Semantic Event Detection with Optimized Vision Language Model.

Challa Santhsoh 11/02/2026

1. Chosen Model

Model: OpenAI CLIP (ViT-B/32)
Checkpoint: openai/clip-vit-base-patch32

CLIP (Contrastive Language–Image Pre-training) is a vision-language model that maps images and text into a shared feature space.

Why CLIP?

- Supports zero-shot detection (no task-specific training required)
- Can detect semantic concepts like:
    - Person walking
    - Vehicle stopping
    - Crowded scene
- Works well on CPU
- Compatible with dynamic quantization

The ViT-B/32 version provides a good balance between accuracy and computational efficiency, making it suitable for real-time applications.

---

2. Optimization Technique

Method Used: INT8 Dynamic Quantization
Framework: PyTorch
Function: torch.quantization.quantize_dynamic

What is Dynamic Quantization?

Dynamic quantization converts model weights from:

- 32-bit floating point (FP32) to 8-bit integers (INT8)

This reduces model size and improves inference speed.

Key Features

- No retraining required
- No labeled dataset needed
- Applied directly to pre-trained model
- Works efficiently for CPU inference

Only model weights are quantized, while activations are dynamically quantized during runtime.

---

3. Performance Results

Performance was measured using the test video (82 frames).

Metric	Baseline (FP32)	Optimized (INT8)	Improvement
Inference Time (sec/frame)	0.0579	0.0410	29.2% faster
FPS	17.28	24.40	+41.2%
Model Size (MB)	577	106	81.6% reduction

Key Improvements

- 1.41× faster inference
- 81.6% reduction in model size
- 29.2% lower latency per frame
- Stable performance across all frames

The optimized model achieves real-time performance at 24.4 FPS on CPU.

---

4. Trade-offs

Advantages

- Much smaller model size
- Faster CPU inference
- No retraining required
- Easy and quick deployment
- Minimal accuracy loss

Limitations

- Quantization is irreversible (original FP32 model must be saved)
  Small numerical precision loss
- Slight possibility of minor accuracy variation in rare edge cases
- GPU speed improvement may differ from CPU results

5. Conclusion

INT8 dynamic quantization is an effective and practical optimization technique for deploying CLIP ViT-B/32 on CPU-based systems.

The optimization achieved:

- 81.6% model size reduction
- 1.41× speed improvement
- Real-time inference capability
- Minimal accuracy degradation

The optimized model is suitable for real-time semantic event detection in resource-constrained environments.