Multi-modal video retrieval using Dilated Pyramidal Residual network

La Ngoc Thuy An, Nguyen Phuoc Dat, Pham Minh Nhut, Vu Hai Quan

Abstract—**Pyramidal Residual Network achieved high accuracy in image classification tasks. However, there is no previous work on sequence recognition tasks using this model. We presented how to extend its architecture to form Dilated Pyramidal Residual Network (DPRN), for this long-standing research topic and evaluate it on the problems of automatic speech recognition and optical character recognition. Together, they formed a multi-modal video retrieval framework for Vietnamese Broadcast News. Experiments were conducted on caption images and speech frames extracted from VTV broadcast videos. Results showed that DPRN was not only end-to-end trainable but also performed well in sequence recognition tasks.**

Keywords—**Dilated Pyramidal Residual Network, video retrieval, multi-modal retrieval, Vietnamese broadcast news**

1. INTRODUCTION

Deep Convolutional Neural Network played an important role in solving complex tasks of automatic speech recognition (ASR) [1], natural language processing (NLP) [2], and optical character recognition (OCR) [3], etc. Its capability could be controlled by varying number of stacked layers (depth), receptive window size, and stride (breadth). Recent researches have focused on the depth and lead remarkable results on the challenging ImageNet dataset [4]. Residual Network (ResNet), introduced by He [5], was the most successful deep network following this approach for filter stacking. By addressing the degradation problem, it could easily gain the accuracy from increased layers. According to the research of Han [6], layer sizes increasing gradually as a function of the depth, like a pyramid structure, could improve the performance of the model. Han proposed a modification and a new model based on ResNet: Deep Pyramidal Residual Network (PyramidNet).

Both ResNet and PyramidNet have been evaluated on ImageNet dataset. Here, we considered extending PyramidNet for sequence recognition tasks. Instead of a single label, recognizing a sequence-like object produced a series of labels. That required alignment first, while image classification tasks could be directly processed by neural network models. This is where the CTC (Connectionist Temporal Classification) loss function [7] kicks in. We use PyramidNet as hidden layers and feed it to the loss function, so that our model was an end-to-end network. In addition, we modified PyramidNet by applying dilated convolution [8], allowing it to aggregate context rapidly, which was important for sequence recognition tasks, because of the need to remember many details of the sequence at hands.

To show that our extended version of PyramidNet could work on sequence recognition tasks, we opted in the application of multi-modal video retrieval which was then decomposed into 2 sub problems: automatic speech recognition (ASR) and optical character recognition (OCR). Experiments were conducted on the Vietnamese broadcast news corpus recorded from the VTV programs.

2. METHOD

Multi-modal video retrieval

Received 10-07-2018, accepted 10-09-2018, published 20-11-2018

La Ngoc Thuy An, Nguyen Phuoc Dat, Pham Minh Nhut, Vu Hai Quan – University of Science, VNU-HCM
**Email: vhquan@vnuhcm.edu.vn*

Video was a self-contained material which carries a large amount of rich information, far richer than text, audio or image. Researches [9-12] have been conducted in the field of video retrieval amongst which content-based retrieval of video event was an interesting challenge. Fig. 1 illustrated a multi-modal content-based video retrieval system which typically combined text, spoken words, and imagery. Such system would allow the retrieval of relevant clips, scenes, and events based on queries which could include textual description, image, audio and/or video samples. Therefore, it involved automatic transcription of speech, multi-modal video and audio indexing, automatic learning of semantic concepts and their representation, advanced query interpretation and matching algorithms, which in turn imposed many challenges to research.

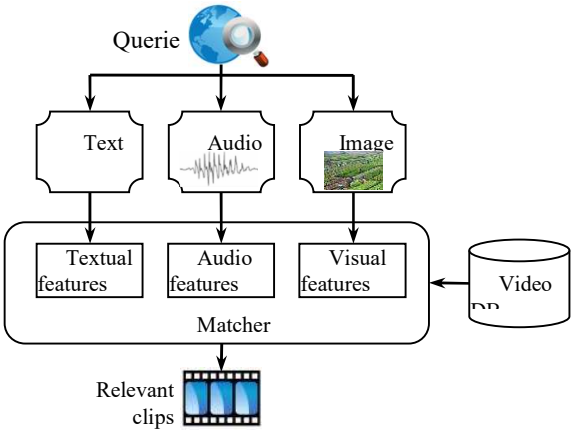

Fig. 1. Multi-modal content-based video retrieval

Indeed, there were a lot of directions to come in. Nevertheless, we chose to tackle the problem by combining both spoken and written information. To be precise, the whole speech channel was transcribed into spoken words, and caption images (Fig. 2) were decoded to written words whenever they appear. While captions represented the news topics and sessions, spoken words beared the detailed content themselves. This combination was the key indexing mechanism for content retrieval, leading us to the domains of ASR and OCR.

Fig. 2. Caption images representing the news topics

Dilated pyramidal residual network

This section provided a detailed specification on how we construct our network – the dilated pyramidal residual network (DPRN). DPRN itself was an extension of PRN, giving it the ability to classify a sequence of samples such as speech and optical characters. This was done by equipping PRN with CTC and dilated convolution to handle sequence recognition. Details were given in the following subsections.

Pyramidal Configuration

Our network consisted of multiple feed-forward neural network groups. Each group was essentially a pyramidal residual network (PyramidNet), using the same definition as in Han [6]. Down-sampling was a convolutional neural network (CNN) with a kernel size of 1, which reduced the sequence length but increased the feature map dimension.

In order to reduce the feature map for CTC loss function, the last block of our entire model was an average pooling module. We called our network Dilated Pyramidal Neural Network (DPRN). A schematic illustration was shown in Fig. 3. We built the network with increasing feature map dimension per block. Instead of following a set formula, we increased them by a flat amount.

Building block

Building block was the core of any ResNet-based architecture. It was actually a feed-forward neural network stack with DCNN (Deep Convolutional Neural Network), ReLUs (Rectified Linear Units) and BN (Batch Normalizations). Hence, in order to maximize the capacity of our model, designing a good building block was essential. We followed the building block of Han, as they provide the best

result with our architecture. We also used zero-padded identity-mapping of all blocks. These concepts were outlined in Fig. 3.

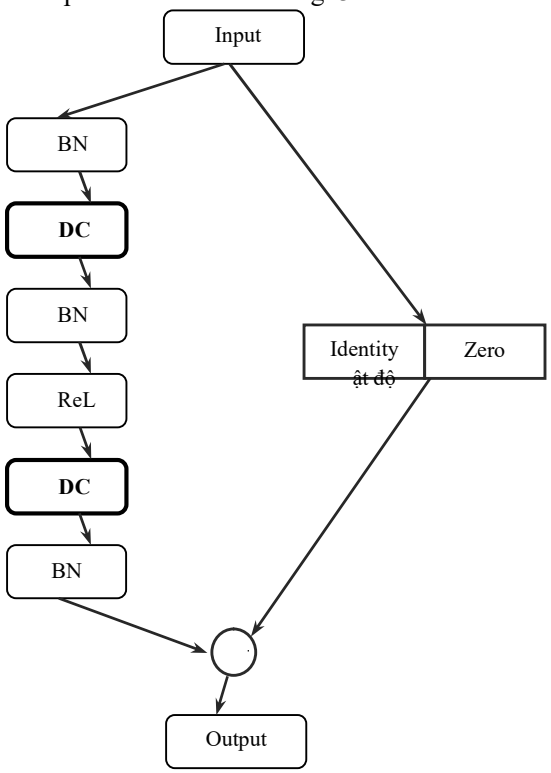

Fig. 3. Building blocks

Dilated Convolution

Dilated convolution was basically the same as normal convolution operation, but with a small difference. The convolution in dilated convolution was defined in a wider receptive field by skipping over n inputs at a time, where n was the width of dilated convolution. Therefore, $n = 1$ was equivalent to normal convolution, and the $n > 1$ convolution could take much larger context than normal convolution.

Stacking multiple dilated layers easily increased maximum learnable context. For instance, the receptive field of each vector after 4 layers of dilated convolution was 31 when each layer had dilation width doubled in the last. In our model, each CNN had a kernel size of 3, starting with dilation rate equals to 1. We doubled the dilation rate after each block until we reached a set number called Max Dilation.

Because CNNs in Sequence Recognition were typically one-dimensional, applied to a sequence of vector representing the vector, we used one-dimensional CNN as well in our network.

Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) [7] was a scoring function between the input, a sequence of observations, and the output, a sequence of labels, which could include "blank" characters. Typically, in speech recognition or sequence recognition problem, the alignment between the speech signal and the label was unknown. CTC provided an effective way to overcome this issue by summing over the probability of all possible alignments between the input and the output. That way, CTC was the best choice to cope our network with sequence recognition.

3. RESULTS

This section presented our experimental setups and results when evaluating DPRN in ASR and OCR. Both were conducted on the same Vietnamese broadcast news corpus collected from VTV programs. We also measured how well these results affected the performance of video retrieval. By default, all variables from our networks were initialized with a value sampled from a normal distribution with mean equals to 0 and standard deviation equals to 1.

OCR evaluation

The OCR dataset included 5602 images with corresponding captions. Each image had a fixed height of 40 pixels with various lengths. We split our dataset into 3 subsets: 4000 images for training, 500 files for validation, and 1102 for testing. In this experiment, we set the Max Dilation to 512. We increased the feature map by 24 in the first CNN. However, every CNN after the first increased the feature map by 30. There was only one group in this model, meaning no down-sampling module.

Table 1. DPRN performance on OCR

CAPTION DECODING		
Model	% CER	% WER
CRNN–CTC	0.78	2.87
DPRN–CTC	0.83	2.27

DPRN was trained using an Adam algorithm [13] with a configuration of 0.02 learning rate,

16 mini-batches, and 50 epochs. We compare our method with a variant of convolutional recurrent neural network (CRNN) by B. Shi [14]. We re-implemented their architectures and test them on our dataset. Character and word error rates were shown in Table 1. With a relative improvement of 20.91% in WER, our method had shown superior accuracy over the approach of CRNN-CTC.

ASR evaluation

The speech dataset, consisting of 39321 audio files (88 hours), was split into 34321/2500/2500 (77h/5.5h/5.5h) for the training set, validation set, and testing set respectively. The feature selection was made by a competition between FBank and MFCC through several configurations. Both candidates included delta and delta-par-delta features of their own.

Table 2. Feature selection for ASR

Feature Type	Number of Features	% CER	% WER
FBank	120	21.82	46.15
FBank	240	26.78	55.48
MFCC	72	15.01	33.78
MFCC	120	13.62	30.55
MFCC	240	13.2	29.09

The networks featured in Table 2 all used max dilation in 1024. The first CNN increases feature map by 16, and by 24 (except the first row, which was 20) for every CNN after that. There were two groups; feature map was increased by 1.5 times after the down-sampling module.

Table 3. DPRN performance on ASR

SPEECH DECODING		
Model	% CER	% WER
KC Engine	-	31.73
Deep Speech 2	16.01	34.02
DPRN–CTC	13.20	29.09

Networks were trained by an Adam algorithm with a configuration of 0.02 learning rate, 16 mini-batches, and 30 epochs. We tried various methods of feature extraction shown in Table 2. MFCC with 80 features attained the best result in our experiments. Hence, it was used in a comparative experiment with other model shown in Table 3. Specifically, we compared with Deep Speech 2 by Amodei [15] and the KCEngine [16] – a conventional HMM-GMM based Vietnamese speech recognizer. We also re-implemented their architectures and test them on our dataset. Table 3 listed the resulted character and word error rates. An absolute record of 29.09% WER surely turned the tide in favor of DPRN over the others. This was a very rewarding outcome considering how much tuning DPRN had gotten.

Video retrieval evaluation

As experimental improvement gave rise on the performance of ASR and OCR tasks, we proceeded to measure the retrieval performance using these indexes. Two hundred (200) name entities were randomly selected from the video database to serve as incoming queries. Attained recalls and precisions were listed in Table 4 in which we aligned DPRN with the duo CRNN-DeepSpeech2 and CRNN-KCEngine.

Table 4. Video retrieval evaluation

Indexing System	#RvE[a]	#RtE[b]	#CRtE[c]	%Recall	%Precision
CRNN + KC Engine	5826	3943	3507	67.68	88.94
CRNN + Deep Speech 2		3750	3296	64.37	87.89
DPRN		4231	3872	72.62	91.52

a. The number of relevant entities in the database; b. The number of retrieved entities; c. The number of correctly retrieved entities

Different entities might result in different recalls and precisions. However, as the number of entities increased, performances should converge to their average point. In this case, DPRN won over the combos of CRNN-DeepSpeech2 and CRNN-KCEngine since its base error rates in ASR/OCR were lower.

4. CONCLUSION

Experiments indeed confirm that our model, DPRN, was fully capable of handling sequence recognition tasks, specifically video speech and news-captions. This modification enabled a powerful model like PyramidNet to operate in an end-to-end sequence mapping manner. It dominated CRNN in an OCR task; however more tuning was needed for the ASR case, despite a win over Deep Speech 2 and KC Engine – those of the current state-of-the-art speech recognition techniques.

Acknowledgment: *This work is part of the VNU key project No. B2016-76-01A, supported by the Vietnam National University HCMC.*

REFERENCES

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1533–1545, 2014.

[2] S. Lai, L. Xu, K. Liu, J. Zhao, "Recurrent convolutional neural networks for text classification", in Proceeding of AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp 2267–2273, 2015.

[3] T. Wang, D.J. Wu, A. Coates, A.Y. Ng, "End-to-End text recognition with convolutional neural networks," International Conference on Pattern Recognition (ICPR), 2012.

[4] J. Deng, et al, "Imagenet: A large-scale hierarchical image database", Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[5] K. He, et al. "Deep residual learning for image recognition", Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[6] D. Han, J. Kim, J. Kim. "Deep pyramidal residual networks", arXiv preprint arXiv:1610.02915 (2016).

[7] A. Graves, et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks", Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[8] F. Yu, V. Koltun, "Multi-scale context aggregation by dilated convolutions", arXiv preprint arXiv:1511.07122 (2015).

[9] A. Amir, et al., "A multi-modal system for the retrieval of semantic video events", *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 216–236, Nov. 2004.

[10] L. Ballan, M. Bertini, A.D. Bimbo, G. Serra, "Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies", *Multimedia Tools and Applications*, vol. 48, no. 2, pp. 313–337, 2010.

[11] A. Fujii, K. Itou, , T. Ishikawa, "LODEM: a system for on-demand video lectures", *Speech Communication*, vol. 48, no. 5, pp. 516–531, May 2006.

[12] A.G. Hauptmann, M.G. Christel, R. Yan, "Video retrieval based on semantic concepts", *Proceedings of the IEEE*, vol. 96, pp. 602–622, 2008.

[13] D. Kingma, J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[14] B. Shi, X. Bai, C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, 2298–2304, 2017.

[15] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, and J. Chen, "Deep speech 2: End-to-end speech recognition in english and mandarin," in International Conference on Machine Learning, pp. 173–182, 2016.

[16] Q.Vu, et al., "A robust transcription system for soccer video database", Proceedings of ICALIP, Shanghai, 2010.

Truy vấn video đa thể thức sử dụng Dilated Pyramidal Residual Network

La Ngọc Thùy An, Nguyễn Phước Đạt, Phạm Minh Nhựt, Vũ Hải Quân

Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
Tác giả liên hệ: vhquan@vnuhcm.edu.vn

Ngày nhận bản thảo 10-07-2018; ngày chấp nhận đăng 10-09-2018; ngày đăng 20-11-2018

***Tóm tắt*—Các dạng mạng neuron đa lớp đã gặt hái được nhiều kết quả đáng ghi nhận trong lĩnh vực phân lớp ảnh, đặc biệt là mạng PRN (Pyramidal Residual Network). Tuy nhiên, ở thời điểm viết báo cáo này, chưa có một công trình chính thức nào áp dụng mạng PRN cho tác vụ phân lớp tín hiệu chuỗi. Chúng tôi đề xuất phương pháp mở rộng kiến trúc PRN, chuyển biến thành một dạng mạng mới với tên gọi DPRN (Dilated Pyramidal Residual Network), đồng thời tiến hành lượng giá hiệu năng của nó trong lĩnh vực nhận dạng tiếng nói và nhận dạng chữ in. Đây là hai tiền tố cần thiết phục vụ cho một ứng dụng trong ngữ cảnh lớn hơn: truy vấn video đa thể thức. Thực nghiệm được tiến hành trên kho ngữ liệu thu thập từ chương trình thời sự của kênh VTV đài truyền hình Việt Nam. Kết quả cho thấy DPRN không chỉ áp dụng được cho tác vụ nhận dạng chuỗi tín hiệu theo thời gian, mà còn cho kết quả vượt trội hơn các giải pháp truyền thống.**

***Từ khóa*—Dilated Pyramidal Residual Network, truy vấn video đa thể thức, nhận dạng tiếng nói tiếng Việt, nhận dạng chữ in**