Decoding Long Short-Term Memory: A Story of Memory and Machines

Sambit K Satapathy

January, 2024

Introduction: Forget me not

Have you ever blanked on someone's name moments after being introduced? Or struggled to remember the answers to a question you just read? We, humans, tend to be masters of short-term memory, juggling conversations and tasks with ease. But our brains can be a bit leaky when it comes to the long haul. Imagine superheroes unable to recall past events; they'd be powerless in the face of ever-changing challenges. This is where the Long Short-term memory (LSTM), our protagonist, a type of artificial intelligence, steps in, ready to become the keeper of your neural network's forgotten thoughts. It's like turning your memory into a superpower, allowing you to recall and process information over extended periods.

Decoding Neural Progress: RNNs to LSTMs

Now, let's delve deeper into the inner workings of LSTMs. Picture our brains as traditional Recurrent Neural Networks (RNNs), information highways where data zips around like rush hour traffic. These networks excel at short-term tasks, seamlessly juggling conversations and remembering what you ate for breakfast. But when it comes to long-term memory, they're like that forgetful friend who consistently loses their keys. This notorious flaw is known as the *'Vanishing Gradient Problem'*. It's like whispers fading down a long hallway or forgotten melodies in a crowded room, with crucial details getting lost as information travels through the network. But fear not, for the LSTM arrives with a toolbox of ingenious solutions. Think of it as a memory palace architect, building a network with special "gates" that control information flow.

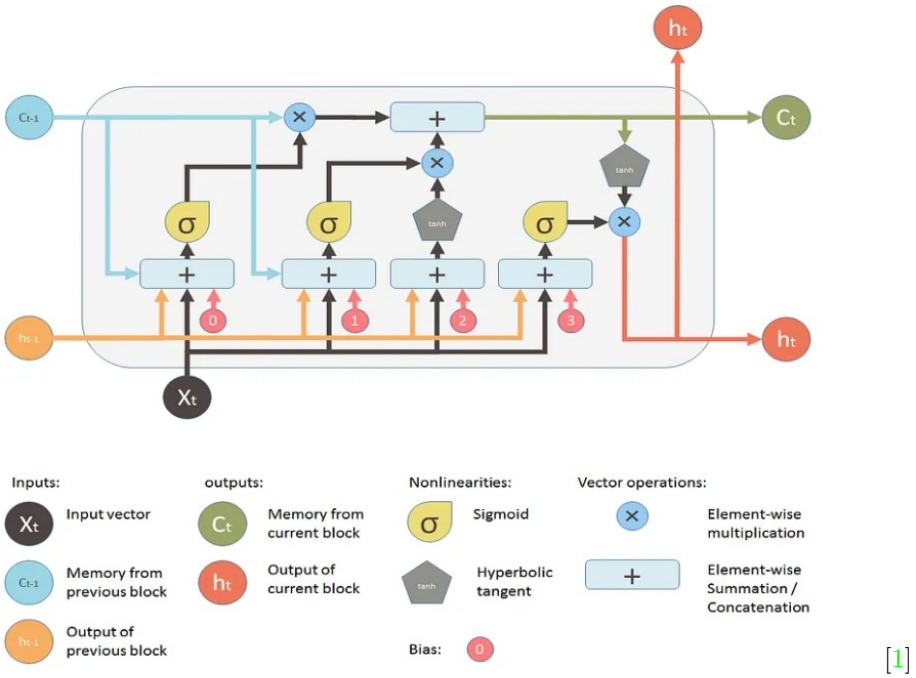
The Birth of LSTM: A Flashback

Transport yourself to the late 1990s, an era marked by the struggles of data scientists grappling with the limitations of existing models in processing sequential data. Picture a landscape where traditional models falter, unable to capture long-term dependencies effectively. In a lab far from the Silicon Valley buzz, two minds, Sepp Hochreiter and Jürgen Schmidhuber, embarked on a groundbreaking journey that would reshape the landscape of sequential data processing. The result? A revolutionary design: the Long Short-Term Memory (LSTM). Like vigilant guards, its gates control information flow, preserving precious memories across time.

LSTM is no ordinary model; it's the superhero with an extraordinary memory. It's the Batman of machine learning, equipped with the ability to remember and learn from the past. This distinguishing feature elevates LSTM into an indispensable force in the realm of sequential data. LSTM's prowess in capturing long-term dependencies and contextual nuances within sequential data makes it particularly valuable. LSTM networks find extensive applications in various real-world scenarios due to their ability to handle sequential data and capture long-term dependencies, including Natural Language Processing (NLP), speech recognition, machine translation, stock market prediction, autonomous vehicles, and medical diagnosis.

Anatomy of LSTM: Unveiling the Superpowers

Dive into the core of the Long Short-Term Memory (LSTM), and you'll discover precisely engineered architecture and a symphony of gates, functions, and pointwise operations. The LSTM's structure is similar to a superhero suit, with each component playing a distinctive role.



[1]

Figure 1: LSTM architecture along with notations

Figure 1 illustrates the internal structure of LSTM along with notations for each element. This may initially seem intimidating, but let me demystify each term for simplicity.

1. **Cell State: The Memory Reservoir**

   - At the core of LSTM lies the cell state, denoted as C_t, acting as a memory reservoir that retains information over time. This reservoir is pivotal to LSTM's ability to capture and preserve long-term dependencies. It can be termed as the *long term memory*.

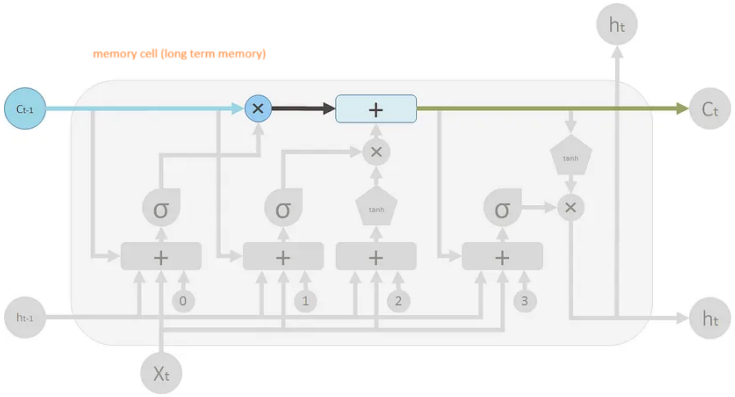


Figure 2: Memory cell (Long term Memory)

2. **Forget Gate: Erasing the Unnecessary**

- The forget gate, denoted as f_t, functions as an editor, erasing irrelevant or outdated information from the cell state. It employs the sigmoid activation function to decide what information to discard.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

- After calculating ft, the forget gate's work isn't done yet. It embarks on another crucial task: meticulously blending ft with the previous cell state, c(t-1), to determine what information will be retained and what will be gracefully forgotten.

$$c_{t-1} * f_t$$

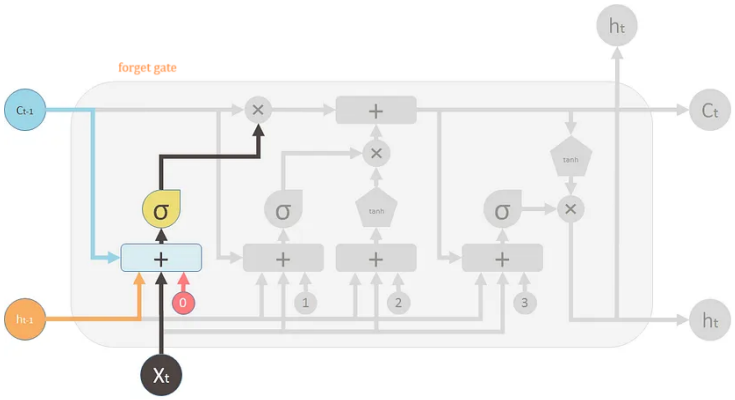


Figure 3: Forget Gate

3. **Input Gate: Decision-Maker Extraordinaire**

- The input gate, denoted as i_t, serves as a decisive force. It determines the relevance of incoming information, employing the sigmoid activation function and pointwise multiplication to decide what data should be stored in the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

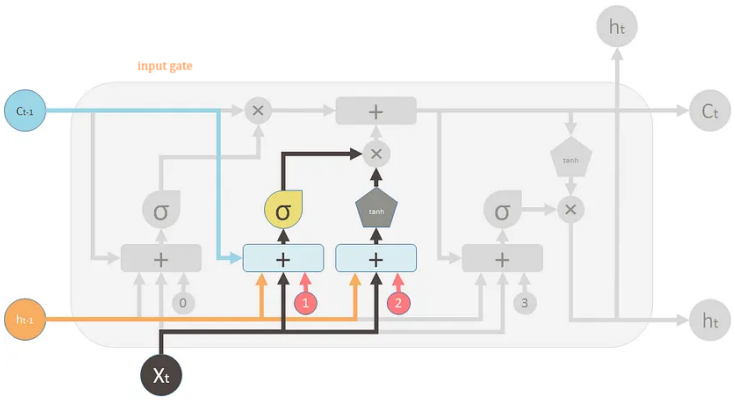


Figure 4: Input Gate

- $\bar{c}_t$ computes a new candidate value for the cell state. It is that information LSTM thinks can be added to the memory cell.

$$\bar{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

- c_t, the cell state, is the heart of LSTMs, where long-term memory is stored and selectively updated. The cell state is updated by meticulously combining the previous cell state c_{t-1}, the forget gate's decision f_t and the candidate cell state $\bar{c}_t$. This delicate balance ensures that only the most relevant information persists while unnecessary details fade away.

$$c_t = (f_t * c_{t-1}) + (\bar{c}_t * i_t)$$

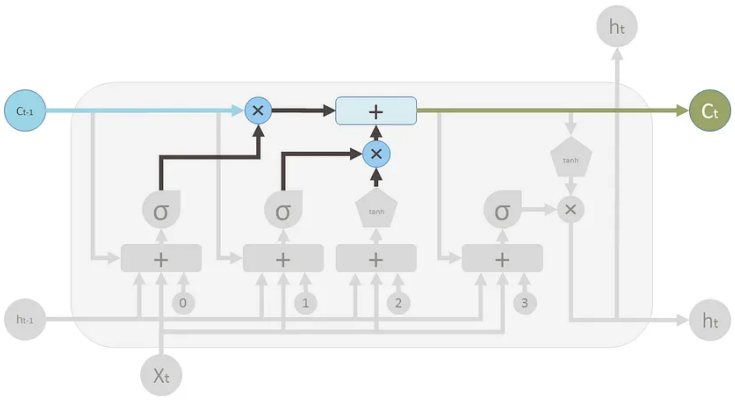


Figure 5: Cell state

4. **Output Gate: Producing the Finale**

- The output gate, denoted as o_t, produces the final prediction based on the refined cell state. It employs the sigmoid and 'tanh' activation functions to control the flow of information and produce the output.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

- h_t produces the final hidden state, influenced by the output gate and cell state candidate, c_t.

$$h_t = o_t * \tanh(c_t)$$

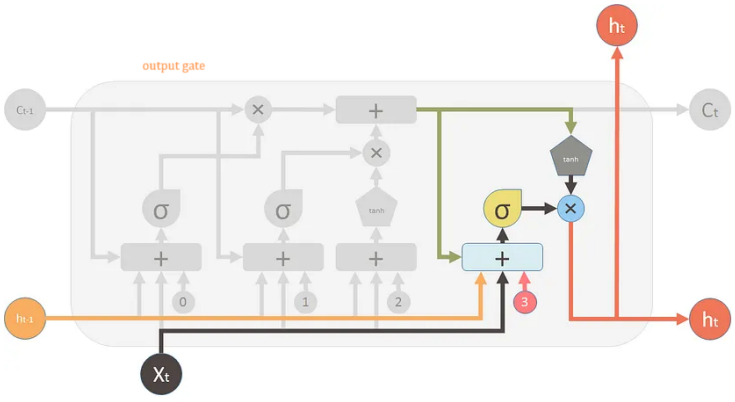


Figure 6: Output Gate

Notations:

- W_i: weight matrices for corresponding gates
- b_i: bias term of corresponding gates
- x_t : current input (at time t)
- h_{t-1} : previous hidden state (at time t-1)
- h_t : current hidden state
- σ : sigmoid activation function
- tanh : Hyperbolic tangent activation function
- $[h_{t-1}, x_t]$: Concatenation of the previous hidden state (h_t) and the current input (x_t).
- $W_c,\ b_c$: Weight matrix and bias term of the cell state

These formulas illustrate the dynamic interactions of gates and pointwise operations, showcasing LSTM's prowess in managing information. [2]

Training an LSTM: Superhero Training Montage

Like a superhero undergoes intense training, an LSTM learns from data through a training montage of backpropagation through time and gradient descent, correcting flaws in each iteration, unraveling the complexities of natural language, anticipating stock market trends, and forecasting weather patterns. It refines its abilities, becoming adept at understanding and predicting sequences, a skill crucial in various real-world applications.

LSTM in Action: Real-Life Superhero Scenarios

Climate change, a global challenge, brings rising temperatures and unpredictable disasters. Introduce the Long Short-Term Memory (LSTM), a quiet hero in our environmental narrative. LSTMs act as climate detectives, analyzing vast datasets to predict extreme weather events, optimize renewable energy use, and track carbon footprints. By learning from historical data, they offer early warnings, enhance energy efficiency, and identify significant polluters. LSTMs, with their analytical prowess, provide hope in our battle against climate change. Perhaps we can script a more optimistic ending for our planet's future through data-driven insights.
However, the applications of LSTMs extend far beyond climate change. These versatile tools can:

- Medical Diagnosis: Early disease detection and personalized treatment plans through patient data analysis.
- Language Translation: Revolutionizing translation by capturing language nuances for accurate and natural-sounding results
- Scientific applications: Modeling complex sequences, aiding research in genomics, drug discovery, and physics, unraveling patterns, and predicting outcomes for groundbreaking advancements.
- Autonomous Vehicles: predicting the movement of other vehicles, pedestrians, and objects. This helps in making real-time decisions and ensuring the safety of the vehicle and its passengers.
- Gesture Recognition: Employed in recognizing gestures and movements, making them valuable in applications like sign language interpretation, virtual reality, and human-computer interaction.

Navigating Challenges in LSTM's Path

While Long Short-Term Memories (LSTMs) possess formidable capabilities, they are not invincible, facing challenges analogous to a superhero's kryptonite. LSTMs may demand substantial computational resources, particularly in deep architectures, resulting in extended training times and resource requirements. Similar to many intricate models, they are prone to overfitting, particularly when dealing with small datasets, potentially leading to poor generalization.

Future of LSTMs: The Next Chapter

As we reach the climax of our story, we peer into the future. While LSTMs have unlocked remarkable possibilities, their future calls for addressing fundamental limitations. Researchers are actively seeking improvements in computational efficiency, interpretability, and user-friendliness. However, inherent constraints such as sequential processing, which hinders parallelization, present significant hurdles. Now, there's a new player on the scene - Transformers. They're changing how we handle sequences, using parallel processing and *attention* tricks. Transformers seem to have answers to some of the problems LSTMs face. So, the story shifts to Transformers, opening a new chapter in the exciting world of machine learning, but that's a story for another time.

Conclusion: A Call to Action

From forgetful friends to climate warriors, LSTMs have taken us on a remarkable journey through the realm of long-term memory. These digital architects have skillfully constructed networks that remember, predict, and learn, influencing everything from predicting the weather to creating music.

While challenges remain, LSTM's impact is undeniable. It has revealed the power of uncovering forgotten information, emphasizing that even the briefest memory holds the key to significant solutions. As we transition to the Transformers' story, embracing attention and parallelization, we eagerly anticipate the remarkable feats ahead in the ever-evolving tale of machine learning. The future holds abundant possibilities, driven by the memories of the past and the innovations of the present.

References

[1] ML Review. Understanding lstm and its diagrams. https://blog.mlreview.com/understanding-lstm-and-its-diagrams-37e2f46f1714.

[2] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *ArXiv preprint arXiv:1402.1128*, 2014.