Speeding up VP9 Intra Encoder with Hierarchical Deep Learning Based Partition Prediction

Somdyuti Paul, Andrey Norkin and Alan C. Bovik
Outline

1. Introduction
2. Related Work
3. Overview of Approach
4. Database Creation
5. H-FCN Model Architecture
6. H-FCN Training
7. Prediction Performance
8. Inconsistency Correction
9. Visualizing Superblock Partitions
10. Encoding Performance
11. Concluding Remarks
12. References
Introduction

- In VP9, $64 \times 64$ superblocks are partitioned recursively, possibly down to $4 \times 4$ blocks at four hierarchical levels.
- The rate-distortion optimization (RDO) based partition decision is a slow process owing to the combinatorial complexity of the partition search space.

![Hierarchical superblock partition at four levels.](image)

Figure 1: Hierarchical superblock partition at four levels.
Several machine learning (ML) based approaches with custom feature design attempted to reduce the computational overhead of the partition search in HEVC [1], VP9 [2] and VVC [3].

Fewer works use deep learning based methods to solve the problem for HEVC [4, 5, 6].

A parallel convolutional neural network architecture was employed in [4] to achieve a speedup of 61.8% for a 2.25% increase in BD-rate in the intra mode of HEVC.

A multi stage ML-framework was used to sequentially make block partition decisions in [2], achieving a speedup of 60.1% over the speed 0 setting of the VP9 encoder with 0.07% increase in BD-rate.
Overview of Approach

Our approach involves a bottom-up block merge prediction using a hierarchical fully convolutional neural network (H-FCN) [7].

Figure 2: VP9 partition prediction approach.

implementation available at https://github.com/Somdyuti2/H-FCN.git
The content for our database comprises 89 movies and 17 television episodes, which were selected from video sources in the Netflix catalog.

Each video content was encoded at three different resolutions (1080p, 720p and 540p) using the reference VP9 encoder from the *libvpx* package.

The contents were encoded in VP9 Profile 0, using speed level 1 and the *good* quality configuration.
A concise description of the partition tree was required for effective learning.

The partition tree was represented in the form of a set of four matrices:

![Matrix representation of the four level partition tree.](image)

Figure 3: Matrix representation of the four level partition tree.
The reference VP9 decoder from the *libvpx* package was modified to extract the superblock partition trees and the corresponding quantization parameter (QP) values from the encoded bitstreams.

The raw pixel data for each superblock was obtained by extracting the luma channels of non-overlapping $64 \times 64$ blocks from the source videos downsampled to the encode resolution.

Our database encompasses internal QP values in the range 8-105.

**Table 1: Summary of VP9 intra-mode superblock partition database**

<table>
<thead>
<tr>
<th>Database</th>
<th>Contents</th>
<th>% of CGI content</th>
<th># of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>62 (M) + 12 (E)</td>
<td>12.16</td>
<td>11 990 384</td>
</tr>
<tr>
<td>Validation</td>
<td>27 (M) + 5 (E)</td>
<td>12.50</td>
<td>4 698 195</td>
</tr>
</tbody>
</table>
Figure 4: Architecture of H-FCN model having 26,336 parameters and 54,610 FLOPs.
Categorical cross entropy loss

\[ L_q(w) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{i,j} \log(p_{i,j}^q(w)) \quad q = 1, \cdots, 85 \quad (N = 128, K = 4) \]

**Figure 5:** *H-FCN loss with training progress.*
The prediction accuracy at each level was evaluated on $10^5$ randomly drawn samples from the training and validation sets.

Table 2: Prediction accuracy of H-FCN model

<table>
<thead>
<tr>
<th>Level #</th>
<th>Training (%)</th>
<th>Validation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>89.42</td>
<td>90.27</td>
</tr>
<tr>
<td>1</td>
<td>84.42</td>
<td>83.47</td>
</tr>
<tr>
<td>2</td>
<td>86.07</td>
<td>85.13</td>
</tr>
<tr>
<td>3</td>
<td>91.73</td>
<td>91.18</td>
</tr>
</tbody>
</table>
Inconsistency Correction

- At each level, the model predictions are made independently of all other levels.
- Possible inconsistencies between the predictions of any two levels are corrected by a top-down approach.

Figure 6: Top-down inconsistency correction.
Visualizing Superblock Partitions

Figure 7: Superblock partitions predicted by the trained H-FCN model compared with ground truth.
The trained model was integrated with the reference VP9 encoder using the Tensorflow C API.

The predicted partitions were ordered to form a preorder traversal of the partition tree, and subsequently used to replace the RDO based partition decision in a recursive fashion.

The encoding performance was evaluated on 30 test sequences at 3 resolutions in terms of both BD-rate and speedup ($\Delta T$).

Table 3: Encoding performance with respect to RDO baseline

<table>
<thead>
<tr>
<th>Resolution</th>
<th>$\Delta T$ (%)</th>
<th>BD-rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1080p</td>
<td>67.5</td>
<td>1.70</td>
</tr>
<tr>
<td>720p</td>
<td>72.2</td>
<td>1.75</td>
</tr>
<tr>
<td>540p</td>
<td>69.5</td>
<td>1.68</td>
</tr>
<tr>
<td>Overall</td>
<td>69.7</td>
<td>1.71</td>
</tr>
</tbody>
</table>
The speedup and BD-rate of our approach was also compared with speed level 4 of the reference VP9 encoder, the highest recommended speed level for the baseline configuration.

Table 4: Comparison of speedup versus BD-rate tradeoff of our approach with VP9 speed level 4

<table>
<thead>
<tr>
<th>Resolution</th>
<th>ΔT (%)</th>
<th>BD-rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speed 4</td>
<td>H-FCN</td>
</tr>
<tr>
<td>1080p</td>
<td>62.0</td>
<td>67.5</td>
</tr>
<tr>
<td>720p</td>
<td>68.2</td>
<td>72.2</td>
</tr>
<tr>
<td>540p</td>
<td>65.9</td>
<td>69.5</td>
</tr>
<tr>
<td>Overall</td>
<td>65.4</td>
<td>69.7</td>
</tr>
</tbody>
</table>
The benefit offered by our approach in terms of speedup persists across the range of QP values used to learn the H-FCN model.

Figure 8: *Speedup achieved by H-FCN and RDO at speed 4 relative to baseline.*
Our H-FCN based partition prediction approach achieved 69.7% speedup on average at the expense of 1.71% increase in BD-rate.

It achieves 4.3% higher speed up than the speed level 4 of the reference encoder, while incurring 1.44% smaller BD-rate penalty.

Further benefits can possibly be derived by extending the approach to the AV1 codec.
References


References

