RNN, LSTM, and GRU: A Mathematical and Visual Tutorial

Contents

1 Introduction

Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) are designed to model temporal dependencies in sequential data. They are widely used in natural language processing, speech recognition, and time series forecasting.

2 Recurrent Neural Networks (RNNs)

2.1 Concept

RNNs introduce recurrence by reusing hidden states across time steps. This allows them to "remember" past information.

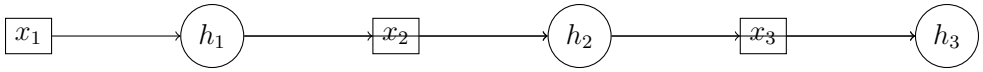
2.2 Mathematical Formulation

Given input x_t at time t, and previous hidden state h_{t-1}:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b)$$

$$y_t = W_{hy}h_t + b_y$$

2.3 Diagram



2.4 Limitations

- **Vanishing gradients**: Difficult to learn long-term dependencies.
- **Sequential computation**: Cannot parallelize across time steps.

3 Long Short-Term Memory (LSTM)

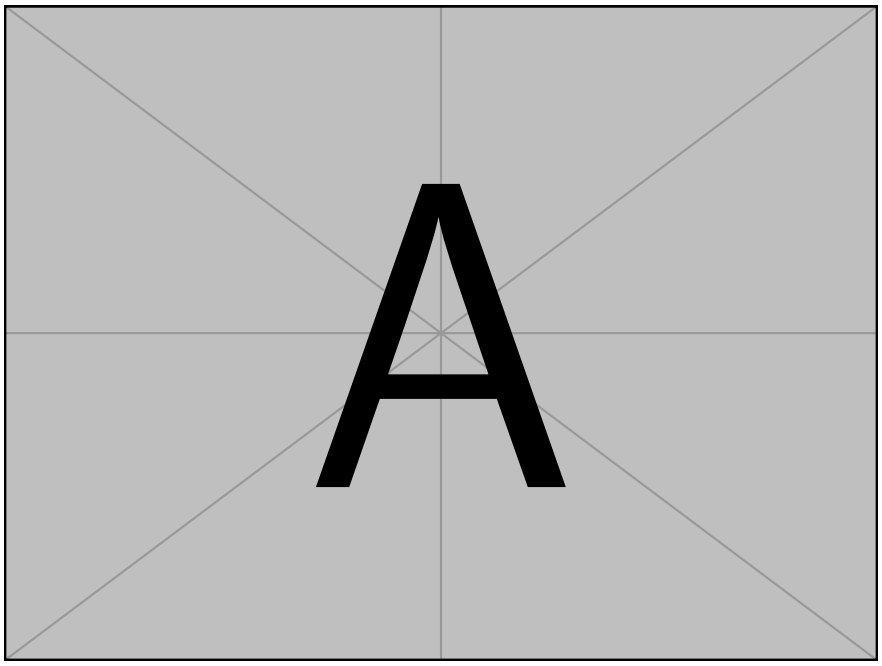
3.1 Intuition

LSTMs use gates to control what information is remembered or forgotten. This solves the vanishing gradient problem by allowing constant error flow through time.

3.2 Equations

$$
\begin{aligned}
f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \quad \text{(forget gate)} \\
i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \quad \text{(input gate)} \\
\tilde{c}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c) \quad \text{(cell candidate)} \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad \text{(cell state)} \\
o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \quad \text{(output gate)} \\
h_t &= o_t \odot \tanh(c_t) \quad \text{(hidden state)}
\end{aligned}
$$

3.3 Diagram (Conceptual)



3.4 Explanation

- **Forget gate**: Decides what information to discard.
- **Input gate**: Decides what new information to store.
- **Cell state**: Memory that flows unchanged across time (modulo gating).
- **Output gate**: Controls what is exposed to the next layer.

4 Gated Recurrent Unit (GRU)

4.1 Simplified Design

GRUs merge the forget and input gates into a single update gate and use a reset gate for memory control.

4.2 Equations

$$
\begin{aligned}
z_t &= \sigma(W_z[h_{t-1}, x_t]) \quad \text{(update gate)} \\
r_t &= \sigma(W_r[h_{t-1}, x_t]) \quad \text{(reset gate)} \\
\tilde{h}_t &= \tanh(W[r_t \odot h_{t-1}, x_t]) \quad \text{(candidate state)} \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t
\end{aligned}
$$

4.3 Advantages

- Fewer parameters than LSTM.
- Often similar performance with less training time.

5 Comparison Table

Feature	RNN	LSTM	GRU
Handles long-term memory	Poor	Excellent	Good
Parameter complexity	Low	High	Medium
Training stability	Low	High	High
Uses memory cell	No	Yes	No
Gate types	None	Forget/Input/Output	Update/Reset

6 Applications

- Language Modeling
- Speech Recognition
- Time Series Forecasting
- Video Captioning