"Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling"

[1412.3555] Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

Objective:

The paper empirically evaluates the performance of advanced gated recurrent units—**Long Short-Term Memory (LSTM)** and **Gated Recurrent Unit (GRU)**—against traditional **tanh units** in recurrent neural networks (RNNs) for sequence modeling tasks. The focus is on polyphonic music modeling and speech signal modeling.

Key Contributions:

1. **Model Comparison**:
   - **LSTM**: Uses input, forget, and output gates to regulate information flow and memory cell updates.
   - **GRU**: Simpler than LSTM, with reset and update gates to modulate candidate activations and hidden states.
   - **tanh**: Traditional RNN unit with no gating mechanisms.
2. **Experimental Setup**:
   - **Datasets**:
     - **Polyphonic Music**: Nottingham, JSB Chorales, MuseData, Piano-midi.
     - **Speech Signals**: Two raw audio datasets (Ubisoft A and B).
   - **Models**: LSTM-RNN, GRU-RNN, and tanh-RNN were sized to have comparable parameters (~20K for music, ~169K for speech).
   - **Training**: RMSProp with gradient clipping and weight noise to stabilize learning.
3. **Results**:
   - **Music Modeling**:
     - GRU slightly outperformed LSTM and tanh on most datasets (e.g., lower negative log-likelihood on JSB Chorales: GRU 6.94 vs. LSTM 8.15).
     - All models performed similarly on Nottingham.

- **Speech Modeling**:
    - Gated units (LSTM/GRU) significantly outperformed tanh (e.g., Ubisoft A test set: LSTM 2.70 vs. tanh 6.44).
    - GRU converged faster in wall-clock time for some datasets.

4. **Key Findings**:
    - **Gated units (LSTM/GRU) consistently surpassed tanh units**, especially in complex tasks like speech modeling.
    - **GRU vs. LSTM**: No clear winner; performance depended on the dataset. GRU often trained faster, but LSTM occasionally achieved better final accuracy.
    - **Additive Gates Matter**: Both LSTM and GRU's additive updates (e.g., memory cell updates in LSTM, interpolation in GRU) mitigated vanishing gradients and improved long-term dependency capture.
5. **Limitations**:
    - Preliminary results; deeper analysis of individual gating components is needed.
    - Performance variability suggests task-specific optimal architectures.

Conclusion:

The study confirms the superiority of gated units (LSTM/GRU) over traditional tanh RNNs for sequence modeling. While GRUs showed faster convergence and competitive performance, LSTMs remained strong contenders. The choice between LSTM and GRU may hinge on specific task requirements. Future work should dissect the roles of individual gating mechanisms to guide architecture selection.