STRUCTURAL SEGMENTATION OF MUSICAL AUDIO WITH REGION PROPOSAL NETWORKS

Christopher Uzokwe
Drexel University
cnu25@drexel.edu

Dr. Youngmoo Kim
Drexel University
ykim@drexel.edu

ABSTRACT

Performance in many Music IR tasks has advanced significantly using deep learning methods, particularly convolutional neural networks (CNNs). The fundamental research behind CNNs has primarily been driven by visual domain problems, such as image recognition and object segmentation, and "standard" CNN architectures are optimized for such visual problems. Our work seeks to leverage these fundamental visual strengths of CNNs by transforming musical structure analysis (MSA) and segmentation into a purely visual task. We use labeled images of self-similarity matrices (SSMs) derived from acoustic features (from popular music examples) as a visual dataset to train a Region Proposal Network (RPN), a state-of-the-art object detection approach, to identify the regions of a song based on visual bounding boxes. This abstract highlights our modifications of the RPN implementation for our SSM dataset, and reports on the fundamental differences between the two tasks that serve as the biggest shortcomings of the approach in its current state.

1. REGION PROPOSAL NETWORKS

To realize our own object detection network (musical sections being the object), we use the Region Proposal Network first introduced in [1]. The system achieved state of the art performance on many image recognition datasets, and its main contribution was the RPN – a set of convolutional layers that could be trained end-to-end to propose object regions. The RPN has since been used as a backbone in varying object detection systems. The RPN first starts with a deep convolutional feature extraction, followed by a sliding kernel search.

The motivation to use Region Proposal Networks stems from their success in traditional object detection systems. We would like to leverage the network in bounding the compelling visual features present in SSMs. Like other MSA systems, RPNs also make use of a standard pipeline of transformations – including sliding kernel search [2], deep feature extractions [3], and layered SSM inference [4]. It is important to note that the RPN wasn't built to layer images, but we can make use of how images themselves are constructed, by mapping different features to the RGB channels of the image.

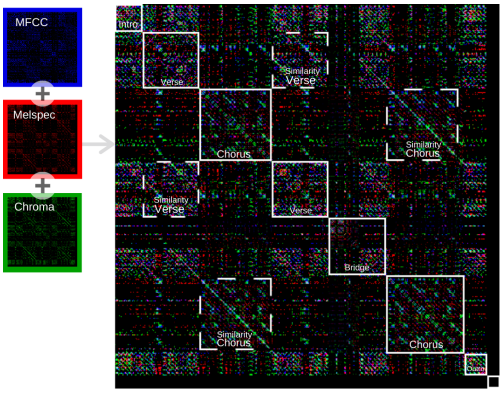

Figure 1. SSMs stacked to create a musical 'image.' Highlighted sections manifest on the diagonal of the matrix, with repeated structures on the off diagonal which denote similarities.

2. CURRENT PROGRESS

Following a previous off-the-shelf implementation done with Facebook AI Research's Detectron2 package [5] (which hosts a full module using the RPN as an boundary detection attention mechanism and a Region-based CNN for object classification), our current work has built from scratch a modified version of the Region Proposal Network, allowing for more precise control of the RPNs main features, the anchor targets to regress on, as well as the proposal layers.

The main modifications to the region proposal network include restricting the proposal layer to make proposals only along the diagonal of the SSM, as well as using only square anchor targets for regression. This is motivated by the way musical sections manifest in the self similarity matrix – strictly along the diagonal of the SSM in square regions.

To assess the performance of the newly created region proposal network, we compare the performance to the base performance of the off-the-shelf Detectron2 module. Using a selection of 550 songs from the SALAMI dataset (popular selections of music) as a training set and 150 songs from the test set, we train for 5 epochs across the

Model	Precision	Recall	F1
Detectron2	0.28	0.19	0.23
Modified RPN	0.32	0.36	0.32

Table 1. Precision, Recall, and F1 boundary score for the Detectron2 and Modified RPN predictions.

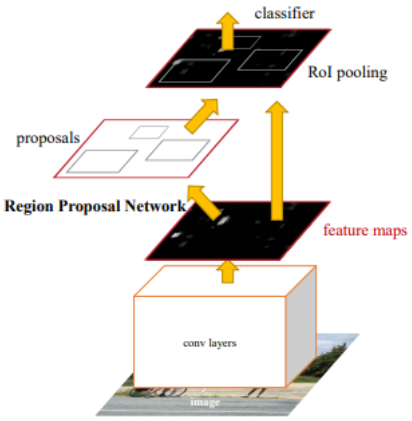

Figure 2. Faster R-CNN architecture figure used in [1]. Images are run through an extractor which the RPN predicts potential regions from.

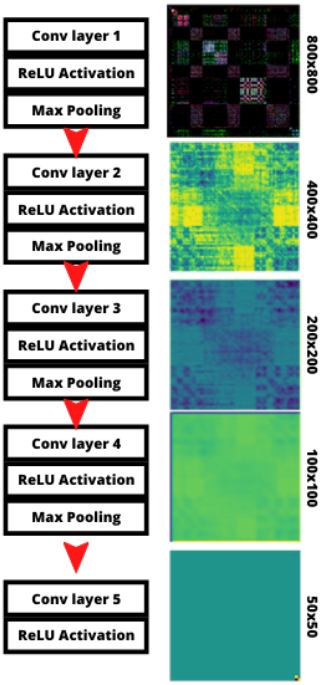

Figure 3. SSM output through the layers of the VGG-16 network. All significant features are lost as the image propagates through the network.

dataset. Ground truth bounding boxes are formed after scaling boundary times according to their frequency scaling, and predictions are converted back the same way. We use the mir_eval [6] package to asses boundary predictions at a 3s threshold. By switching from the detectron2 model to the modified RPN, we saw an increase in F1 score going from 0.23 (Detectron2) to 0.32 (modified RPN).

3. FUTURE CONSIDERATIONS

In the feature extraction portion of the network, we use the VGG-16 model [7]. This very deep convolutional network consists of several sequential 2D convolutional layers, max pooling layers, and ReLU activations. It calls for each image to be scaled to a standard size of 800x800 before entering the network, which is progressively compressed via convolutional layers to 50x50 at the deepest layer.

During this process, the VGG-16 suffers from the vanishing gradient problem. As our image propagates through the network, all of its features go to 0, and no significant information is retained. The network is then unable to make predictions based on the features, and tends to learn one optimal set of predictions across the whole dataset. By viewing the image transformations throughout the network we can see that the 50x50 map goes to 0, and features come back as we propagate upwards through the network in the 100x100, 200x200, and 400x400 feature maps.

Another problem with the compressed features is that it can become difficult for the network to accurately locate a boundary, given that the features are scaled to 800x800 and then scaled even further during the feature extraction. Initial tests have shown that the network performs with increased precision when operating on larger feature maps – an issue that might not be relevant in object detection, but shows more importance in this time-dependent variant.

Compensating for these issues may come at the expense of a larger computational cost and greater training time, but there are other networks that exist such as ResNet [8] – a residual learning network that mitigates the vanishing gradient problem with skip connections while maintaining deep layer transformations.

Because we use anchor targets as ground truth input to the network, the associated IoU thresholds of the positively and negatively labeled targets play a significant role. The IoU thresholds are based on the percentage overlap from the anchor target to the actual section boundary. The original RPN implementation is able to use relatively low positive thresholds for the anchor targets (since bounding an object does not require a perfect outline), but our case requires a tighter scope when we consider the precision necessary in making boundary predictions in time.

Since the IoU thresholds directly affect the number of positive and negative samples we use for training, we also must consider a sampling ratio that allows for an even distribution of samples. Further experiments will work to highlight and describe how such parameter modifications effect the system as a whole, and with this information we hope to optimize the system to produce more accurate predictions.

4. REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region

proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[2] K. Ullrich, J. Schlüter, and T. Grill, "Boundary detection in music structure analysis using convolutional neural networks." in *ISMIR*, 2014, pp. 417–422.

[3] M. C. McCallum, "Unsupervised learning of deep features for music segmentation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 346–350.

[4] T. Grill and J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations." in *ISMIR*, 2015, pp. 531–537.

[5] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[6] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.