Deep Convolutional and Recurrent Neural Network for Object Classification and Localization

Yeha Lee, Kyu-Hwan Jung, Hyun-Jun Kim and Sangki Kim
{yeha.lee, khwan.jung, dannis, sangki.kim}@vuno.co

Motivation

For accurate object classification and localization, considering spatial context and dependency is crucial. However, most popular deep learning architecture for vision task convolutional neural networks(CNN), in its basic form, only considers local information by the concept of 'receptive field'. Therefore, to learn global spatial dependencies among these 'receptive fields', we need to use some form of graphical model which are able to learn 2-dimensional structure of the representation of the image. Among various models, multi-dimensional recurrent neural network, specifically multi-dimensional long-short term memory(MD-LSTM) ([2]) has shown promising results and can be naturally integrated and trained 'end-to-end' fashion. However, when we try to learn the structure with very low level representation such as input pixel level, the dependency structure can be too noisy or spatially long-term dependency information can be vanished while training. Therefore, we propose to use 2D-LSTM layer on top of convolutional layers by taking advantage of convolution layers to extract high level representation of the image and 2D-LSTM layer to learn global spatial dependencies. We call this network as convolutional LSTM(CLSTM) and investigate the characteristics of this model with conventional CNN architecture.

Models

[Classification]

For classification, we trained 4 models, 2 CNNs and 2 CLSTMs. Our baseline CNN models are our own implementation of 'model C' of [1]. Our CLSTM models are constructed by replacing the last two convolution layer of CNN with two 2D LSTM layers. Since we used multidirectional 2D LSTM, there are 2^2 directional nodes for each location of feature map. For training, we used scale jittering by making the shorter side of image to be between 256 and 512 and randomly cropped and flipped 224x224 image.
For testing, we used multi-scale tests with shorter side size to be [256, 384, 512]. We also used sliding window approach to obtain spatial classification map and averaged all location and scale to produce final classification result. The learning rate of 2D LSTM layer is 1/10 of other layers such that it starts with 1e-3 and multiplied by 1/10 when error curve start to saturate.

CNN Model	CLSTM Model
(7x7) x 96	(7x7) x 96
(2x2) maxpool	(2x2) maxpool
(3x3) x 384 (3x3) x 384 (3x3) x 384 (3x3) x 384 (3x3) x 384 (3x3) x 384 (2x2) maxpool	(3x3) x 384 (3x3) x 384 (3x3) x 384 (3x3) x 384 (3x3) x 384 (3x3) x 384 (2x2) maxpool
(3x3) x 768 (3x3) x 768 (3x3) x 768 (3x3) x 768 (3x3) x 768 (3x3) x 768 (2x2) maxpool	(3x3) x 768 (3x3) x 768 (3x3) x 768 (3x3) x 768 (3x3) x 768 (3x3) x 768 (2x2) maxpool
(3x3) x 896 (3x3) x 896 (3x3) x 896 (3x3) x 896 (3x3) x 896 (3x3) x 896	(3x3) x 896 (3x3) x 896 (3x3) x 896 (3x3) x 896 **(2dLSTM x 4) x 224** **(2dLSTM x 4) x 224**
SPP(7,3,2,1)	SPP(7,3,2,1)
FC 4096 FC 4096 FC 1000	FC 4096 FC 4096 FC 1000

[Localization]

For localization, we used per-class-regression approach and replaced the 1000-dimensional output layer of CNN with 4000-dimensional output layer which produces bounding box coordinate for each class([3]). We also replaced the SPP layer of classification model with maxpooling layer. We trained the model with shorter side size between 254 and 384 and tested with size [256, 320]. We generated multiple bounding boxes by applying sliding window on the final feature map before fully-Connected layer and the boxes are merged by simply merging boxes with closest center point. The confidence of each box is the softmax output of ensemble of 2 CNNs and 2 CLSTMs and accumulated while boxes are merged.

Results

For ILSVRC 2015, we submitted two models; CNN-CNN and CNN-CLSTM. CNN-CNN is simple combination of two independently trained CNNs and CNN-CLSTM is combination of CNN and CLSTM network.

[Classification]

Below, we show the classification performance of ILSVRC 2015 validation dataset with possible combination of CNN and CLSTM. As we can see, CLSTM alone or combination of CLSTM networks shows poorer performance than CNNs. However, when we combine CNNs with CLSTMs, they outperform each homogeneous combinations. By visually observing the examples that CLSTM predicts correctly while CNN fails, we claim that CLSTM outperforms CNN when the object can be recognized by the context surrounding the object and CNN outperforms CLSTM when there is a dominant object. Since most of the images in ILSVRC is mixture of these two extremes, their combination is beneficial for classification. In ILSVRC 2015 CLS-LOC task, our result **ranked 5th in classification with test error 5.03%**

[Validation Error]

Model	Top-5 Error
Single CNN	5.71%
Single CLSTM	6.11%
2 CNNs	5.41%
2 CLSTMs	5.75%
CNN+CLSTM	5.29%
2CNN + 2CLSTM	**5.12%**

[ILSVRC 2015 Results]

Team	Top-5 Error
MSRA	3.57%
ReCeption(Google)	3.58%
Trimps-Soushen	4.58%
Qualcomm Research	4.87%
VUNO	**5.03%**

[Localization]

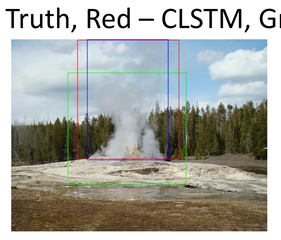
In localization task, we can again observe that combination of CNN with CLSTM outperforms their homogeneous combinations. When we visually investigated the examples that CLSTM finds the bounding box correctly while CNN fails, we could see the object to be localized is highly relevant with the surrounding context. Below, we show our performance with the aforementioned examples. In ILSVRC 2015 CLS-LOC task, our result **ranked 10th in localization with test error 25.30%.** Since we have not used any object proposal method and used relatively simple bounding box merging algorithm, we think we can improve the localization performance with more careful approaches.

[Validation Error]

Model	Top-5 Error
2 CNNs	25.77%
2 CLSTMs	25.80%
CNN+CLSTM	**25.34%**

[Sample Images]

Sample images from validation set the object which are correctly localized by CLSTM while fails by CNN. (Blue – Ground Truth, Red – CLSTM, Green – CNN)




[ILSVRC 2015 Results]

Team	Top-5 Error
MSRA	9.01%
Trimps-Soushen	12.29%
Qualcomm Research	12.55%
MCG-ICT-CAS	14.69%
Lunit-KAIST	14.73%
Tencent-Bestimage	15.55%
ReCeption(Google)	19.58%
CUimage	21.21%
CIL	23.19%
VUNO	**25.30%**

Conclusion

In this work, we proposed CLSTM network to be combined with state-of-the-art CNN to capture contextual information for classification and localization. Through the multi-dimensional recurrent structure, CLSTM can learn wide range spatial dependencies end-to-end manner.
By experiment and visual investigation, we could observed that CNNs and CLSTMs shows different strengths which supports our claim. Using our approach, we ranked 5th in classification and 10th in localization task
at ILSVRC 2015 CLS-LOC task.

References

[1] He, Kaiming, Xiangyu Zhang Shaoqing and Ren Jian Sun "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." arXiv preprint arXiv:1502.01852 (2015).
[2] Alex Graves, Santiago Fernandez and Jurgen Schmidhuber. Multidimensional recurrent neural networks. In ´ Proceedings of the 2007 International Conference on Artificial Neural Networks, Porto, Portugal, September (2007).
[3] Simonyan Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).