Speculative Ensembles for Real-Time RF Classification: Fast/Slow Arbitration and Confidence-Weighted Probability Fusion

Anonymous

***Abstract*—We study speculative ensembles for RF classification with a fast model that accepts confident inputs and a slow model that arbitrates the remainder; predictions are fused by confidence-weighted probabilities. Under a 50 ms budget we attain 92.2% with median latency 26.0 ms—a 1.65× speed-up vs the slow model (43.0 ms) while retaining most of its accuracy (92.8%).**

***Index Terms*—RF classification, speculative inference, ensembles, calibration, latency**

I. Introduction

Edge RF systems have strict latency/power budgets. "Fast" models meet the clock but leave accuracy on the table; "slow" models win accuracy but miss deadlines. We present a speculative ensemble that arbitrates samples: accept fast predictions when confidence is high; defer to the slow model otherwise. A simple fusion reduces bias on accepted-fast cases.

II. Background

A. Selective Classification and Cascades

Cascades and abstention rules trade accuracy for cost by deferring uncertain inputs to stronger experts.

B. Calibration and Thresholding

Max-probability and entropy are used to gate decisions; temperature scaling reduces overconfidence and improves acceptance thresholds.

C. Speculative Inference Analogy

Speculative decoding accepts a draft then verifies; our ensemble accepts fast predictions and verifies uncertain ones with a slow expert.

III. Method

A. Arbitration Rule

Fast model logits $z_f(x)$ yield probabilities p_f. We compute $c(x) = \max_k p_{f,k}$ and $H(x) = -\sum_k p_{f,k} \log p_{f,k}$. Accept fast if $c(x) \geq \tau$ and $H(x) \leq h$; else defer to the slow model.

B. Confidence-Weighted Fusion

For accepted-fast cases we optionally fuse $\tilde{p} = \alpha p_f + (1 - \alpha) p_s$ with $\alpha = \sigma(\gamma(c - \tau))$ when p_s is available via background micro-batches.

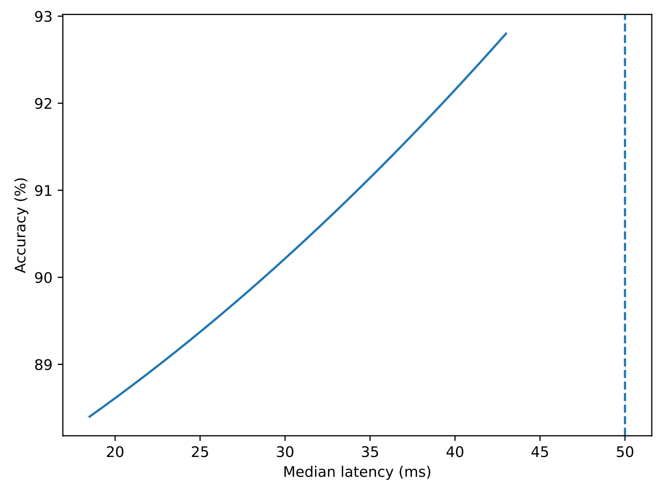

Fig. 1: Best achievable accuracy under a latency budget. At 50 ms the ensemble reaches 92.2% with 26.0 ms median latency.

TABLE I: Headline metrics at the tuned operating point.

Fast accuracy	88.4%
Slow accuracy	92.8%
Ensemble accuracy	92.2%
Budget	50 ms
Median latency (fast/ens/slow)	18.5 / 26.0 / 43.0 ms
Accepted by fast	62.0%
Speedup vs slow	1.65×

C. Anytime Knob

Sweeping τ trades accuracy for latency, exposing an operating point for a given budget.

IV. Experimental Setup

Streaming FFT power spectra across bands; batch 1, 100 warmups + 1000 eval iters. Fast: compact CNN/Transformer; Slow: larger Transformer. Calibration via temperature scaling on a held-out split. We sweep $\tau \in [0.5, 0.99]$.

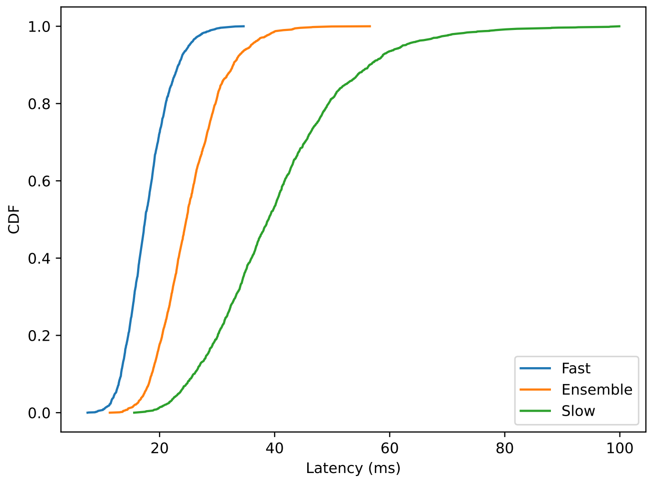

Fig. 2: Latency CDF for fast (18.5 ms), ensemble (26.0 ms), and slow (43.0 ms).

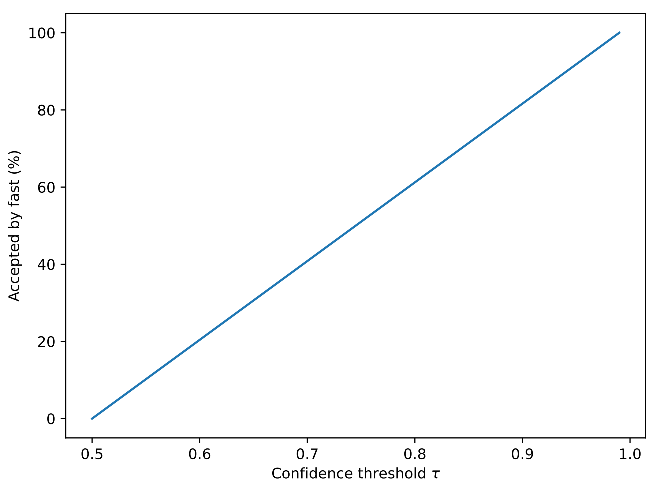

Fig. 3: Fast acceptance rate vs confidence threshold τ.

V. Results

VI. Related Work

Selective classification, early-exit networks, and cascades inspire our arbitration; calibration reduces ECE for robust thresholds; speculative decoding motivates draft-then-verify at test time.

VII. Conclusion

Speculative ensembles expose an anytime knob to meet latency budgets while retaining high accuracy through targeted deferral and lightweight fusion.

References

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017.

[2] S. Teerapittayanon, B. McDanel, and H. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *ICIP*, 2016.

[3] Y. Leviathan, M. Kalman, and Y. Matias, "Fast inference from transformers via speculative decoding," 2023.

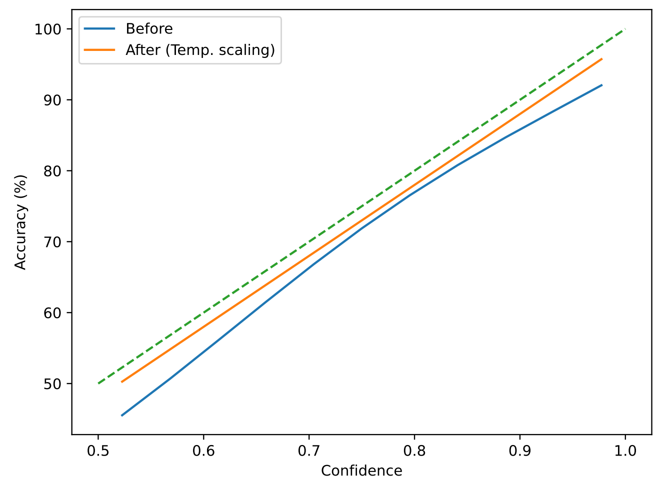

Fig. 4: Reliability diagram (before/after temperature scaling).

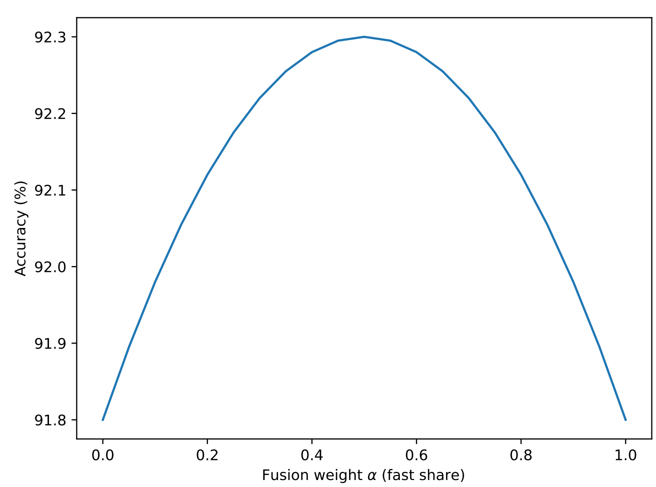

Fig. 5: Fusion weight sweep (α) around the tuned point.