DEEP TRANSFORMERS FOR FAST SMALL INTESTINE GROUNDING IN CAPSULE ENDOSCOPE VIDEO

Xinkai Zhao[1], Chaowei Fang[2], Feng Gao[3], De-Jun FAN[3], Xutao Lin[3], Guanbin Li[1]

[1]School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China
[2]School of Artifical Intelligence, Xidian University, Xi'an, China
[3]The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

ABSTRACT

Capsule endoscopy is an evolutional technique for examining and diagnosing intractable gastrointestinal diseases. Because of the huge amount of data, analyzing capsule endoscope videos is very time-consuming and labor-intensive for gastrointestinal medicalists. The development of intelligent long video analysis algorithms for regional positioning and analysis of capsule endoscopic video is therefore essential to reduce the workload of clinicians and assist in improving the accuracy of disease diagnosis. In this paper, we propose a deep model to ground shooting range of small intestine from a capsule endoscope video which has duration of tens of hours. This is the first attempt to attack the small intestine grounding task using deep neural network method. We model the task as a 3-way classification problem, in which every video frame is categorized into esophagus/stomach, small intestine or colorectum. To explore long-range temporal dependency, a transformer module is built to fuse features of multiple neighboring frames. Based on the classification model, we devise an efficient search algorithm to efficiently locate the starting and ending shooting boundaries of the small intestine. Without searching the small intestine exhaustively in the full video, our method is implemented via iteratively separating the video segment along the direction to the target boundary in the middle. We collect 113 videos from a local hospital to validate our method. In the 5-fold cross validation, the average IoU between the small intestine segments located by our method and the ground-truths annotated by broad-certificated gastroenterologists reaches 0.945.

Index Terms— Capsule Endoscopy, Small Intestine Grounding, Transformer, Convolutional Neural Network

1. INTRODUCTION

Capsule endoscope (CE) is a disposable wireless imaging device which is widely used for disease diagnosis in the entire gastrointestinal (GI) tract. Convenience and noninvasiveness are the main superiorities of capsule endoscope, compared to enteroscopy [1, 2]. In particular, the small intestine mucosa is clearly visualized in videos captured by capsule endoscopes.
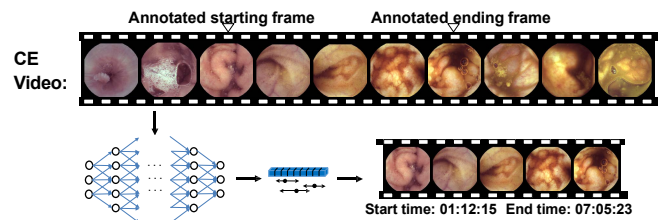

Fig. 1: Illustration of the small intestine grounding task. Given a video of capsule endoscope, the grounding task aims at identifying the starting and ending time positions of the small intestine.

There are lots of lesions that may exist in the small intestine, such as Crohn's disease, ulcers, angioectasias, polyps, and bleeding lesions. With the help of the capsule endoscope, medicalists are able to look into the inner environment of the small intestine and make more accurate disease diagnosis. However, a full CE video usually occupies over 10 hours and contains over 100,000 frames. It is very time-consuming to screen out lesions from the input video. Devising automatic machine learning machines to locate the small intestine is valuable as it can greatly reduce the time consumption of subsequent related disease diagnosis. With the development of artificial intelligence, extensive studies have reported the promising performance of this novel technology for the recognization of various small intestine diseases. Aoki *et al.* [3] use CNNs to detect erosions and ulcerations in CE images. Ding *et al.* [4] use CNNs to identify abnormalities in CE images. [5] shows that CNNs can reduce the reading time of endoscopists. In this paper, we focus on settling the grounding task of the small intestine in a full CE video as illustrated in Fig. 1. This is the first attempt to attack this task via deep learning.

To solve the small intestine grounding problem, we propose an efficient search approach on the basis of a convolutional neural network (CNN). First of all, we separate the whole video into three components, esophagus/stomach, small intestine and colorectum. A CNN model is built up to fulfill the 3-way video frame classification task. It is com-

posed of three core modules. The backbone of an existing CNN classification such as ResNet [6] and DenseNet [7] is regarded as the encoder for extracting feature representations of video frames. The category of every frame is highly related to the categories of preceding and subsequent neighboring frames. For example, the small intestine never comes after the large intestine. Hence, a Transformer module [8] is adopted to aggregate features from neighboring frames. Finally, a classification head is used for predicting the category confidence scores. Considering there exists tens of thousands of frames in a full capsule video, it is time consuming to infer the classes of all frames and then seek the small intestine section in brute force. We apply a search algorithm to locate the left and right boundaries of the small intestine. Starting from the middle frame of the whole video, our devised algorithm approaches the target boundary iteratively, reducing the potential boundary positions by half. A dataset containing 113 CE videos is collected from a local hospital, and our method achieves promising grounding results with average IoU of 0.945 under the 5-fold cross validation.

2. METHOD

The target of this paper is to ground the small intestine in a full capsule endoscopy video. Provided a video consisting of T frames $\{\mathbf{I}_t | t = 1, \cdots, T\}$, we propose a deep learning based algorithm to locate the starting and ending temporal positions (denoted as t_s and t_e respectively) of the small intestine. To solve this problem, we first define a 3-way classification problem. Practically, all video frames are classified into 3 categories, including esophagus/stomach, small intestine and colorectum. For frame $\mathbf{I}_t$, we denote the ground-truth class as $y_t \in \{1, 2, 3\}$ and the predicted category confidences as $\mathbf{p}_t \in \mathcal{R}^3$. $y_t = 1$, $y_t = 2$ and $y_t = 3$ indicates frame $\mathbf{I}_t$ belongs to esophagus/stomach, small intestine and colorectum respectively. To explore inter-frame relationship, deep Transformer is employed to enhance the feature representation of every frame with its neighboring frames. Furthermore, a search algorithm is devised to locate the small intestine. The overall pipeline of our method is shown in Fig.2.

2.1. Network Architecture

First of all, we adopt the backbone of an existing classification network as an encoder for extracting the feature representation of every single frame. For example, the ResNet [6] or DenseNet [7] can be modified as the encoder after removing the last fully connected layer for class probability prediction. Specifically, we first train a classification network to obtain the features of the frame. Given a video frame $\mathbf{I}_t$, we denote the extracted feature representation as $\mathbf{f}_t$.

To capture the temporal inter-frame dependencies, the Transformer module is used to enhance the feature representation of $\mathbf{I}_t$ with multiple neighboring frames. Here, the features of $2N + 1$ frames $\{\mathbf{f}_i\}_{i=t-N}^{t+N}$ are regarded as the input of the Transformer module. Then, we use three fully connected layers to generate the key, query and value vectors for all frames. We denote the key, query, value vector calculated from $\mathbf{f}_i$ as $\mathbf{k}_i$, $\mathbf{q}_i$, and $\mathbf{v}_i$, respectively. Assume A be the $(2N + 1) \times (2N + 1)$ attention matrix depicting the relation between every pair of frames. The relation attention between the i-th and j-th frame is estimated as,

$$A_{i,j} = \mathbf{q}_i^T \mathbf{k}_j / \sqrt{m}, \qquad (1)$$

where m is the dimension of the feature representation. Every row of $\mathbf{A}$ is normalized by the softmax function. Then, a new value vector of the i-th frame is generated through a matrix multiplication, $\hat{\mathbf{v}}_i = \sum_j A_{i,j} \mathbf{v}_j$.

To better enhance attention, we use multiple self-attention heads in the transformer. Specifically, 8 heads are adopted to produce 8 new value vectors for every frame, $\{\hat{\mathbf{v}}_i^m\}_{m=1}^8$. These vectors are concatenated, and compressed into a new feature representation $\hat{\mathbf{f}}_i$ for the i-th frame via a fully connected layer and a residual block consisting of two fully connected layers. The features of $2N + 1$ frames, $\{\hat{\mathbf{f}}_i\}_{i=t-N}^{t+N}$, are fed into another two fully connected layers, yielding the final prediction $\mathbf{p}_t$ of the t-th frame.

The training process is composed of two stages. First, the encoder is optimized as a single frame classification model. A linear layer is used to predict category confidences from the feature of every single frame. In the second stage, the parameters of the encoder is frozen and the parameters of the transformer is optimized. The cross entropy function is adopted to calculate training losses for both stages.

2.2. Search Algorithm

To locate the shooting range of the small intestine, we can infer the categories of all frames and then find out the starting and ending frames of the small intestine in brute force. However, this process is inefficient and source-intensive. To increase the efficiency of locating the small intestine with little accuracy loss, we devise a fast and fault-tolerant searching algorithm. With this algorithm, even if some frames are missclassified, the locating accuracy is still satisfactory.

The procedure for searching the ending position is illustrated in Algorithm 1. Starting from the middle point of the whole video, the searching position t is repeatedly updated to approach the switching point between the left small intestine and the colorectum. The updating direction depends on the inferred category of the image at current position. If $\mathbf{I}_t$ belongs to the colorectum, the target position should be later than t; otherwise, the target position is earlier than t. The searching interval d, which is initialized as $T/2$, is decayed by α after every step. In order to ensure the algorithm can tolerate some classification errors, the key is the α must be greater than 0.5.

In practice, the confidence value of the inferred class is used to weigh the updating stride. If the inferred class is un-
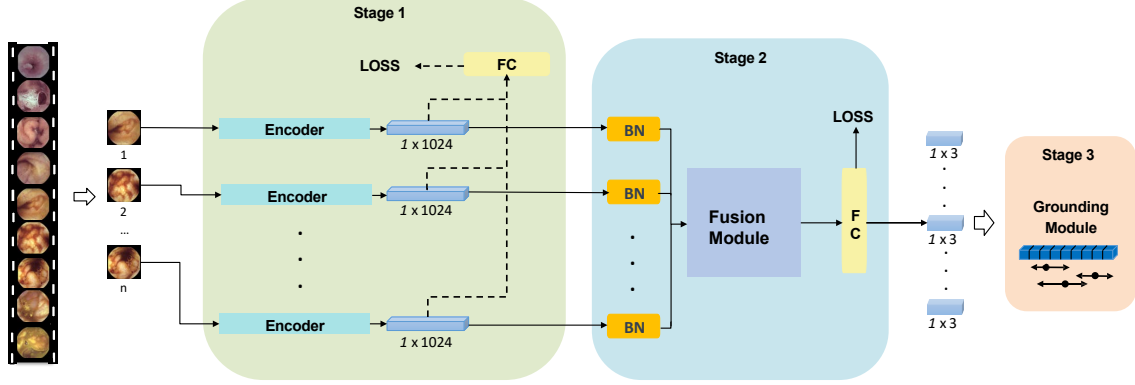
Fig. 2: The entire pipeline can be divided into three stages, and after each stage, the trained parameters will be fixed and directly applied to the next stage.

Algorithm 1: Algorithm for searching the ending frame of small intestine.

Input: CE video frames, $\{\mathbf{I}_i\}_{i=1}^{T}$.
Output: The ending frame index, t_e.
Init: Frame index, $t = roundint(0.5 * T)$; searching interval, $d = roundint(0.5 * T)$.

repeat

    Use the network in Section 2.1 to predict the category confidence $\mathbf{p}_t$, regarding $\{\mathbf{I}_i\}_{i=t-N}^{t+N}$ as the input;

    $c_t \leftarrow \arg\max_{c \in \{1,2,3\}} p_t[c]$;

    if $c_t == 1$ *or* $c_t == 2$ **then**

        | $t \leftarrow \text{roundint}(t + d * (max(p_t[c_t] - \theta, 0) + \epsilon))$

    else

        | $t \leftarrow \text{roundint}(t - d * (max(p_t[c_t] - \theta, 0) + \epsilon))$

    end

    $d \leftarrow \text{roundint}(\alpha * d)$

until $d < 1$;

$t_e \leftarrow t$

---

confident, a small stride is adopted to update the searching position. This is beneficial for increasing the robustness a searching algorithm against incorrect predictions. We determine the category of $\mathbf{I}_t$ as, $c_t = \arg\max_{c \in \{1,2,3\}} p_t[c]$ where $p_t[c]$ is the confidence value of the c-th class. The updating stride of the searching position is defines as the following formulation,

$$\Delta t = \text{round}(\alpha * d * (\max(p_t[c_t] - \theta, 0) + \epsilon)), \quad (2)$$

where θ and ϵ are constants.

The algorithm of searching the starting position of small intestine can be easily obtained via adjusting the conditions of deciding the updating direction.

3. EXPERIMENTS

3.1. Dataset

We collect 113 CE videos from the Sixth Affiliated Hospital, Sun Yat-sen University. These videos are captured by Miro-Cam®. The frame rate is 3fps and the average length is about 11 hours 35 minutes. The resolution of every video frame is $320 * 320$. Each video is composed of the shooting ranges of esophagus, stomach, small bowel, and large bowel in chronological order.

We use 5-fold cross validation to evaluate the performance of the grounding method. We counted the distribution of labels for the 113 samples. In terms of frame numbers, the esophagus and stomach occupy about 7.2% of the frames, the small intestine 44.9%, and the large intestine 47.9%.

3.2. Implementation Details

In the first phase, we choose a learning rate of 10^{-3} and a weight decay of 10^{-2}, and in the second phase, we choose a learning rate of 10^{-5} and a weight decay of 10^{-5} with $N = 6$ samples to train the fusion module. Adamw optimizer is used in both phases, and the loss function is cross-entropy loss. In the third phase, α is set to 0.9 , ϵ is set to 0.01, and the threshold θ is set to 0.5.

3.3. Quantitative Analysis

We present several experiment results comparing other models, i.e. VGG, MobileNet, without inter-frame feature fusion. We also conduct inner comparisons based two backbones including DenseNet and ResNet. We put the features of $2N+1$ frames $\{\mathbf{F}_i\}_{i=t-N}^{t+N}$ into bi-directional long short-term memory (LSTM) and transformer encoder (TFE) to compare the efficacy of fusion methods for classification. The results of multi-class classification are presented in Table 1. Accuracy in Table 1 means micro-averages and accuracy* means

Method	IoU	$Accuracy$	$Accuracy*$
VGG	0.601 ± 0.229	0.724	0.570
MobileNet	0.800 ± 0.276	0.805	0.571
DenseNet	0.926 ± 0.124	0.896	0.872
ResNet	0.932 ± 0.092	0.887	0.885
DenseNet + LSTM	0.932 ± 0.089	0.897	0.890
ResNet + LSTM	0.944 ± 0.065	0.930	0.893
DenseNet + TFE	0.938 ± 0.114	0.917	0.891
ResNet + TFE	0.945 ± 0.086	0.908	0.911

Table 1: IoU of small-bowel grounding and accuracy of classification

		Ground Truth		
		1	2	3
	1	89.4	4.6	0.3
ResNet	2	8.8	92.9	11.3
	3	1.9	2.5	88.4
Prediction				
	1	92.4	2.5	0.3
ResNet-TFE	2	7.2	93.1	7.3
	3	0.4	4.4	92.3

Table 2: Confusion Matrix of ResNet and ResNet-TFE. The numbers represent percentages of images.

macro-average. Micro-average and macro-average accuracies are calculated in image level and category level respectively. If the i-th class has n_i samples, of which c_i samples are correctly predicted,

$$accuracy = \frac{\sum_{i=1}^{3}(c_i)}{\sum_{i=1}^{3}(n_i)}, \quad accuracy* = \frac{1}{3}\sum_{i=1}^{3}(\frac{c_i}{n_i}) \quad (3)$$

Table 2 shows the confusion matrix for ResNet and ResNet-TFE, respectively. The results show that the performance is good and balanced. Moreover, the results of ResNet-TFE are slightly better than those of ResNet.

3.4. Grounding Results

Our proposed algorithm costs 109 seconds averagely to locate the small intestine in full CE videos. Table 1 also shows the grounding results for each method. Intersection over Union (IoU) is defined as the intersection of the prediction and ground truth divided by the union between them. The results show that the methods devised by us have better performance.

In Fig. 3, we present the boxplot about the absolute deviation between the predicted cutoff point and the ground truth. The results show that the error between the predicted and true values is within 100 frames, for most of the sample. The points represents the means, which are between 500 and 1200 because of some outlier points. And the median is less than 50 for both start and end points. This indicates that our method achieving promising results in small intestine grounding.
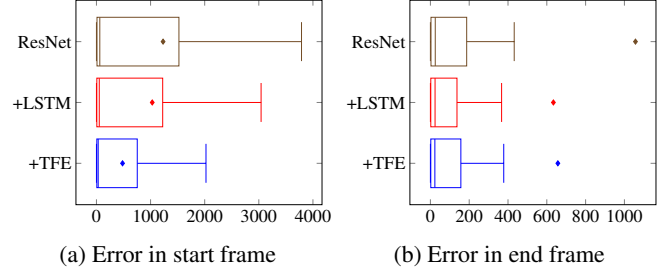

(a) Error in start frame   (b) Error in end frame
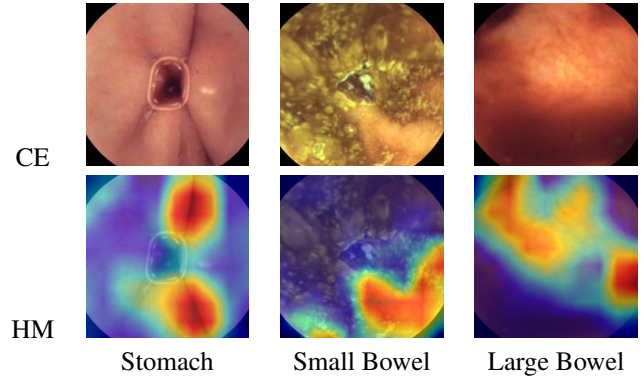
Fig. 3: Error in start and end frame

Table 3: Examples of three types of labels and corresponding heat maps (HM) are shown. In these examples, the results for the stomach are more relevant to the folds, the small intestine is judged to be largely dependent on the mucosa, and the large intestine is more concerned with the texture of the inner wall.

3.5. Model Interpretability

In this section, we use class activation map (CAM) to visualize which part of information is discriminative to the classification network. Table 3 shows the CAM of some samples, and it can be seen that the network mainly judges the small bowel based on textural information on the mucosa, which is consistent to manual judgment.

4. DISCUSSION AND CONCLUSION

We propose a novel algorithm locate the shooting range of the small intestine in a full CE video. To the best of our knowledge, it is the first work to solve this problem using deep neural networks. In our approach, we used the transformer encoder for feature fusion between the features of different frames. Besides, we present an efficient algorithm to search the starting and ending frame of the small intestine based on a classification network, which achieves good results in terms of efficiency and accuracy.

153

5. REFERENCES

[1] Gavriel Iddan, Gavriel Meron, Arkady Glukhovsky, and Paul Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, pp. 417–417, 2000.

[2] Stuart L Triester, Jonathan A Leighton, Grigoris I Leontiadis, Suryakanth R Gurudu, David E Fleischer, Amy K Hara, Russell I Heigh, Arthur D Shiff, and Virender K Sharma, "A meta-analysis of the yield of capsule endoscopy compared to other diagnostic modalities in patients with non-stricturing small bowel crohn's disease," *American Journal of Gastroenterology*, vol. 101, no. 5, pp. 954–964, 2006.

[3] Tomonori Aoki, Atsuo Yamada, Kazuharu Aoyama, Hiroaki Saito, Akiyoshi Tsuboi, Ayako Nakada, Ryota Niikura, Mitsuhiro Fujishiro, Shiro Oka, Soichiro Ishihara, et al., "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal endoscopy*, vol. 89, no. 2, pp. 357–363, 2019.

[4] Zhen Ding, Huiying Shi, Hao Zhang, Lingjun Meng, Mengke Fan, Chaoqun Han, Kun Zhang, Fanhua Ming, Xiaoping Xie, Hao Liu, et al., "Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model," *Gastroenterology*, vol. 157, no. 4, pp. 1044–1054, 2019.

[5] Tomonori Aoki, Atsuo Yamada, Kazuharu Aoyama, Hiroaki Saito, Gota Fujisawa, Nariaki Odawara, Ryo Kondo, Akiyoshi Tsuboi, Rei Ishibashi, Ayako Nakada, et al., "Clinical usefulness of a deep learning-based system as the first screening on small-bowel capsule endoscopy reading," *Digestive Endoscopy*, vol. 32, no. 4, pp. 585–591, 2020.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

Acknowledgments

Compliance with Ethical Standards

This is a numerical simulation study for which no ethical approval was required.