactions between text, layout, and images in a single multi-modal framework. The results show that the performance of the new layoutLMV2 model is significantly better than that of layoutLM, with F1 Score values increasing from 0.79 to 0.84 on FUNSD, 0.95 to 0.96 on CORD, 0.95 to 0.98 on SROIE, and 0.83 to 0.85 on Kleister-NDA [8]. Chia Wen Kuo and Zsolt Kira proposed that some detectors are not good at encoding the connection among the targets and image or scene level messgae, and the conditional connection among the objects and graphs lacks optimization. They proposed adding auxiliary information to represent the missing object relationship and using the multimodal trained model Clip to retrieve contextual descriptions to improve the performance of the model. The improved new model has a CIDEr of up to 7.5%, which is 1.3% higher than the existing baseline model [9]. The Nicola Messina team developed a new model to find the correct title among a large number of available titles using image related data. The team constructed a cascaded model of Mcprop and Crane models based on the XLM Robertsa and Clip models. The nDCG value of the new model is as high as 0.533, ranking fifth in the Kaggle Wikipedia Challenge [10].

Further, for the study of VQA related task, Huijuan Xu, Kate Saenko et al. believe that convolutional recursive networks have been applied to image visual Q&A, but have not been able to model spatial reasoning. Therefore, a recursive neural network model with explicit attention mechanism was proposed. The SMem-VQA Two Hop model with the best performance achieved an accuracy of up to 40.07% on the dataset DAQUAR. The results improved by 2.27% and 3.35% on the test dev and test standard of the VQA dataset compared to the baseline model, respectively [11]. Chen Zhu and Yanpeng Zhao et al. proposed that visual Q&A involves multi regional interaction, but existing models find it difficult to effectively encode this cross regional relationship. Based on this, the team proposed a structured attention mechanism. The new model incorporating attention mechanism improved by 9.5% contrast to the best baseline model on the CLEVR dataset and decreased by 1.25% contrast to the best baseline model on the VQA dataset [12].Zhengyang Wang and Shuiwang Ji believe that VQA has different requirements for text representation compared to NLP, so they propose using CNN instead of RNN to achieve text feature extraction. The experiment shows that the accuracy of the new model using "CNN Inception+Gate" is 61.33%, which is 0.98% higher than the baseline model LSTM [13]. Feng Liu believes that the iVQA model requires a high level of understanding of images and problem generation is a multi-modal dynamic reasoning process. Therefore, the team developed an model called iVQA which are supposed to modify the focal point step by step. The results showed that the CIDEr, ROUGE-L, and METEOR values of the new model on the test set reached 1.714, 0.468, and 0.205, respectively, all higher than those of various baseline models [14]. Minjoon Seo et al. found that unidirectional attention has been applied to machine understanding tasks, but this attention mechanism only focuses on local information and lacks the establishment of global relationships. Therefore, a model with bidirectional attention flow was proposed. The results showed that the new model achieved 76.3 and 76.9 val and test values for CNN in cloze tests, and 80.3 and 79.6 val and test values for Daily Mail, respectively, achieving the best results [15].

III. METHOD

A. Model

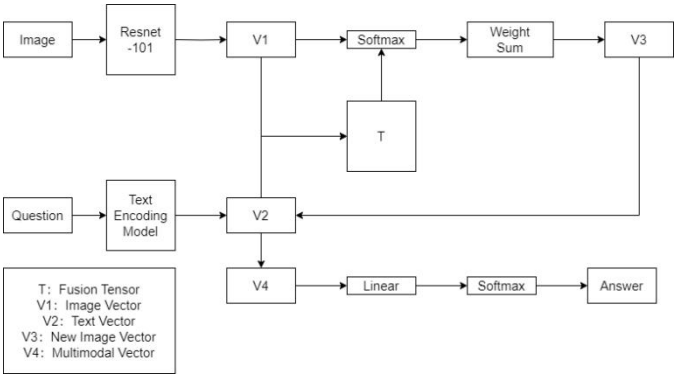

Fig. 2. The main structure of the whole model including the Resnet-101 mode, the text encoding model, feature fusion and the answer generation

Firstly, input the images and problems from the dataset into the Resnet-101 model and the text encoding model, respectively, to extract image feature vector 1 and text feature vector 2. Next, the method of calculating the sum of ReLU activation and the opposite number of the square of the difference between the two feature vectors is used to fuse and obtain the multi-modal feature tensor T. Afterwards, softmax function is used to obtain the probability distribution and a new image feature vector 3 is obtained by weighted sum. The information in vector 3 can better represent the image area related to the problem. Fusion feature vector 3 with text feature vector 2 to obtain the final multi-modal feature vector 4. Finally, the final vector is processed sequentially using the Linear layer and Softmax function to obtain an answer that includes a probability distribution. In this study, the main task is to attempt to select different text encoding models to update the overall model, in order to obtain a model with stronger performance. At the same time, L2 regularization is added based on the overfitting of these models in the training procedure to diminish this problem to further enhance the accuracy of the selected models.

B. Resnet-101

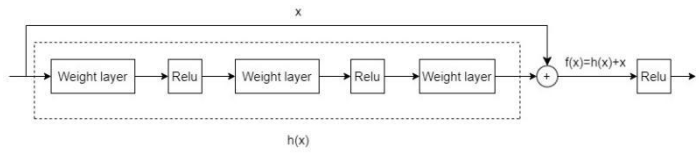

Fig. 3. The unit of Resnet-101. The model introduces residual block with a total of 101 convolutional layers. For there may be layers where the derivative values are difficult to converge. By introducing residual blocks, the gradient can skip this layer to solve the problem

This study uses the Resnet-101 model to extract image feature vectors. The Resnet-101 model belongs to the convolutional neural network architecture, which can encode input images into fixed length and dimensional tensors. The characteristic of this model compared to conventional convolutional neural networks is that by introducing residual blocks, skip connections can be achieved when gradients are transmitted to a certain layer, thereby solving the problems of gradient vanishing and representation bottlenecks to improve the accuracy of the results. The Resnet-101 model is a specific implementation of the Resnet model, with a depth of 101 layers and excellent performance in image feature extraction tasks. It's one of main operating step is convolution.

$$y(i,j) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} x(i+m, j+n) * k(m,n)$$

Where x is the input image; y is the output image; k is the convolutional kernel weight. In it, i and j are the coordinates in the output image while m and n are the coordinates in the convolutional kernel weight. M and N are the height and width of the convolutional kernel.

C. Text Encoding Model

1) RNN: RNN is one of the most basically used models for processing the text sequences. The traditional RNN model consists of three layers, output, input and hidden, characterized by a hidden state that can be continuously updated with sequence processing. The advantage of RNN is that each neuron in the structure receives the output of the last neuron as input and loops through this process, enabling RNN to effectively process data based on context information. The disadvantage is that there may be gradient vanishing and explosion problems when processing long sequence data.

*2) LSTM: T*he abbreviation of the Long Short-Term Memory. It is a special kind of recurrent neural network. It helps to decrease the gradient explosion and vanishing in traditional recurrent neural networks when processing long sequence data by introducing three gating devices, including forgetting gates, input gates, and output gates, to control information flow. In addition, compared to RNN, LSTM also has better generalization ability and stronger memory ability. Therefore, this experiment considers using LSTM to process text data in order to expect higher model performance.

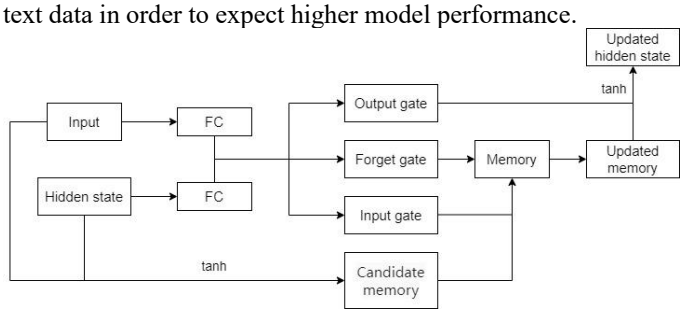

Fig. 4. The structure of LSTM.The input and hidden state are processed by the full connection layer and transferred to the output, forget and input gates. They also be processed by tanh active function to get the candidate memory. Memory based on the forget, input gate and candidate memory to calculate the updated memory. And output gate uses the updated memory to get the new hidden state

$$J_t = \sigma(X_t W_{xj} + H_{t-1} W_{hj} + b_j)$$

$$K_t = \sigma(X_t W_{xk} + H_{t-1} W_{hk} + b_k)$$

$$L_t = \sigma(X_t W_{xl} + H_{t-1} W_{hl} + b_l)$$

$$\widetilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$$

$$C_t = K_t \odot C_{t-1} + J_t \odot \widetilde{C}_t$$

$$H_t = L_t \odot \tanh(C_t)$$

Where J_t is the input gate; K_t is the forget gate; L_t is the output gate; W is the weight parameter; b is the bias parameter; H_t is the updated hidden state; H_{t-1} is the hidden state of the previous time step; $\widetilde{C}_t$ is the candidate memory cell; C_t is the updated memory cell; tanh is the active function.

3) GRU: The abbreviation of the Gated Recurrent Unit. It also belongs to the architecture of recurrent neural network. It introduces the gating mechanism of update gate and reset gate, which can capture the dependencies in sequence data, thus transforming the input text information into feature vectors. Compared with LSTM model, GRU model has fewer parameters and simpler logic, so it has advantages in time and Space complexity. In this experiment, the GRU model is used to convert the input text information into a fixed length feature vector for subsequent processing.

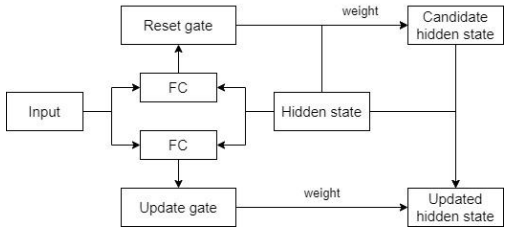

Fig. 5. The structure of GRU. The input and hidden state are processed by the full connection layer and transferred to the reset and update gates for weight calculation. Candidate hidden state and hidden state based on the weight values from update gate to get the result as new hidden state

$$M_t = \sigma(X_t W_{xm} + H_{t-1} W_{hm} + b_m)$$

$$N_t = \sigma(X_t W_{xn} + H_{t-1} W_{hn} + b_n)$$

$$\widetilde{H}_t = \tanh(X_t W_{xh} + (M_t \odot H_{t-1}) W_{hh} + b_h)$$

$$H_t = N_t \odot H_{t-1} + (1 - N_t) \odot \widetilde{H}_t$$

Where M_t is the reset gate; N_t is the updated gate; X_t is the input; W is the weight parameter; b is the bias parameter; H_t is the updated hidden state; H_{t-1} is the hidden state of the previous time step; $\widetilde{H}_t$ is the candidate hidden state.

4) Transformer Encoder: The Transformer model is currently widely used in VQA tasks. The encoder part of this model mainly utilized the vectors from embedding to obtain the features vectors of the information. Unlike traditional sequence processing models, the transformer adopts a self-attention mechanism, which can better capture long-distance dependencies in the sequence and make information extraction more accurate. This experiment uses the encoder part of the Transformer as the text encoding model to extract the text feature vector. The study are looking forward to get better performance while using the Transformer encoder with certain number of heads of multi-head attention.

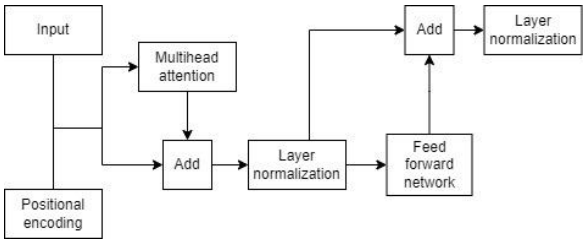

Fig. 6. The structure of Transformer encoder to obtain features vectors. The encoder part will processes the input and transmits the output to the multi-head attention and add module. The add module will send message to the layer normalization and make the result processed by the feedforward network. It will obscure a text feature vector eventually for later operation.

D. L2 Regularization

L2 regularization can to some extent alleviate the problem of reduced generalization ability caused by overfitting in the model, resulting in improved accuracy on the validation set. The method of L2 regularization is to set a norm penalty term during model training. In this experiment, multiple comparative experiments were conducted to determine the optimal performance with a setting of $1e^{-5}$.

$$J(\theta) = L(\theta) + \alpha R(\theta)$$

$$R(\theta) = \sum_i^n \theta_i^2$$

Where $J(\theta)$ is the loss function of L2 regularization term; $L(\theta)$ is the original loss function; $R(\theta)$ is the L2 regularization term; α is the regularization strength parameter.

E. Attention Mechanism

The Additive attention mechanism is one of the most basic attention mechanisms, commonly used in the processing of text sequences. Using this attention mechanism can enable the model to calculate the attention distribution based on the weighted sum of different elements when processing text, thereby we are supposed to improve the model's ability by apply it to RNN, LSTM and GRU model.

The Multi-head attention mechanism is an improved self attention mechanism that can use multiple attention mechanisms in parallel to capture different levels of text information by setting the number of heads in the Transformer model. The Transformer model using multi-head attention mechanism can help the model better understand contextual information, which to some extent helps to improve the performance of the model.

F. Answer Generation

Softmax is a commonly method as the Activation function in visual question answering tasks for answer generation. It can map the input vectors to a probability distribution between 0 and 1 and propose the best answer to the question according to the highest probability output. In the experiment, using the attention multi-modal feature vectors as input in order to allow the softmax function to calculate the value of possibility in order to get the final accuracy later.

$$softmax(v)_i = e^{v_i} \Big/ \sum_{j=1}^{k} e^{v_j}$$

Where v is the input vector; v_i is the i-th element in the input vector v; v_j represents the j-th element in the input vector v; k is the number of categories in the neural network.

IV. Result

The dataset used in this experiment is the VQA-V2 dataset, which contains over 100000 images and 600000 question answer pairs. The images are mainly sourced from the Microsoft-COCO dataset and are highly representative and valuable for research in VQA tasks.

In terms of evaluation metrics, commonly used Accuracy is used to measure model performance. Considering that overfitting may lead to higher accuracy in the training set but poor actual performance during model utilization, the accuracy on the training set is not considered, and the accuracy on the validation set is used as the only evaluation indicator of model performance.

A. Performance Of Different Text Encoding Models

TABLE I. The validation accuracy in three categories: Number, Count, and All on the VQA-V2 data set while using different text encoding models.The highest accuracy in the three categories is shown as rough

Model	Number(%)	Count(%)	All(%)
RNN	33.78	39.17	49.13
RNN(Additive Attention)	43.84	50.80	60.35
LSTM	47.75	55.18	**65.03**
LSTM(Additive Attention)	47.22	54.49	64.65
GRU	45.54	52.54	64.77
GRU(Additive Attention)	**48.47**	**56.06**	64.79
Transformer Encoder(2-Head)	42.21	48.97	59.59
Transformer Encoder(4-Head)	36.29	42.13	53.67
Transformer Encoder(6-Head)	29.72	34.26	42.19

- **RNN:** Use the RNN model to encode the problem. Use the Resnet-101 model to encode the graphics. The performance of this model on the VQA-V2 dataset is very poor, with an overall accuracy of only 49.13%.
- **RNN(Additive Attention):** Compared to traditional RNN model, RNN with additive attention mechanism has significantly improved the accuracy of problem text encoding in all three categories, with an improvement of about 10%.
- **LSTM:** The conventional LSTM model achieved the highest accuracy of 65.03% on All, and also achieved accuracy of 47.75% and 55.18% on the other two categories.
- **LSTM(Additive Attention):** The additive attention mechanism is added to the LSTM model to encode the problem, the performance of the model remains unchanged compared to conventional LSTM.
- **GRU:** Simplifying the model on the basis of LSTM and using GRU with simpler logic to encode the problem did not perform as accurately as the LSTM model, although it achieved good performance.
- **GRU(Additive Attention):** The GRU with additive attention mechanism has achieved significant performance improvement in accuracy in counting related issues. The highest accuracy of 48.47% was achieved in answering questions in the Number category; Achieved the highest score of 56.06% in the Count category.
- **Transformer Encoder(2-Head):** Attempt to use Transformer Encoder to partially replace RNN class models to encode text problems and fuse them with image features to obtain multi-modal feature vectors and predict answers.Setting the attention heads in the Transformer Encoder model to 2 results in model performance.
- **Transformer Encoder(4-Head):** Setting the attention heads in the Transformer Encoder model to 4 results in model performance.
- **Transformer Encoder(6-Head):** Setting the attention heads in the Transformer Encoder model to 6 results in model performance.

B. Performance Of Training With L2 Regularization

TABLE II. The validation accuracy in three categories: Number, Count, and All on the VQA-V2 data set by setting the weight decay to $1e^{-5}$. The data highlighted in red indicates an improvement(increase less than 0.5% ignored) in performance

Model	Number(%)	Count(%)	All(%)
RNN(Additive Attention)	45.72	52.93	61.47
LSTM	48.26	55.79	64.69
GRU(Additive Attention)	48.42	55.91	64.67
Transformer Encoder(2-Head)	43.82	50.77	60.04

In this stage of the experiment, one model with the best performance was selected from each series models in the previous part to participate in this part. Add L2 regularization during model training to obtain the accuracy results.

The LSTM model, RNN model with additive attention mechanism, and Transformer Encoder model with 2-head attention mechanism have all achieved varying degrees of accuracy improvement in numerical related problems such as Number and Count.

In terms of Number issues, they have increased by 0.51%, 1.88%, and 1.61% respectively; In terms of Count issues, they have increased by 0.61%, 2.11%, and 1.80%, respectively. Only the RNN model showed a significant improvement of 1.12% in ALL class problems. For the GRU model, the accuracy of all three types of problems did not show significant changes.

V. Discussion

A. Performance Differences Between Models

From the results of 4.1, it can be found that using different language models to encode the text sequence of the VQA-V2 dataset results in different accuracy of the overall model in answering questions. Due to the use of the same Resnet-101

model in the image encoding section and the bilinear fusion method for multi-modal feature fusion, there are only differences in text processing between different models, so this experiment can be considered as a control experiment. The experimental results mainly indicate the following conclusions:

1) The performance of the model using recurrent neural networks (RNN, LSTM, GRU) is generally better than that of the model using Transformer Encoder. Therefore, Transformer Encoder should not be considered as the text feature extraction model for this experiment. The possible reason is that the structure of the Transformer model is too complex, and its performance in processing long sequence data using parallel computing text is not as good as that of a simple structured recurrent neural network. Secondly, the recurrent neural network structure helps to better capture the long-distance dependencies of text problems, while the Transformer Encoder mainly captures these dependencies in limited contexts. Of course, the setting of some hyperparameters such as iteration count, batch size, learning rate, etc. may also greatly affect the performance of Transformer Encoder.

2) Adding additive attention mechanisms to certain recurrent neural networks (RNN, GRU) can significantly improve model performance. This may be due to the additive attention mechanism, which makes the model pay more attention to important regions in the text sequence, making the information in the feature vectors more abundant. At the same time, as one of the basic attention mechanisms, the additive attention mechanism has a simple structure and is not easy to make the model too complex, leading to a decrease in its accuracy.

3) The accuracy of the LSTM model in recurrent neural networks did not continue to increase after adding attention mechanisms, possibly due to the limitations of the basic architecture of the recurrent neural network model. The highest accuracy of 65.03% as the experimental limit is already the upper performance limit of this type of model. Under the influence of the barrel effect, adding attention mechanisms cannot compensate for the shortcomings of the recurrent neural network model itself, which can be supported by the latest performance of traditional RNN and GRU models with attention mechanisms not exceeding 65.03%. In the future, it is possible to consider optimizing the basic architecture of the LSTM model or using other non recurrent neural network models for text encoding, which is expected to further improve the accuracy of the model.

4) Finally, the performance of the Transformer Encoder is not ideal, and there is a significant gap between it and the recurrent neural network model. Considering its complex structure, large data computation, long single training time, and poor capture of long-distance text relationships compared to recurrent neural networks. More importantly, this experiment only used the Encoder part of the Transformer model to extract text features, and the Decoder part was not replaced, which may affect the performance of the model to some extent. Secondly, from the Transformer Encoder data, it can be seen that as the number of heads with multi head attention increases from 2 to 4 and then to 6, the performance of the model decreases significantly in sequence. This may be due to the information redundancy problem caused by the excessive number of heads, and the further increase in computational complexity of the model, resulting in a decrease in performance.

B. Performance Of Models With L2 Regularization

From the results of the second part, it can be observed that the performance of the model has changed after adding L2 regularization to the model with higher performance. This change is mainly reflected in the accuracy improvement of two types of numerical related problems, Number and Count. The possible reason is that the overfitting phenomenon in the model on digital issues is relatively severe, so the accuracy on the validation set has been significantly improved by adding weight attenuation. However, the overfitting phenomenon does not exist or is very mild in answering other categories of questions, resulting in little improvement in accuracy on All.

As discussed in section 4.1, the barrel effect of the LSTM model has not been further improved in all accuracy despite the use of L2 regularization. However, its performance in digital problems has improved slightly.

In addition to the GRU model, RNN, Transformer Encoder, LSTM, and other models have significantly improved their performance by adding L2 regularization to prevent overfitting. This also indicates that overfitting is a widespread problem in the process of model training, and incorporating L2 regularization into the training module is at least a decision that will not lead to performance degradation.

VI. CONCLUSION

This experiment uses VQA-V2 as the dataset, and uses RNN, LSTM, GRU and their models with additive attention mechanisms, as well as Transformer Encoder models with different heads to encode text problems. Fusion the text feature vector with the image feature vector encoded by the Resnet-101 model and ultimately generate the answer. The experimental results show that the text encoding model using GRU with Additive Attention achieved the highest performance in both numerical and countable problems, with accuracy of 48.47% and 56.06%, respectively; The text encoding model using LSTM achieved the highest performance on All class problems, with an accuracy of 65.03%.

Incorporating L2 regularization during the training process can improve the performance of some models on numerical and counting problems, but this effect is not significant for models with already good performance (GRU with Additive Attention). This experiment provides a reference for the selection of text encoding models for future research on other datasets, also indicates L2 regularization terms can be considered in the training when studying counting and numerical problems.

In the future study, on the one hand, we will try to use different image encoding models (such as VGG, DenseNet, ViT, etc.) to conduct in-depth research using similar methods

to further improve model performance. On the other hand, we are willing to start with multi-modal feature fusion methods and try to use different feature fusion methods (such as bilinear fusion, Gating-based Fusion, Tucker fusion, etc.) to improve the performance of the model. In summary, these methods that may improve model performance will be included in our future experimental scope.

REFERENCES

[1] Kafle, Kushal, and Christopher Kanan. "Visual question answering: Datasets, algorithms, and future challenges." Computer Vision and Image Understanding 163 (2017): 3-20.

[2] Chen, Kan, et al. "Abc-cnn: An attention based convolutional neural network for visual question answering." arXiv preprint arXiv:1511.05960 (2015).

[3] Wan, Kun, et al. "Reconciling feature-reuse and overfitting in densenet with specialized dropout." 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019.

[4] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. PMLR, 2015.

[5] Rennie, Steven J., et al. "Self-critical sequence training for image captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[6] Cornia, Marcella, et al. "Meshed-memory transformer for image captioning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[7] Pan, Yingwei, et al. "X-linear attention networks for image captioning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[8] Xu, Yang, et al. "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding." arXiv preprint arXiv:2012.14740 (2020).

[9] Kuo, Chia-Wen, and Zsolt Kira. "Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[10] Messina, Nicola, et al. "Transformer-Based Multi-modal Proposal and Re-Rank for Wikipedia Image-Caption Matching." arXiv preprint arXiv:2206.10436 (2022).

[11] Xu, Huijuan, and Kate Saenko. "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. Springer International Publishing, 2016.

[12] Zhu, Chen, et al. "Structured attentions for visual question answering." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[13] Wang, Zhengyang, and Shuiwang Ji. "Learning convolutional text representations for visual question answering." Proceedings of the 2018 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2018.

[14] Liu, Feng, et al. "ivqa: Inverse visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[15] Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." arXiv preprint arXiv:1611.01603 (2016).