

Corrected Analysis: Why Results Don't Match Architecture Complexity

Critical Discovery

Actual Parameter Counts (from your data):

Baseline: 3.5M parameters (Simplest architecture)
Attention: 8.0M parameters (Medium complexity)
Transformer: 8.5M parameters (Similar to Attention!)
Improved: 15M parameters (LARGEST - contrary to expectations!)

Expected Pattern:

Baseline < Improved < Attention < Transformer
3.5M < 15M < 8M < 8.5M
↑ THIS IS BACKWARDS!

Something is wrong with your implementation!

The Real Problem

Why Improved Baseline Has 15M Parameters

Looking at the models.py code, the "Improved Baseline" has:

```
python

# Deeper CNN (4 layers instead of 3)
Conv2d(1, 64)    # 64 filters
Conv2d(64, 128)  # 128 filters
Conv2d(128, 256) # 256 filters
Conv2d(256, 512) # 512 filters ← EXTRA LAYER

# BiLSTM
LSTM(512*4, hidden_dim//2, bidirectional=True)

# Decoder concaten
```