Learning Stabilization Techniques

🔥 Learning Rate Warmup

Gradually increase learning rate initially to stabilize training

```
lr = d_model^(-0.5) × min(step^(-0.5), step × warmup^(-1.5))
```

🎯 Label Smoothing

Reduces **overconfidence** in predictions

- ✓ Typical value: **ε = 0.1**
- ✓ Improves generalization

✂️ Gradient Clipping

Prevents **exploding gradients** during training

- ✓ Caps gradient norm at threshold
- ✓ Ensures training stability

💧 Dropout

Regularization through random neuron dropout

- ✓ Applied to **attention weights**
- ✓ Applied to **FFN outputs**

⚖️ Weight Initialization

Proper initialization for stable gradients

- ✓ **Xavier** initialization
- ✓ **He** initialization

⚡ Mixed Precision Training

Uses **FP16** for faster computation

- ✓ Reduces memory usage
- ✓ Speeds up training

🎓 Training Best Practices

Combine these techniques for **stable and efficient training** of Transformer models