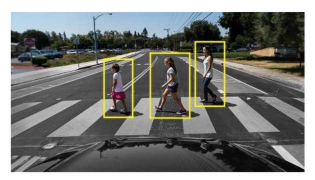
PedeScan: A Plug-n-Play Quantized Quality-Aware Modality Balanced Pedestrian Detector

Kapu Nirmal Joshua[1], Vijay Pandey[1], Tushar Sandhan[1]

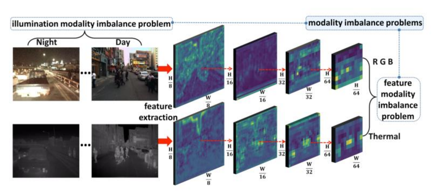
[1] Department of Electrical Engineering
Indian Institute of Technology Kanpur

INTRODUCTION AND MOTIVATION

Pedestrian Detection is an task critical to many applications such as autonomous driving and navigation!



RGB Images works well mainly in the daytime, and fails during the night. This is where thermal imaging comes into importance. However we must now balance both modalities for accurate detection, as they give us same information.

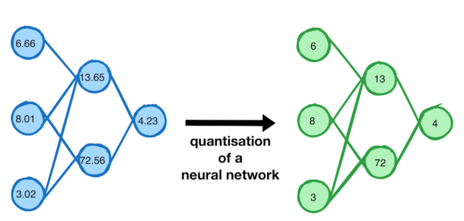


The MBNet Architecture[1] proposed by *Kailai Zhou, Linsen Chen, Xun Cao* addresses exactly this issue. It uses a two module approach: the DMAF *(Differential Modality Aware Fusion)* Module and the IAFA (Illumination Aware Feature Alignment) Module to address this exact issue.

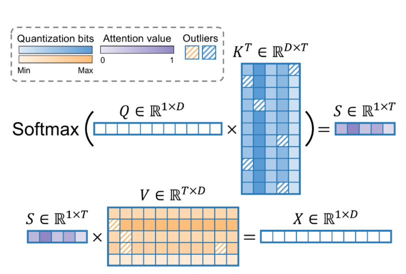
D_{RGB} D_T (RGB + Thermal Datasets) → F_T F_R (RGB Thermal Feature Maps) → DMAF Module (aligns the feature maps) → F'_T F'_R (Enhanced Feature Maps) → IAFA Module → (Pedestrians Detected)

The *MBNet* Pipeline (shown above) takes a significant amount of time during inference (~10 minutes for 1000 pictures). This can be dangerous for time-sensitive applications such as autonomous driving.

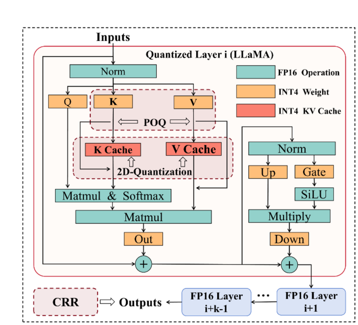
PEDESCAN SYSTEM DETAILS: A QUALITY AWARE QUANTIZATION BASED APPROACH



Inspired by works such KVQuant[2] and QAQ[3], we extend the concepts of quality aware quantization our MBNet Framework to reduce latency while maintaining comparable miss rates.



We extend the key concepts of quality aware KV Cache quantization – including non uniform channel quantization, and the statistical allocation of non uniform number of bits for each channel.



We use a three-stage pipeline for our system as shown below:

Multimodal QAQ Stage

Let $\mathbf{F}_R \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{F}_T \in \mathbb{R}^{H\times W\times C}$ represent RGB and Thermal feature maps. Quantize using functions $Q_R(\cdot)$ and $Q_T(\cdot)$:

$$\hat{\mathbf{F}}_R = Q_R(\mathbf{F}_R), \quad \hat{\mathbf{F}}_T = Q_T(\mathbf{F}_T). \tag{1}$$

Quantization noise:

$$\boldsymbol{\epsilon}_m = \hat{\mathbf{F}}_m - \mathbf{F}_m, \quad m \in \{R, T\}. \tag{2}$$

Noise variance:

$$\sigma^2_{\epsilon_m} = \frac{\Delta_m^2}{12}, \tag{3}$$

Constraint:

$$\sigma^2_{\epsilon_m} \le \left(\sigma^{(m)}_{\max}\right)^2. \tag{4}$$

Number of levels:

$$L_m = 2^{b_m}. \tag{5}$$

Step size:

$$\Delta_m = \frac{F^{(m)}_{\max} - F^{(m)}_{\min}}{L_m - 1}. \tag{6}$$

Minimum bits:

$$b_m \ge \log_2\left(\frac{F^{(m)}_{\max} - F^{(m)}_{\min}}{\sigma^{(m)}_{\max}\sqrt{12}} + 1\right). \tag{7}$$

Use attention maps to allocate bits, ensuring important regions retain higher precision.

Let $\mathbf{A} \in \mathbb{R}^{H\times W}$ be the attention map. Allocate bits as:

$$b_m(i,j) = b^{(m)}_{\min} + \left(b^{(m)}_{\max} - b^{(m)}_{\min}\right)\cdot \mathbf{A}(i,j). \tag{8}$$

Non Uniform Bit Allocation Stage

Assign bits based on channel variance to preserve important features.

Quantization noise:

$$\sigma^2_{\epsilon^{(m)}_c} = \frac{\left(F^{(c,m)}_{\max} - F^{(c,m)}_{\min}\right)^2}{12\cdot 2^{2b^{(m)}_c}}. \tag{9}$$

Constraint:

$$\sigma^2_{\epsilon^{(m)}_c} \le \left(\sigma^{(c,m)}_{\max}\right)^2. \tag{10}$$

Minimum bits:

$$b^{(m)}_c \ge \frac{1}{2}\log_2\left(\frac{F^{(c,m)}_{\max} - F^{(c,m)}_{\min}}{\sigma^{(c,m)}_{\max}\sqrt{12}}\right). \tag{11}$$

Identify and handle outliers to prevent distortion in quantization.

Kurtosis:

$$\kappa^{(m)}_c = \frac{\mathbb{E}\left[\left(F^{(m)}_c - \mu^{(m)}_c\right)^4\right]}{\left(\sigma^{(m)}_c\right)^4}. \tag{12}$$

Optimize bit allocation to minimize total bits while controlling quantization noise.

$$\min_{b_m, b^{(m)}_c(i,j)} \sum_m \sum_c \sum_{i,j} b^{(m)}_c(i,j) \tag{13}$$

$$\text{subject to} \quad \sigma^2_{\epsilon^{(m)}_c}(i,j) \le \left(\sigma^{(c,m)}_{\max}\right)^2, \quad \forall m, c, i, j. \tag{14}$$

Integration Stage

Combine quantized features with DMAF and IAFA for enhanced fusion and alignment.

Compute differential feature:

$$\hat{\mathbf{F}}_D = \hat{\mathbf{F}}_R - \hat{\mathbf{F}}_T. \tag{15}$$

Fusion weights:

$$\mathbf{V}_w = \tanh\left(\mathrm{GAP}\left(\hat{\mathbf{F}}_D\right)\right). \tag{16}$$

Recalibrate features:

$$\hat{\mathbf{F}}'_R = \hat{\mathbf{F}}_R + \mathbf{V}_w \odot \hat{\mathbf{F}}_T, \tag{17}$$

$$\hat{\mathbf{F}}'_T = \hat{\mathbf{F}}_T + \mathbf{V}_w \odot \hat{\mathbf{F}}_R. \tag{18}$$

Adjust based on illumination:

$$w_R = \frac{I+\epsilon}{1+2\epsilon}, \quad w_T = \frac{1-I+\epsilon}{1+2\epsilon}, \tag{19}$$

Fused feature:

$$\mathbf{F}_{\text{fused}} = w_R\hat{\mathbf{F}}'_R + w_T\hat{\mathbf{F}}'_T. \tag{20}$$

Train the network to handle quantization effects using fake quantization during training.

$$\mathbf{F}_{\text{QAT}} = \mathrm{Quantize}(\mathbf{F}) + (\mathbf{F} - \mathrm{Quantize}(\mathbf{F}))\cdot \mathrm{StopGradient}. \tag{21}$$

Evaluate memory and computational savings from the quantization scheme.

$$\text{Compression Ratio} = \frac{\sum_m HWCb_{\text{float}}}{\sum_m \sum_c \sum_{i,j} b^{(m)}_c(i,j)}, \tag{22}$$

Multimodal QAQ Stage | Non Uniform Bit Allocation Stage | Integration Stage

PEDESCAN SYSTEM FLOWCHART

D_{RGB} D_T (RGB + Thermal Datasets) → F_T F_R (RGB Thermal Feature Maps) →

Constrained Quantization Module → Step Size Calculation Module → Non Uniform Bit Allocator → Attention Guided Quantizer → Channel Wise Bit Allocator

(Quality Aware Quantization Module)

$F\hat{}_T = Q_T(F_T)$

$F\hat{}_R = Q_R(F_R)$

$\varepsilon = F\hat{} - F$

Constraint: $\sigma^2_\varepsilon \le \sigma^2_{\max}$

$$b_m \ge \log_2\left(\frac{F^{(m)}_{\max} - F^{(m)}_{\min}}{\sigma^{(m)}_{\max}\sqrt{12}} + 1\right).$$

$$b^{(m)}_c \ge \frac{1}{2}\log_2\left(\frac{F^{(c,m)}_{\max} - F^{(c,m)}_{\min}}{\sigma^{(c,m)}_{\max}\sqrt{12}}\right).$$

Channel Wise Bit Allocator → Outlier Suppression (Kurtosis) → Optimising Number of Bits Allocated → Integration with DMAF and IAFA Modules → Quantizing Aware Training → (Pedestrians Detected) → Complexity Analysis

$$\min_{b_m, b^{(m)}_c(i,j)} \sum_m \sum_c \sum_{i,j} b^{(m)}_c(i,j) \tag{13}$$

$$\text{subject to} \quad \sigma^2_{\epsilon^{(m)}_c}(i,j) \le \left(\sigma^{(c,m)}_{\max}\right)^2, \quad \forall m, c, i, j. \tag{14}$$

$$\mathbf{F}_{\text{QAT}} = \mathrm{Quantize}(\mathbf{F}) + (\mathbf{F} - \mathrm{Quantize}(\mathbf{F}))\cdot \mathrm{StopGradient}. \tag{21}$$

Adjust based on illumination:

$$w_R = \frac{I+\epsilon}{1+2\epsilon}, \quad w_T = \frac{1-I+\epsilon}{1+2\epsilon}, \tag{19}$$

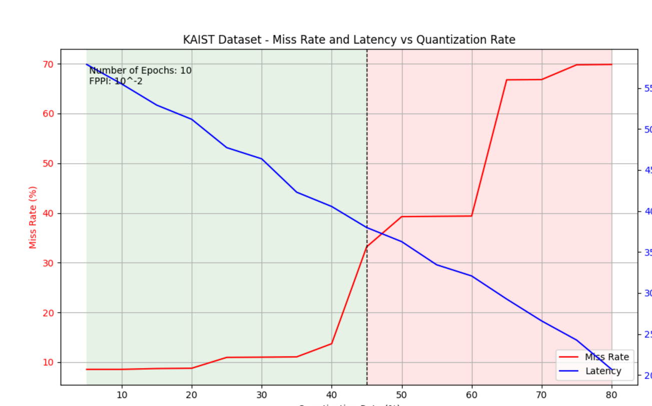
Fused feature:

$$\mathbf{F}_{\text{fused}} = w_R\hat{\mathbf{F}}'_R + w_T\hat{\mathbf{F}}'_T. \tag{20}$$

$$\text{Compression Ratio} = \frac{\sum_m HWCb_{\text{float}}}{\sum_m \sum_c \sum_{i,j} b^{(m)}_c(i,j)}, \tag{22}$$

(used for calculating the quantization rate – an important hyper parameter)

RESULTS ON *CVC14* AND *KAIST* DATASETS



(Results on the KAIST Dataset)

Table 2: *KAIST* Dataset

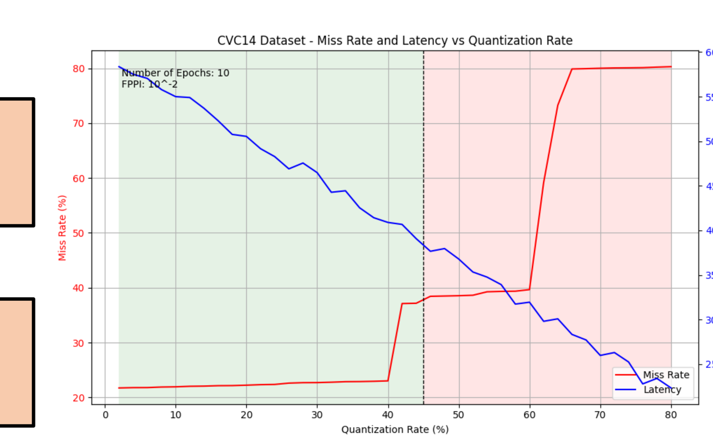
Quantization Noise (%)	Miss Rate (%)	Latency (ms)
5.0	21.85	258.59
10.0	21.96	222.67
15.0	22.08	222.29
20.0	22.17	222.56
25.0	22.43	221.62
30.0	22.74	221.26
35.0	22.95	222.28
40.0	23.05	222.53
45.0	**37.24***	222.60
50.0	38.57	222.17
55.0	39.33	222.08
60.0	39.41	222.49
65.0	73.30	221.35
70.0	79.98	222.25
75.0	80.10	222.55
80.0	80.30	222.81

~50% latency reduction achieved with just a ~1% increase in MR[2] %

Latency reduction is achieved up to a peak quantization noise of ~45% (θ_0)

We define the green region where $\sigma^2_\varepsilon \le \theta_0$ as the "viable" region where quality aware latency reduction is possible.

We define the red region where $\sigma^2_\varepsilon \ge \theta_0$ as the "un-viable" region where latency reduction is not possible.



(Results on the CVC-14 Dataset)

Table 1: *CVC14* Dataset

Quantization Noise (%)	Miss Rate (%)	Latency (s)
5.0	21.80	570.48
10.0	21.94	550.05
15.0	22.15	523.07
20.0	22.24	505.45
25.0	22.61	469.03
30.0	22.71	464.78
35.0	22.90	425.33
40.0	23.02	408.92
45.0	**38.43***	376.54
50.0	38.55	367.88
55.0	39.33	339.14
60.0	39.66	319.41
65.0	79.87	283.28
70.0	80.00	259.71
75.0	80.10	227.78
80.0	80.29	223.22

REFERENCES

[1] Kailai Zhou, Linsen Chen, Xun Cao. Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems. ECCV 2020.
[2] Coleman Hooper, Sehoon et al.. KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization. arXiv preprint. 2024.
[3] Shichen Dong, Wen Cheng, Jiayu Qin, Wei Wang. QAQ: Quality Adaptive Quantization for LLM KV Cache. arXiv preprint. 2024

Scan this QR code to access the GitHub Repository and learn about the code behind PedeScan!