Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 147 (2019) 374–380

Procedia Computer Science

www.elsevier.com/locate/procedia

2018 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2018

Image To Latex with DenseNet Encoder and Joint Attention

Jian Wang, Yunchuan Sun, Shenling Wang*

College of Information Science and Technology, Beijing Normal University, Beijing, 100875, P. R. China

Abstract

Mathematical formula structural analysis usually converts mathematical formulas in images into Latex codes. It has been named as Image2Latex by OpenAi. At present, many researchers use the model in the field of image captioning for image2latex and have achieved good results. In this paper, we propose some improvements to the baseline model which is a sequence-to-sequence model used in image caption. We improve the encoder by employing densely connected convolutional network (DenseNet) because it can strengthen feature extraction and facilitate gradient propagation. We propose to use a more effective joint attention mechanism which include both spatial attention and channel-wise attention to solve this problem. We conducted experiments on the dataset im2latex-100k. Experimental results showed that our model improved the performance of formula analysis.

Keywords: Image2latex; DenseNet; Saptial attention; Channel-wise attention;

1. Introduction

Optical Character Recognition (OCR) is commonly used to identify natural language from images. However, as early as the work of [1], there has been researchers working to convert images into structured languages or tags. The primary target for this research is OCR for mathematical expressions, and how to handle presentational aspects such as sub and superscript notation, special symbols, and nested fractions.

* Corresponding author. Tel.:+8613120198518.
E-mail address: slwang@bnu.edu.cn

10.1016/j.procs.2019.01.246

Nowadays, more researchers focus on recognition of complex structured markup language which has important theoretical meaning for the construction of modern OCR systems. Actually, recognition of complex structured markup language has been listed as a Requests for Research by OpenAi named as Image2Latex.In the past few years, some advances have been made in the areas of OCR in natural scenes [2] and image caption generation. In the image caption problem, the most commonly used model is called the sequence-to-sequence model wherein an encoder encodes a source sequence into feature vectors, which a decoder employs to produce the target sequence. In recent years, this architecture has been augmented with an attention and alignment model which selects a subset of the feature vectors for decoding. [3] proposed a novel attention based encoder-decoder model. Its general application inspires researchers that mathematical expression recognition can also be one proper application.

The model is completely data driven and does not require specific domain knowledge and can be widely applied to many similar problems. Many researchers have applied the image caption model to this problem of image2latex and achieved good results. In this paper, we use a baseline model which proposed in [4]. It employs a CNN+LSTM architecture with classic attention model. In order to improve the recognition effect, we improve the encoder and attention model by some novel approach. Experiments have shown our improvements are indeed useful. The main contributions of the paper include:

- Using densely connected convolutional networks (DenseNet) proposed by [5] to replace the CNN network in the model, enabling the model to encode images more efficiently.

- Using a joint attention mechanism called C-S attention model [6] to promote the attention mechanism in the model. Experiments have shown that the joint attention mechanism has a better performance compared to the classic attention model.

2. Related Work

In the past few years, there have been many important breakthroughs in computer vision and natural language processing. Many researchers have attempted to take breakthroughs on the task of recognizing mathematical expressions. [7] has proven that techniques combining neural network and standard NLP approaches like Conditional Random Field is very effective to recognize words in images and [8] used convolutional neural networks to recognize words in images. The sequence-to-sequence model based on the recurrent neural networks is often used for machine translation, speech recognition, and other fields. The introduction of attention by [9] improves the accuracy of the sequence-to-sequence model, and eventually established a new standard in Machine Translation systems. Inspired by the idea of sequence-to-sequence model, the authors in [10] construct the model in the fully end-to-end manner, which contains RNN with an attention module. It is a proper application in the field of OCR.

The model used in image caption have applied to Image2Latex. The image2latex model can generate the latex code corresponding to the input image of formulas. Inspired by recent work in Optical Character Recognition and image captioning, some end-to-end system is utilized which implements the recognition of image formulas. [11] took the similar approach and applied it to generate latex code of formulas from images. [12] presented a neural encoder-decoder model to convert images into presentational markup based on a scalable coarse-to-fine attention mechanism. [13] proposed a model namely WAP which learns to encoder input expression images and decode them into latex code.

Recently, The DenseNet has shown excellent performance on image classification task as it strengthens feature extraction and facilitates gradient propagation. In the work of [14], they improve the encoder by employing DenseNet and present a novel multi-scale attention model based encoder-decoder model for handwritten mathematical expression recognition. Based on the baseline model, we were trying to replace CNN encoder with DenseNet encoder and use a joint attention mechanism called C-S attention so as to achieve a better performance. Specifically, the C-S attention includes spatial attention and channel-wise attention.

3. Model

In the baseline model, the encoder employs a CNN network which consists of 6 convolution layers and the decoder is a LSTM with a general attention mechanism. We employ the CNN encoder by employing DenseNet and improve the attention mechanism to get a better performance.

3.1. DenseNet Encoder

We use the DenseNet to get the feature maps of the input image. DenseNet is a novel network which connects each layer to every other layer in a feed-forward fashion. In DenseNet, the feature maps of all previous layers are used as input and its own feature maps are used as input of all subsequent layers.

DenseNet has been proved that it performs better than some typical CNN networks in image classification such as ResNet. Specific to image2latex, baseline CNN encoder is hard to distinguish the difference of some tiny symbols while DenseNet have better performance. DenseNet have several advantages including: alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters. DenseNet improve the information flow between layers by introducing direct connections from any layer to all subsequent layers.

The feature maps of the l^{th} layer represent as:

$$x_l = H_l([x_0, x_1, \ldots, \; x_{l-1}]) \tag{1}$$

$[x_0, x_1, ..., x_{l-1}]$ is the input of the l^{th} layer and it refers to the concatenation of the feature-maps produced in layers $0, ..., l-1$. The function $H_l(\bullet)$ donates represents a composite function of three consecutive operations: batch normalization (BN), followed by a rectified linear unit (ReLU) and a convolution (Conv).

As an essential part of convolutional networks, down-sampling layers can change the size of feature-maps to increase receptive field and improve invariance. DenseNet introduce two substantial gradients to facilitate down-sampling in model. One of the gradient is called Dense Block which is multiple densely connected and the other is called Transition Layer which do convolution and pooling. We can reduce the number of feature-maps by half at transition layers. The network structure of DenseNet we used is shown in Figure 1.

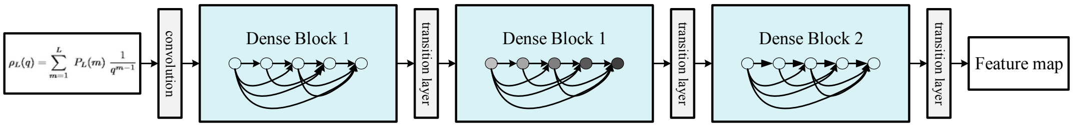

Fig. 1. DenseNet with three dense blocks

3.2. Decoder

The output of encoder is a feature map with size of $H \times W \times C$ .We can produce the latex tokens with a recurrent neural network once we receive the feature map. We use the LSTM network as our decoder because LSTMs have shown to be very efficient to capture long term dependencies and facilitate the back-propagation of gradients.

Formally, suppose that we want to generate the t^{th} word of the target latex tokens. We should take a hidden vector h_{t-1} and the previous token y_{t-1} as the input of out LSTM. In the structure of LSTM, the hidden state include both the hidden state and the memory.

The LSTM then give the probability of the next token with a recursive formula:

$$p(y_t) = f(y_{t-1}, h_{t-1}, c_t) \tag{2}$$

c_t is an attention vector which is computed at each time step and depends on the image and the next hidden vector. More specifically, we also compute another output state o_t used to compute the distribution probability over the formula as follows:

$$h_t = LSTM(h_{t-1}, [Ey_{t-1}, o_{t-1}]) \tag{3}$$

$$c_t = Attention(h_t, V) \tag{4}$$

$$o_t = \tanh(W^c[h_t, c_t]) \tag{5}$$

$$p(y_{t+1} \mid y_1, ..., y_t, V) = soft\max(W^{out} o_t) \tag{6}$$

Where W are weight matrix and E is an embedding matrix. V is encoder image by reshaping the feature maps. c_t only stand for the attention vector which is computed by attention mechanism. If we want to start generate tokens by LSTM, we should process the feature map with the following rule to enable input as an LSTM:

$$h_0 = \tanh(W \bullet (\frac{1}{n} \sum_{i=1}^{n} e_i) + b) \tag{7}$$

Where we learn an independent matrix W and bias b for each of the hidden states. We use two special tokens START and END. Once we get the initial state h_0 and the START token we can generate the next token step by step until our decoder predicts the END token.

3.3. Attention

We use a joint attention mechanism which integrated spatial attention and channel-wise attention to enhance the performance of the decoder. It was proved to help convergence in the case of image captioning.

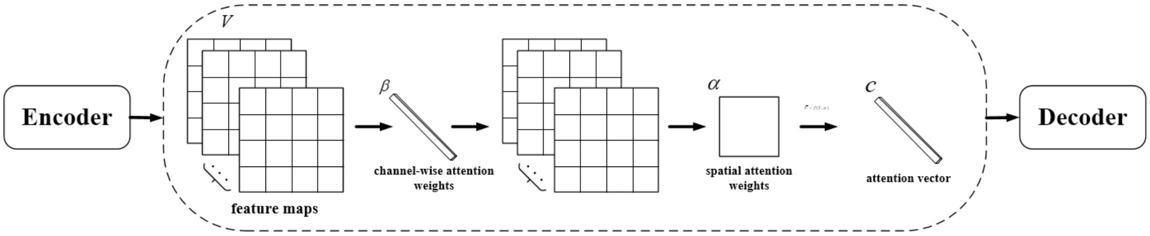

Fig. 2. C-S attention model

In general, a latex code usually relates to partial regions of an image of formula. Spatial attention model attempts to focus on the semantic-related regions. We represent the original feature map as $V = [v_1, v_2, ..., v_m]$, where $v_i \in \mathbb{R}^c$ and $m = W \times H$. We can regard v_i as the visual feature of the i^{th} location. We can get an attention distributions given the previous time step LSTM hidden state h_{t-1} .The spatial attention model can be expressed as:

$$a = \tanh((W_s V + b_s) + W_{hs} h_{t-1}) \tag{8}$$

$$\alpha\text{=softmax}(\text{W}_{is}\text{a+b}_{is}) \tag{9}$$

Actually, the model is a single-layer neural network followed by a softmax function. The transformation matrices $W_s \in \mathbb{R}^{k\times C}$,$W_{hs} \in \mathbb{R}^{k\times d}$,$W_{is} \in R^k$ is used to map features and hidden state to a same dimension. $b_s \in \mathbb{R}^k$,$b_{is} \in \mathbb{R}^1$ are model biases.

The focus of the channel-wise attention mechanism is different from the focus of the spatial attention mechanism. The feature map that extracted by encoder are usually multi-channel and each channel of a feature map is a response activation of the corresponding convolution filter. The channel-wise attention can help select semantic attributes. In channel-wise attention, we should reshape the original feature map V firstly. We reshape V to U , and $U=[u_1,u_2,...,u_C]$, where $u_i \in \mathbb{R}^{W\times H}$ donates the i^{th} channel of the feature map V , and C is the total number of channels. We obtain the channel feature v by applying mean pooling for each channel:

$$v=[v_1,v_2,...,v_C], v\in\mathbb{R}^C \tag{10}$$

v_i is a scalar which is the mean of vector u_i . It represents the i^{th} channel features. The channel-wise model can be defined as follows:

$$b=\tanh((W_c\times v+b_c)+W_{hc}h_{t-1}) \tag{11}$$

$$\beta = soft\max(W_{ic}b+b_{iC}) \tag{12}$$

Where $W_C \in \mathbb{R}^{k\times C}$,$W_{hC} \in \mathbb{R}^{k\times d}$,$W_{iC} \in \mathbb{R}^k$ are transformation matrices and $b_c \in \mathbb{R}^k$,$b_{iC} \in \mathbb{R}^1$ are model biases.

There exists two types of mechanism according to different implementation order of channel-wise attention and spatial attention. The first type dubbed Channel-Spatial(C-S) and the second type denoted as Spatial-Channel(S-C). In our experiment, we apply channel-wise attention before spatial attention.

The C-S model calculates two weights in order, and then computes the attention vector based on these two weights. At each time step of the decoding process, the attention mechanism attends to the feature map V and computes a weighted vector. φ_c donates the channel-wise attention model in the formula (13) and we obtain the channel-wise attention weights β from φ_c . Then we obtain the spatial attention weights α by feeding the channel-wise weighted feature map to the spatial attention model φ_s . In the formula (15), we compute a weighted average of these vectors which is also called attention vector. v_i represents each pixel position in the feature maps. Hence, our Decoder attends to the attention vector c at each time step of the decoding process. The calculation process of the C-S model can be expressed as figure 2.

$$\beta=\varphi_c(h_{t-1},V) \tag{13}$$

$$\alpha=\varphi_s(h_{t-1},f_c(V,\beta)) \tag{14}$$

$$c=\sum_{i=1}^{H\times W}\alpha,v_i \tag{15}$$

4. Experiment:

4.1. Dataset and Metric

We conducted experiments on im2latex-100k which includes a total of ~100k formulas and images. We splitted the dataset into train (~84k), validation (~9k) and test (~10k) sets. There exists a particular challenge because half of the formulas in im2latex-100k contain more than 50 tokens and the generated tokens depends on the beginning of the sentence and the previous tokens.

We evaluate our result by exact match (EM), BLEU score and edit distance. If the tokens of both a predicted formula and ground truth formula are exactly the same the exact match quantity is one. The approach of BLEU(Bilingual Evaluation understudy) considers that if the translation of the system is closer to the artificial translation result, the quality of its translation is higher. It captures the overlap of n-grams which appears in both system translations and reference translations. It has been shown to be the most correlated with human judgement and it is a standard metric in translation. We also use edit distance to evaluate the percentage of the reconstructed text that matches the original. The edit distance between a predicted formula and the ground truth tells us how many characters we should add/remove/change in one formula to obtain the other one. The edit distance is the percentage of the reconstructed text that matches the original. A perfect match has an edit distance of 1.

4.2. Comparison with baseline model

We tune the number of dense block and growth rate in DenseNet model to design two types DenseNet model. The most important difference between the two models is that the number of dense blocks is different. All of the growth rate in the two models was set to 24. More details of our two models is showed in Table 1.

Table 1. Details of two types of Dense Net model

Layers	DenseNet(2 blocks)	DenseNet(3 blocks)
Convolution	$\mathrm{conv}(7 \times 7)$	
Pooling	$\max\mathrm{pool}\ (3 \times 3)$	
Dense Block(1)	$\{\mathrm{conv}(1 \times 1), \mathrm{conv}(3 \times 3)\} \times 8$	$\{\mathrm{conv}(1 \times 1), \mathrm{conv}(3 \times 3)\} \times 8$
Transition Layer(1)	$\mathrm{conv}(1 \times 1) + \mathrm{average\ pool}(2 \times 2)$	
Dense Block(2)	$\{\mathrm{conv}(1 \times 1), \mathrm{conv}(3 \times 3)\} \times 16$	$\{\mathrm{conv}(1 \times 1), \mathrm{conv}(3 \times 3)\} \times 16$
Transition Layer(2)	$\mathrm{conv}(1 \times 1) + \mathrm{average\ pool}(2 \times 2)$	
Dense Block(3)	$\{\mathrm{conv}(1 \times 1), \mathrm{conv}(3 \times 3)\} \times 32$	
Transition Layer(3)	$\mathrm{conv}(1 \times 1) + \mathrm{average\ pool}(2 \times 2)$	
Decoder	LSTM	

The comparison among the proposed model and others on im2latex-100k is listed in table 2. The embedding size is set to 80. We use a small warm-up learning rate of 1e-4 for the two first epochs in order to solve high variance and high gradients at the beginning of the training. The size of beam is 5. DenseNet + C-S donates the model whose encoder is DenseNet with 2 blocks and attention mechanism is C-S attention model. The Model with DenseNet (2 blocks) only change the encoder compared to the baseline model. The Model with C-S attention only change the attention model in the baseline model. The decoders for all models in table 2 are LSTM .We use 10 epochs for a training on the whole dataset and testing on the test set. The EM, BLEU score and edit distance increase %8.41, %2.54 and %1.32 respectively by employing both DenseNet encoder and C-S attention model.

Table 2. Comparison of four models.

Model	EM(%)	BLEU(%)	Edit distance(%)
Baseline model	28.68	85.71	90.25
Model with DenseNet(2 blocks)	35.68	87.82	91.38
Model with C-S attention	31.79	86.54	90.75
DenseNet + C-S	**37.09**	**88.25**	**91.57**

5. Conclusion:

In this paper we improve the performance of caption-generation system applied to image2latex task by introducing the DenseNet encoder and C-S attention model. Both the DenseNet and C-S attention mechanisms can improve the performance of the seq2seq model in the image2latex task especially when the number of iterations is relatively small. However, the use of DenseNet and C-S attention model will consume more computing resources. We want to optimize the network structure for higher computing efficiency in the future work.

Acknowledgements

This research is sponsored by National Natural Science Foundation of China (No.61371185, 61401029, 61472044, 61472403, 61571049£¬61601033, 11401028) and the Fundamental Research Funds for the Central Universities (No.2014KJJCB32, 2013NT57) and by SRF for ROCS, SEM. and China Postdoctoral Science Foundation Funded Project (No.2016M590337).

References

[1] Anderson, Robert H. Syntax-directed recognition of handprinted two-dimensional mathematics. In Symposium on Interactive Systems for Experimental Applied Mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium, pp. 436–459. ACM, 1967.

[2] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading Text in the Wild with Convolutional Neural Networks[J]. International Journal of Computer Vision, 2016, 116(1):1-20.

[3] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[4] Genthial, G., & Sauvestre, R. (2016) Image to Latex.

[5] Huang G, Liu Z, Laurens V D M, et al. Densely Connected Convolutional Networks[J]. 2016:2261-2269.

[6] Chen L, Zhang H, Xiao J, et al. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning[J]. 2017:6298-6306.

[7] Jaderberg M, Simonyan K, Vedaldi A, et al. Deep Structured Output Learning for Unconstrained Text Recognition[J]. Eprint Arxiv, 2014, 24(6):603–611.

[8] Wang T, Wu D J, Coates A, et al. End-to-end text recognition with convolutional neural networks[C]// International Conference on Pattern Recognition. IEEE, 2013:3304-3308.

[9] Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation[J]. Computer Science, 2015.

[10] Lee C Y, Osindero S. Recursive Recurrent Nets with Attention Modeling for OCR in the Wild[C]// Computer Vision and Pattern Recognition. IEEE, 2016:2231-2239.

[11] Deng Y, Kanervisto A, Rush A M. What You Get Is What You See: A Visual Markup Decompiler[J]. 2016.

[12] Deng Y, Kanervisto A, Ling J, et al. Image-to-Markup Generation with Coarse-to-Fine Attention[J]. 2017.

[13] Zhang J, Du J, Zhang S, et al. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition[J]. Pattern Recognition, 2017, 71: 196-206.

[14] Zhang J, Du J, Dai L. Multi-scale attention with dense encoder for handwritten mathematical expression recognition[J]. arXiv preprint arXiv:1801.03530, 2018.