Clownfish: Edge and Cloud Symbiosis for Video Stream Analytics

Vinod Nigade, Lin Wang, Henri Bal
Vrije Universiteit Amsterdam
Motivation: Video Stream Analytics

- Applications such as **augmented reality**, **public safety** at airport need **accurate analytics in real time**

- **Higher accuracy** due to advanced (DNN-based) computer vision algorithms

- **Increased computational complexity** of DNNs hurts real-time objective
Motivation: Video Stream Analytics

- Applications such as **augmented reality**, **public safety** at airport need **accurate** analytics in **real time**

- **Higher accuracy** due to advanced (DNN-based) computer vision algorithms

- **Increased computational complexity** of DNNs hurts real-time objective

(a) Frame-based inference

(b) Window-based inference
Motivation: Design Choices

Edge-only

- Faster response time
- Resource limitations; Smaller models, often lower accuracy

Cloud-only

- Higher accuracy
- Streaming over WAN; Highly variable and long response time

[Chameleon, SIGCOMM’18], [AWStream, SIGCOMM’18], [Nexus, SOSP’19]

[ParkMaster, SEC’17], [Efficient-3DCNNs, CVPR’19], [Skynet, MLSys’20]
Motivation: Design Choices

• How to benefit from both worlds?

Edge Hybrid Cloud

• Fast response time
• High accuracy

[Glimpse, SenSys’15], [Neurosurgeon, ASPLOS’17], [FilterForward, SysML’19]

5
Motivation: Leverage Temporal Correlations

- Video has significant temporal correlation across frames
e.g., an action may span across several frames

- Common frames across overlapping windows in window-based inference
Motivation: Leverage Temporal Correlations

- Video has significant temporal correlation across frames, e.g., an action may span across several frames.
- Common frames across overlapping windows in window-based inference are measured using Cosine similarity.
Clownfish: Architecture

Goal:
• Achieve symbiosis between edge and cloud for real-time video stream analytics

Challenges:
• How to fuse the cloud analytics results with the edge results?
• Which frames to send to the cloud?
Clownfish: Components

**Edge node:**
- **Window Manager**
  Generates frame windows
- **Local**
  Runs optimized (or smaller) DNN model
- **Filter**
  Filters out windows to be sent to cloud
- **Fusion**
  Fuses analytics results from cloud with that of edge

**Cloud node:**
- **Remote**
  Runs complete (or bigger) DNN model

Windows generated by **Window Manager**
Clownfish: Fusion Method

- A lightweight method that runs on the edge node
- Exponential Smoothing (ES) approach to fuse past result and current local result
- $\alpha_t \in [0, 1]$ is a weight (correlation) parameter in ES for previous fused result and current local result
- Two main procedures,
  - **FUSE**: Used for real-time results fusion
  - **REINFORCE**: Updates state when remote result becomes available

**FUSE**

$$\tilde{p}_f(t) = \begin{cases} p(t), & \text{if } t = 1, \\ \alpha_t \tilde{p}_f(t-1) + (1 - \alpha_t) \tilde{p}(t), & \text{otherwise,} \end{cases}$$

Where,
- $\tilde{p}_f(t-1)$, Fused result for the past window $w_{t-1}$
- $\tilde{p}(t)$, Local result for window $w_t$

**REINFORCE**

Update $\tilde{p}_f(t - N), ..., \tilde{p}_f(t - 1)$,

$$\tilde{p}_f(i) = \begin{cases} g(\tilde{p}_f(i), \tilde{p}_r(i)), & \text{if } i = t - N, \\ \alpha_i \tilde{p}_f(i-1) + (1 - \alpha_i) \tilde{p}(i), & \text{otherwise.} \end{cases}$$

Where,
- $\tilde{p}_r(t)$, Remote result for window $w_{t-N}$
- $i \in [t - N, t - 1]$
Fusion Method: Estimating Temporal Correlations

How to set value $\alpha_t$ ?

- **Estimate correlation parameter** using similarity score between two consecutive windows.

- **When score is high**, windows have similar context:
  - Assign relatively larger value for $\alpha_t$, i.e., larger weight to the previous fused result.

- **Traditional similarity functions** based on vector distance such as Cosine, Euclidean may give:
  - Low correlation for the same context
  - High correlation for different contexts at context transition.
Fusion Method: Estimating Temporal Correlations

How to set value $\alpha_t$?

- Estimate correlation parameter using similarity score between two consecutive windows.

- When score is high, windows have similar context.
  - Assign relatively larger value for $\alpha_t$, i.e., larger weight to the previous fused result.

- Traditional similarity functions based on vector distance such as Cosine, Euclidean may give:
  - Low correlation for the same context.
  - High correlation for different contexts at context transition.

- Context similarity function using learning-based approach to capture (dis)similarity of contexts.
Clownfish: Filter

When to send windows to remote cloud?

Two context-aware policy,
• Send a window at the start of context.
  • Leverage similarity score to identify context transition, i.e., $\rho_t - \rho_{t-1} \geq 0.5$

• Periodically send windows within same context and restart periodic timer at context transition
Evaluation

How effective is our SimiNet-based fusion method?

Setup:
• Local model: 3D Resnet-18
• Remote model: 3D Resnext-101
• Dataset: PKU-MMD
• Task: Action Recognition
Evaluation

Setup:
• Local model: 3D Resnet-18
• Remote model: 3D Resnext-101
• Dataset: PKU-MMD
• Task: Action Recognition

How effective is our SimiNet-based fusion method?

• Our SimiNet-based fusion method performs close to remote and accuracy gap is within 2%
• Substantial bandwidth reduction with limited penalty on accuracy
How does network latency affect accuracy of Clownfish?

- Network latency has a **negligible impact** on the achieved accuracy of Clownfish.
Evaluation

How do bandwidth conditions affect Clownfish?
Evaluation

How do bandwidth conditions affect Clownfish?

Accuracy

- Accuracy is comparable to that of cloud-only solution
Evaluation

How do bandwidth conditions affect Clownfish?

**Accuracy**

- Accuracy is **comparable** to that of cloud-only solution
- Maintains **stable throughput** (FPS) similar to the edge-only solution

<table>
<thead>
<tr>
<th>Bandwidth (5 Mbps)</th>
<th>Bandwidth (7.5 Mbps)</th>
<th>Bandwidth (No shaping)</th>
</tr>
</thead>
<tbody>
<tr>
<td>56.69</td>
<td>57.09</td>
<td>58.69</td>
</tr>
<tr>
<td>58.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge-only</td>
</tr>
<tr>
<td>Clownfish</td>
</tr>
<tr>
<td>Cloud-only</td>
</tr>
</tbody>
</table>

(A: no shaping, B: 7.5Mbps, C: 5Mbps)
Evaluation

How does Clownfish perform when compared to filtering-based approach, e.g., EarlyDiscard\(^1\)?

---

Evaluation

How does Clownfish perform when compared to filtering-based approach, e.g., EarlyDiscard\(^1\)?

<table>
<thead>
<tr>
<th>Solution</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge-only</td>
<td>51.16%</td>
</tr>
<tr>
<td>EarlyDiscard</td>
<td>55.48%</td>
</tr>
<tr>
<td>Clownfish (7.5Mbps)</td>
<td>57.09%</td>
</tr>
<tr>
<td>Cloud-only</td>
<td>58.98%</td>
</tr>
</tbody>
</table>

Evaluation

How does Clownfish perform when compared to filtering-based approach, e.g., EarlyDiscard\(^1\)?

<table>
<thead>
<tr>
<th>Solution</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge-only</td>
<td>51.16%</td>
</tr>
<tr>
<td>EarlyDiscard</td>
<td>55.48%</td>
</tr>
<tr>
<td>Clownfish (7.5Mbps)</td>
<td>57.09%</td>
</tr>
<tr>
<td>Cloud-only</td>
<td>58.98%</td>
</tr>
</tbody>
</table>

- Clownfish outperforms EarlyDiscard in terms of accuracy and throughput

Summary

• **Clownfish**, a hybrid framework for real-time video stream analytics that takes the benefits of edge and cloud

• Clownfish fusion method based on exponential smoothing exploits temporal correlation categorized using learning-based similarity model

• Clownfish always operates in real time like an edge-only solution and achieves high accuracy comparable to a cloud-only solution
Summary

• **Clownfish**, a hybrid framework for real-time video stream analytics that takes the benefits of edge and cloud

• Clownfish fusion method based on exponential smoothing exploits temporal correlation categorized using learning-based similarity model

• Clownfish always operates in real time like an edge-only solution and achieves high accuracy comparable to a cloud-only solution

For more details,
• Source code: https://github.com/vuhpdc/clownfish
• Contact: v.v.nigade@vu.nl

Thank You!