राष्ट्रीय प्रौद्योगिकी संस्थान दिल्ली

**NATIONAL INSTITUTE OF TECHNOLOGY DELHI**

**Department of Electrical Engineering**

**Project Synopsis**

|  |  |  |
| --- | --- | --- |
| ***Student & Supervisor Details*** | | |
| Name of Student | **Sourabh Shiroti** | |
| Roll Number | **221230061** | |
| Course (Branch), Semester | **B. Tech./M. Tech (EE)- …3……**  **Semester-6** | **Project Code- EEP 353** |
| Supervisor’s Name | **Dr. Gautam Kumar Sir** | |
|  |  | |
| ***Project Details*** | | |
| Title of the Project | NLP Transformer-based Models | |
| Base Paper (Main Reference Paper) | The **base paper** (main reference paper) for your project on transformer-based models in Natural Language Processing (NLP) would typically be the foundational work that introduced the transformer architecture, as it set the stage for most of the subsequent developments in this area.  The **main reference paper** for transformer-based models is:  **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., “Attention is all you need,” *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008. [Online]. Available:** [**https://arxiv.org/abs/1706.03762**](https://arxiv.org/abs/1706.03762)  This paper introduces the **Transformer architecture**, which relies solely on self-attention mechanisms (and discards the recurrent layers commonly used in prior architectures like RNNs and LSTMs). The transformer has become the backbone of many state-of-the-art NLP models, including **BERT**, **GPT**, **T5**, and **XLNet**. | |
| Brief Introduction | Transformer-based models have revolutionized Natural Language Processing (NLP) by enabling deep contextual understanding and efficient parallel processing. Introduced in the paper *"Attention Is All You Need"* (2017) by Vaswani et al., the **Transformer architecture** uses **self-attention mechanisms** and **positional encodings** to process text efficiently.  Unlike traditional recurrent models (RNNs, LSTMs), Transformers handle **long-range dependencies** better and allow **parallel computation**, making them highly scalable for large datasets.  Popular Transformer-based models include **BERT (Bidirectional Encoder Representations from Transformers)** for contextual understanding, **GPT (Generative Pretrained Transformer)** for text generation, and **T5 (Text-to-Text Transfer Transformer)** for various NLP tasks. These models power applications like chatbots, translation, summarization, and sentiment analysis. | |
| Motivation of Project | Traditional NLP models, such as RNNs and LSTMs, struggle with long-range dependencies, sequential processing, and context retention. The **Transformer architecture** solves these issues using self-attention mechanisms, enabling faster, more accurate, and scalable NLP applications. With the rise of **large-scale text generation, machine translation, and sentiment analysis**, Transformer-based models like **BERT, GPT, and T5** have become industry standards. This project aims to leverage Transformer-based models for [specific application: text classification, chatbot, summarization, etc.], enhancing efficiency, accuracy, and user experience. | |
| Related Work (Literature Review with reference) | Transformer-based models have significantly advanced the field of **Natural Language Processing (NLP)** by overcoming limitations of traditional models like RNNs and LSTMs. Below is a review of key research papers and developments in this field.   1. **BERT (Bidirectional Encoder Representations from Transformers)**   ***Reference:* Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. *Contribution:***   * **Introduced BERT (Bidirectional Encoder Representations from Transformers).** * **Utilized Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) for unsupervised learning.** * **Significantly improved NLP tasks like question answering, sentiment analysis, and named entity recognition (NER)**  1. **RoBERTa (Robustly Optimized BERT Pretraining Approach)**   ***Reference:* Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *Contribution:***   * **Improved BERT by removing NSP, using larger datasets, and applying longer training cycles.** * **Achieved better performance on NLP tasks like GLUE, SQuAD, and RACE benchmarks.**  1. **DistilBERT (Distilled version of BERT)**   **DistilBERT (Distilled BERT) is a smaller, faster, cheaper version of BERT, designed to retain most of its performance while reducing computational cost. It was introduced by Hugging Face in the paper:**  ***Reference:* Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.**   1. **ALBERT (A Lite BERT)**   **ALBERT (A Lite BERT) is a variant of BERT designed to reduce the model size and computational cost while maintaining its performance on NLP tasks. It was introduced by Google Research and Toyota Technological Institute in the paper:**  ***Reference:* Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations.**   1. **XL Net**   **XLNet is an advanced model that combines the benefits of autoregressive models (like GPT) and autoencoding models (like BERT) to achieve state-of-the-art performance on various NLP tasks. It was introduced by Google AI and CMU in the paper:**  ***Reference:* Yang, Z., Dai, Z., Yang, Y., Salakhutdinov, R., & Cohen, W. W. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding.**   1. **T5 (Text-to-Text Transfer Transformer)**   **T5 (Text-to-Text Transfer Transformer) is a model introduced by Google Research that reimagines all NLP tasks as text-to-text tasks, where both the input and output are treated as text strings. This unification allows T5 to handle a wide variety of NLP tasks with the same model and architecture, making it a versatile and powerful solution. It was presented in the paper:**  ***Reference:* Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.** | |
| Research Gap | Despite the impressive advancements in **transformer-based models** such as **BERT**, **GPT**, **T5**, and **XLNet**, there remain several challenges and open research gaps in the field of Natural Language Processing (NLP). These gaps present opportunities for further innovation and improvement in transformer-based architectures. | |
| Objective | The main objective of this project is to **explore, evaluate, and enhance transformer-based models** for Natural Language Processing (NLP) tasks.  The project aims to **enhance the understanding and utility of transformer-based models** in solving complex NLP challenges, making them more **efficient**, **interpretable**, and **suitable** for a wide range of applications. The outcomes of the project could contribute to improving NLP technology and make it more accessible and reliable for various industries. | |
| Tentative Work Plane/Methodology | To achieve the objectives of the project on **transformer-based models for NLP tasks**, the work plan can be structured into clear phases, with each phase focusing on specific tasks and deliverables. Below is a proposed methodology for the project:   | **Phase** | **Duration** | **Key Tasks** | | --- | --- | --- | | Phase 1: Literature Review | 1-2 weeks | Review transformer models, NLP tasks, and research gaps | | Phase 2: Model Selection | 1-2 weeks | Select models, collect datasets, preprocess data | | Phase 3: Fine-Tuning | 3-4 weeks | Fine-tune models, evaluate performance, hyperparameter tuning | | Phase 4: Optimization | 2-3 weeks | Model pruning, quantization, long-range dependency handling | | Phase 5: Multimodal & Cross-Lingual | 2-3 weeks | Multimodal integration, cross-lingual transfer learning | | Phase 6: Ethical Considerations | 1-2 weeks | Bias detection and mitigation, ethical implications | | Phase 7: Evaluation & Documentation | 1-2 weeks | Final evaluation, report writing, presentation preparation | | Phase 8: Conclusion | 1 week | Review work, provide conclusion, suggest future work | | |
| References (In IEEE format) |  **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.**, “Attention is all you need,” *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008. [Online]. Available: <https://arxiv.org/abs/1706.03762>   **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.**, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186. [Online]. Available: <https://arxiv.org/abs/1810.04805>   **Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I.**, “Improving language understanding by generative pre-training,” *OpenAI Blog*, 2018. [Online]. Available: <https://openai.com/research/language-unsupervised>   **Sanh, V., Debut, L., Chaumond, J., & Wolf, T.**, “DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019. [Online]. Available: <https://arxiv.org/abs/1910.01108>   **Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R.**, “ALBERT: A lite BERT for self-supervised learning of language representations,” *Proceedings of ICLR*, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7aEtvS   **Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V.**, “XLNet: Generalized autoregressive pretraining for language understanding,” *Proceedings of NeurIPS*, 2019, pp. 5754–5764. [Online]. Available: <https://arxiv.org/abs/1906.08237>   **Raffel, C., Shinn, C., Collobert, R., Dehghani, M., & Lee, L.**, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1-67, 2020. [Online]. Available: <https://arxiv.org/abs/1910.10683>   **Liu, X., He, P., Chen, W., & Gao, J.**, “Multi-task deep neural networks for natural language understanding,” *Proceedings of ACL*, 2019, pp. 1-10. [Online]. Available: <https://arxiv.org/abs/1901.11504>   **Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D.**, “Electra: Pre-training text encoders as discriminators rather than generators,” *Proceedings of ICLR*, 2020. [Online]. Available: <https://arxiv.org/abs/2003.10555>   **Sun, T., Qiu, X., Xu, Y., & Huang, X.**, “How to fine-tune BERT for text classification?,” *Proceedings of the 2019 Chinese Computational Linguistics Conference (CCL)*, 2019. [Online]. Available: <https://arxiv.org/abs/1905.05583> | |

Signature of Student with Date Signature of Supervisor with Date