Deep Learning and Reinforcement Learning

A Report on Deep Learning Trend - Transformers

1. Introduction to Modern Deep Learning Trend: Transformers

In recent years, a powerful new model called the **Transformer** has completely changed how computers understand and work with language. This model is mainly used in a field called **Natural Language Processing (NLP)**, which involves teaching machines to read, write, and understand human language — like answering questions, translating text, or analysing the mood of a review.

Before Transformers, we used models like **RNNs (Recurrent Neural Networks)** and **LSTMs (Long Short-Term Memory networks)**. These older models read text **one word at a time**, in order. While this helped in understanding the meaning, it was **slow** and **struggled to remember long sentences**.

Then in 2017, a group of researchers led by **Vaswani et al.** introduced a new idea in their famous paper **"Attention is All You Need"**. They created the **Transformer model**, which brought a major change:

➔ Instead of reading one word at a time, Transformers read the whole sentence at once.

➔ They use something called **self-attention**, which allows the model to focus on the most important words, no matter where they appear in the sentence.

⇨ **<u>Example</u>:**

If the sentence is:

"The cat that was chased by the dog ran away."

The word "cat" is the one actually doing the running — even though "dog" comes right before "ran".

Traditional models might get confused here, but Transformers can understand this relationship clearly by comparing all words with each other using self-attention.

2. Core Idea Behind Transformers

The Transformer architecture is built upon three fundamental components:

- **Self-Attention Mechanism**:
  - ➪ It allows the model to focus on important words in a sentence.
  - ➪ Each word looks at every other word and decides how much attention it should give to them.
  - ➪ **Example**:

    In the sentence *"The girl who won the prize was happy"*,

    the word "was" should be more connected to "girl" rather than "prize".

    Self-attention helps the model figure this out automatically.
- **Positional Encoding**:
  - ➪ It adds special numbers to each word to indicate its position (1st, 2nd, 3rd, etc.)
  - ➪ This helps the model understand **word order and sentence structure**, which is important for meaning.
  - ➪ **Example**:

    "The dog chased the cat" is different from *"The cat chased the dog"* —

    Even though the words are the same, their positions change the meaning.
- **Multi-head Attention and Feedforward Layers**:
  - ➪ **Multi-head attention** allows the model to look at the sentence in different ways at the same time — like seeing the same sentence from multiple perspectives.
  - ➪ **Feedforward layers** are simple networks that help the model learn complex patterns after attention is calculated.

  Unlike RNNs, which process text word-by-word, Transformers operate on the **entire sequence at once**, making them more efficient and better at capturing long-term dependencies.

---

3. Our Project: Text Classification with Transformers

In this project, we worked on classifying movie reviews into two categories:

Positive / Negative

We used a famous dataset called IMDB movie reviews, which contains 25,000 real user reviews. The goal was to train computers to understand the sentiment (feeling or emotion) behind each review using deep learning models.

To do this, we compared two different models:

- ❖ Model 1: **BiLSTM (Bidirectional Long Short-Term Memory)**
  - ⇨ This is a type of Recurrent Neural Network (RNN) that reads text from both directions – from left to right and right to left.
  - ⇨ It helps the model understand the full context of a sentence.
  - ⇨ We built this model from scratch using Keras, a deep learning library.
  - ⇨ It uses word embeddings (numerical representations of words) and a BiLSTM layer to make predictions.
  - ● Advantage: Understands the order of words well.
  - ● Limitation: Slower and less accurate with long reviews or large data.

- ❖ Model 2: DistilBERT (Transformer-Based)
  - ⇨ DistilBERT is a smaller, faster version of the well-known BERT model.
  - ⇨ It comes pre-trained — meaning it already understands a lot about language by learning from millions of sentences before we even use it.
  - ⇨ We used the Hugging Face Transformers library to fine-tune this model on our IMDB dataset.
  - ⇨ It's based on the Transformer architecture, which looks at the entire sentence at once using self-attention.
  - ● Advantage: Faster training, better accuracy, and great performance on longer texts.
  - ● Limitation: Needs more memory than basic models (but less than full BERT).

- ❖ **What We Did**:

1. Trained both models on the same IMDB dataset.
2. Tested their accuracy and performance.

3. Used explainability tools like **LIME** and **SHAP** to understand what the models were focusing on during prediction.
4. We also created a special metric called Conflict Score to detect high-risk errors — predictions where the model is very confident but still wrong. These are especially important in real-world use cases.

❖ What We Found:

1. DistilBERT performed better in accuracy and confidence.
2. BiLSTM showed overconfidence in some wrong predictions.
3. Conflict score and explainability tools helped identify failure modes.

---

4. Key Concepts Covered

❖ **Transformers:** Understanding Language as a Whole

- ⇨ Transformers are a powerful deep learning architecture.
- ⇨ Unlike older models (like RNNs or LSTMs), Transformers look at the entire sentence all at once instead of word by word.
- ⇨ This helps the model understand context better and process longer text faster using a mechanism called self-attention.

❖ **Fine-Tuning Pre-trained Models:** Save Time, Improve Results

- ⇨ Instead of training a model from scratch (which takes a lot of data and time), we used a pre-trained model called DistilBERT.
- ⇨ It already understood English well because it was trained on large amounts of data like books and Wikipedia.
- ⇨ We only needed to "fine-tune" it — that means training it a bit more on our own dataset (IMDB movie reviews) so it could learn our specific task: sentiment classification.

- **Hugging Face Library:** The Toolkit We Used
  - Hugging Face is a popular open-source library for NLP.
  - It provides:
    - Pre - trained models like BERT, RoBERTa, DistilBERT
    - Easy-to-use tools to train, test, and use these models
    - Tokenizers to prepare text input for the models

- **Hugging Face Tokenizer:** Converting Text to Model Input
  - Machines can't read raw text — they need numbers.
  - The DistilBertTokenizerFast helped us:
    - Break sentences into smaller parts (tokens)
    - Convert those tokens into numerical IDs
    - Make sure all inputs are of equal length using padding/truncation
- **DistilBERT (via Hugging Face):** The Pre-trained Brain
  - We used TFDistilBertForSequenceClassification, a smart model already trained to understand English.
  - We just fine-tuned it to classify reviews as positive or negative.

---

5. Key Applications of Transformers

Transformers are now ubiquitous across multiple domains:

Domain	**Application**
NLP	Text classification, summarization, translation
Healthcare	Biomedical literature mining, diagnostics
Finance	Sentiment analysis, fraud detection
Vision	Vision Transformers (ViTs) for image tasks
Gaming & Robotics	Strategy learning, motion planning

6. Future Potential of Transformers

Transformers are evolving beyond NLP into broader, smarter AI systems. Some key trends include:

- Multimodal Transformers

Models that process text, images, and audio together — like GPT-4 and Gemini — enabling tasks such as describing images or understanding videos.

- Efficient Transformers

New designs like Longformer and Linformer reduce memory and speed up training, making Transformers usable on longer texts and smaller devices.

- Instruction-Tuned Models

Models like ChatGPT and LLaMA can follow plain-language instructions (e.g., “Summarize this article”), making AI more user-friendly.

- Edge Deployments

Lightweight Transformer models are being adapted to run on mobile devices and IoT systems, allowing smart applications without needing internet/cloud access.

---

7. Conclusion

This project demonstrated the power of Transformer models for text classification. We compared an older BiLSTM model with DistilBERT, a pre-trained Transformer from the Hugging Face library, fine-tuned on the IMDB dataset.

Key Outcomes:

- Higher accuracy and faster training with DistilBERT
- Better understanding of long-range dependencies
- Explainable AI using LIME and SHAP to visualize predictions