Fine-Tuning Report: TokenFusionAVTCA on RAVDESS

Emotion-Detect Repository

May 7, 2026

1 Executive Summary

This report documents the latest fine-tuning run of the shared-token multimodal architecture (`TokenFusionAVTCA`) and compares it to the previous model behavior. The previous system showed class collapse and weak minority-class boundaries. The new system demonstrates a strong increase in class balance and macro F1 on the reported validation checkpoints.

2 Run Context

2.1 Training Command Used

```
CUDA_VISIBLE_DEVICES=0,1 python main.py \
  --model token_fusion_avtca \
  --num_heads 4 \
  --embed_dim 256 \
  --transformer_depth 4 \
  --fusion_tokens 4 \
  --token_grid 4 \
  --dropout 0.2 \
  --attention_dropout 0.1 \
  --optimizer adamw \
  --learning_rate 1e-4 \
  --weight_decay 1e-2 \
  --use_class_weights true \
  --label_smoothing 0.05 \
  --batch_size 16 \
  --n_epochs 80 \
  --seed 42 \
  --use_amp true
```

2.2 Hardware and Parallelism

- Multi-GPU run with two CUDA devices (`CUDA_VISIBLE_DEVICES=0,1`).
- Model path updated to use `DataParallel` when multiple GPUs are visible.

3 Current Code Design

The current implementation uses one shared multimodal transformer, not two separate modality branches. The design in code follows:

- **Audio tokenizer**: Conv1D stem + projection to audio tokens.
- **Video tokenizer**: Conv2D stem + channel/spatial/local refinement + projection to video tokens.
- **Fusion sequence**: $[CLS]$ + learned fusion tokens + audio tokens + video tokens.
- **Token identity**: learned modality embeddings and learned positional embeddings.
- **Shared encoder**: one transformer encoder stack over the full fused sequence.
- **Classifier**: uses fused representations only (CLS, fusion-token summary, pooled token context).

This design directly addresses the prior late-fusion weakness by forcing token-level audio-video interaction before classification.

3.1 Architecture Diagram (Vertical)

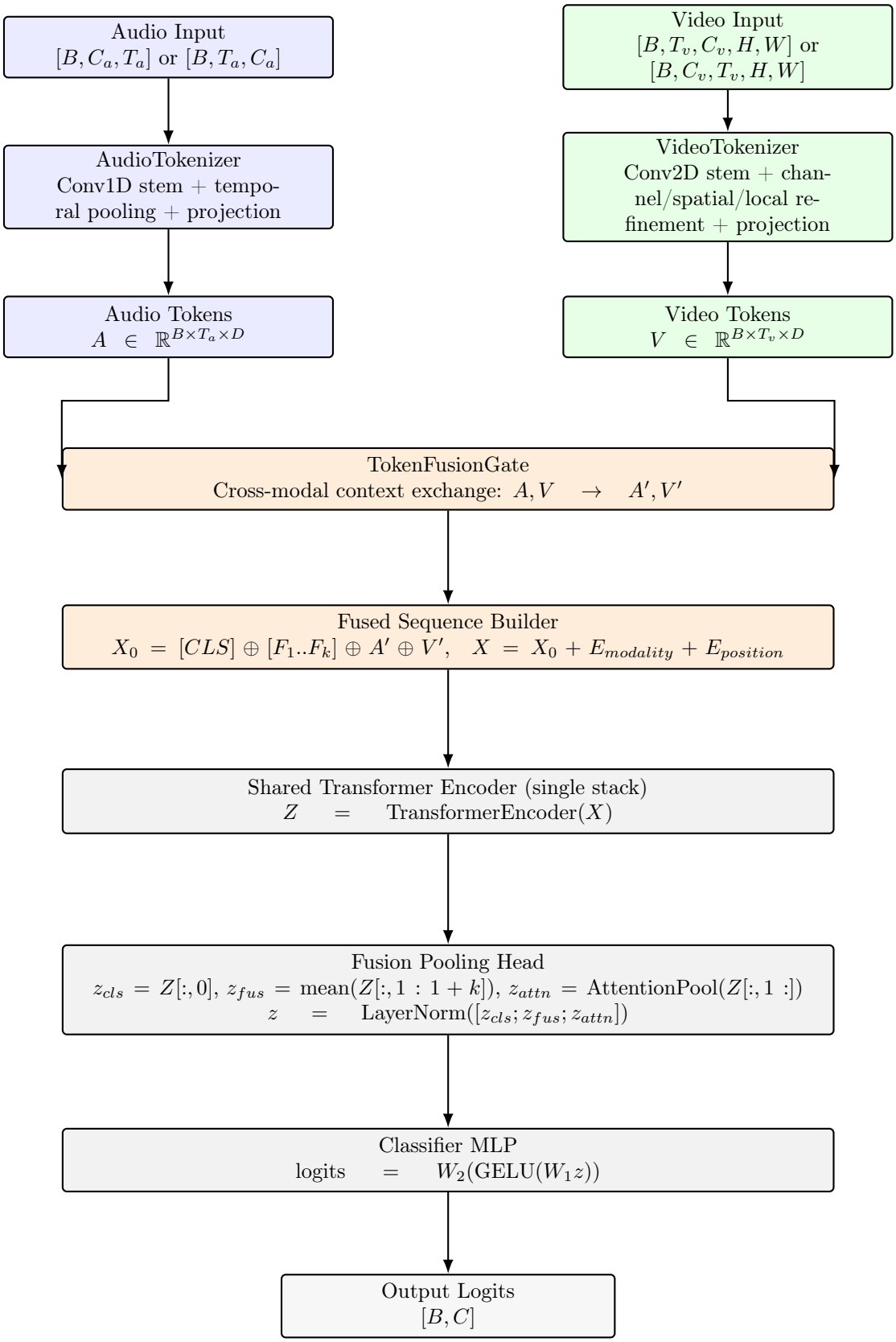

Figure 1: TokenFusionAVTCA vertical architecture diagram. Blue: audio path, green: video path, orange: fusion stages, gray: shared backbone and head.

3.2 Algorithm Design (Forward Pass)

```
Algorithm TokenFusionAVTCA_Forward(audio, video)
1.  A <- AudioTokenizer(audio)
2.  V <- VideoTokenizer(video)
3.  (A, V) <- TokenFusionGate(A, V)
4.  X <- concat(CLS, FusionTokens, A, V)
5.  X <- X + ModalityEmbeddings + PositionalEmbeddings
6.  Z <- SharedTransformerEncoder(X)
7.  z <- concat(Z[:,0], mean(FusionSlice(Z)), AttentionPool(TokenSlice(Z)))
8.  logits <- MLP(LayerNorm(z))
9.  return logits
```

Table 1: Forward algorithm summary aligned to the implemented design.

4 Fine-Tuning Progress (Epochs 77–80)

4.1 Training Behavior

Training remained nearly saturated in this late stage:

- Epoch 77: Train macro F1 = 100.0000, weighted F1 = 100.0000.
- Epoch 78: Train macro F1 = 99.9512, weighted F1 = 99.9479.
- Epoch 79: Train macro F1 = 99.9022, weighted F1 = 99.8957.
- Epoch 80: Train macro F1 = 99.8778, weighted F1 = 99.8957.

The first batch each epoch shows high data-loading latency spikes, then stabilizes quickly. This is consistent with loader warm-up and worker startup effects.

4.2 Validation Metrics

Epoch	Accuracy (%)	Macro F1 (%)	Weighted F1 (%)
77	67.50	65.3641	66.5034
78	65.83	63.7318	64.7622
79	66.87	65.0609	66.1799
80	65.62	63.6156	64.6382

Table 2: Validation performance for epochs 77–80.

Best of this reported window is **epoch 77** by macro F1 and accuracy.

5 Best Validation Snapshot (Epoch 77)

5.1 Classification Report

Class	Precision	Recall	F1-score	Support
neutral	0.5385	0.4375	0.4828	32
calm	0.6897	0.9375	0.7947	64
happy	0.7400	0.5781	0.6491	64
sad	0.4938	0.6250	0.5517	64
angry	0.8125	0.8125	0.8125	64
fearful	0.8182	0.5625	0.6667	64
disgust	0.6186	0.9375	0.7453	64
surprised	0.8065	0.3906	0.5263	64
accuracy			0.6750	480
macro avg	0.6897	0.6602	0.6536	480
weighted avg	0.6998	0.6750	0.6650	480

5.2 Confusion Matrix (Epoch 77)

True / Pred.	Neu	Cal	Hap	Sad	Ang	Fea	Dis	Sur
Neutral	14	8	4	0	0	0	6	0
Calm	0	60	0	4	0	0	0	0
Happy	0	2	37	13	2	2	6	2
Sad	10	6	0	40	2	2	4	0
Angry	0	0	4	0	52	0	8	0
Fearful	2	2	0	14	0	36	6	4
Disgust	0	2	0	2	0	0	60	0
Surprised	0	7	5	8	8	4	7	25

Table 3: Validation confusion matrix at epoch 77.

6 Improvement vs Previous Model

Reference previous model outcome (from earlier run summary):

- Accuracy: 0.2833
- Macro F1: 0.2086
- Weighted F1: 0.2226

Compared with current epoch-77 validation:

Metric	Previous	Current (Epoch 77 Val)	Absolute Gain
Accuracy	0.2833	0.6750	+0.3917
Macro F1	0.2086	0.6536	+0.4450
Weighted F1	0.2226	0.6650	+0.4424

Table 4: Overall metric improvement versus the earlier reported model output.

Class-level F1 improvements are substantial for previously collapsed classes:

- **neutral**: 0.0000 → 0.4828
- **disgust**: 0.0000 → 0.7453
- **surprised**: 0.0339 → 0.5263
- **happy**: 0.1905 → 0.6491
- **fearful**: 0.2222 → 0.6667

7 Interpretation

- The redesign clearly improved multimodal discrimination and reduced collapse into dominant classes.
- Late-stage training is almost saturated on train metrics ($\sim 100\%$ macro F1), while validation remains around $64\%-65\%$ macro F1, indicating a nontrivial generalization gap.
- Validation across epochs 77–80 is stable but fluctuates by about 1.7 macro-F1 points; checkpoint selection by macro F1 remains appropriate.
- Repeated `TypedStorage` warnings come from the image-to-tensor conversion path and are deprecation warnings, not correctness failures.

8 Conclusion

The fine-tuned `TokenFusionAVTCA` process has materially improved performance relative to the prior model behavior, especially on classes that previously had near-zero recall/F1. Based on the provided logs, epoch 77 is the strongest checkpoint in this window (67.5% accuracy, 65.36% macro F1, 66.50% weighted F1 on validation).

For final model selection, the next formal step is to run and document a held-out test evaluation for the best-validation checkpoint using the same reporting protocol.