SMARTICS Journal, Vol.10 No. 1 2024. p33-38
ISSN online: 2476-9754, ISSN print: 2623-0429
DOI : https://doi.org/10.21067/smartics.v10i1.8351

Tinjauan Perkembangan Kecerdasan Buatan Berbasis Arsitektur *Transformer*

Bayu Firmanto[a,*], As'ad Shidqy Aziz[a], Jendra Sesoca[a]
[a]Universitas Wisnuwardhana, Jl.Danau Sentani No 99, Malang, Indonesia
[*]correspondence email : bayufirmanto@wisnuwardhana.ac.id

***Abstract*— Artificial Intelligence, especially technique utilizing machine learning using transformer architecture has experienced rapid progress. The transformer architecture was first introduced in 2017 and laid the foundation for the development of larger and more accurate models in NLP, some of which use BERT and GPT. This review examines five studies that have made significant contributions to the development of the transformer architecture, including research by Vaswani, Devlin, Brown, and Dai. The results of this study shows that the transformer architecture is capable of improving training efficiency, accuracy, and long-context understanding in various NLP tasks. However, there are still some issues with this technology that need to be addressed further.**

***Index Terms*— artificial intelligence, transformer, natural language processing, BERT, GPT-3**

***Abstrak*—Teknologi kecerdasan buatan khususnya teknik berbasis pembelajaran mesin menggunakan arsitektur transformer telah mengalami kemajuan pesat. Arsitektur transformer pertama kali diperkenalkan pada 2017 dan membangun pondasi pengembangan model yang lebih besar dan lebih akurat dalam masalah NLP, beberapa diantaranya menggunakan BERT dan GPT. Tinjauan ini mengkaji lima penelitian yang berkontribusi signifikan dalam perkembangan arsitektur transformer, antara lain penelitian oleh Vaswani, Devlin, Brown, dan Dai. Hasil kajian ini menunjukkan bahwa arsitektur transformer mampu meningkatkan efisiensi pelatihan, akurasi, serta pemahaman konteks panjang pada berbagai tugas NLP. Namun, masih ada beberapa permasalahan dengan teknologi ini yang perlu diatasi lebih lanjut.**

***Kata Kunci*—kecerdasan buatan, transformer, *natural language processing*, BERT, GPT-3**

I. PENDAHULUAN

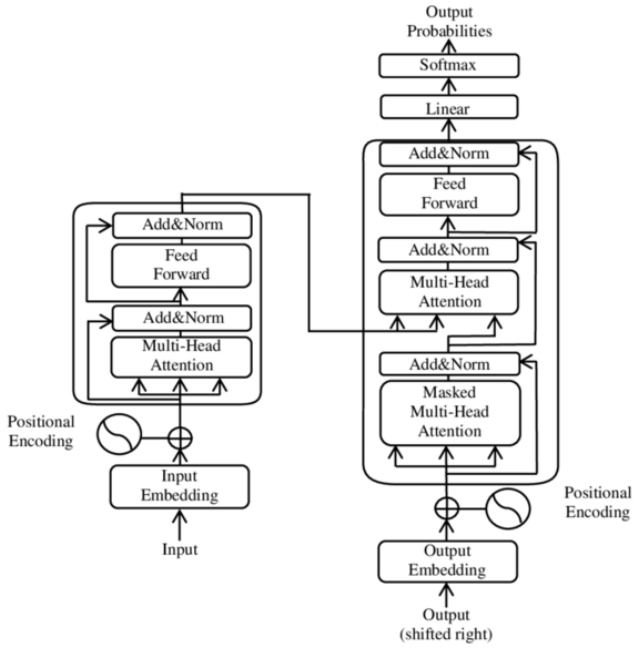
Perkembangan teknologi di bidang kecerdasan buatan terjadi dengan sangat pesat, terutama dalam kurun waktu satu dekade terakhir. Perkembangan kecerdasan buatan dan implementasinya dalam berbagai bidang telah membantu manusia menyelesaikan masalah secara efektif dan pada berbagai kasus, secara lebih efisien. Salah satu cabang kecerdasan buatan yang berkembang sangat pesat dalam beberapa tahun terakhir adalah bidang dengan pendekatan pembelajaran mesin (*machine learning)* dan pengolahan bahasa alami (*natural language processing)* [1].

Perkembangan yang paling berpengaruh terhadap bidang kecerdasan buatan adalah dengan diperkenalkannya arsitektur transformer. Transformer merupakan sebuah arsitektur model kecerdasan buatan yang pertama kali diusulkan pada 2017 [2], untuk menggantikan penggunaan pendekatan sekuensial yang digunakan dalam model-model kecerdasan buatan yang telah banyak dilakukan sebelumnya, antara lain pendekatan dengan menggunakan arsitektur RNN dan LSTM [3]-[5].

Arsitektur transformer sendiri telah berdampak besar dalam bidang pengolahan bahasa alami (*Natural Language Processing*), dimana arsitektur ini secara *de facto* menggantikan pendekatan sekuensial yang telah ditawarkan sebelumnya, dan menggunakan mekanisme parallel dan *self-attention* yang memungkinkan model untuk mengakses input secara langsung. Model ini meningkatan akurasi prediksi, dan efisiensi pelatihan, sehingga membuka pintu bagi implementasi model yang memiliki ukuran lebih besar dan lebih kompleks semisal BERT[6] dan GPT[7]. Dengan perkembangan tersebut, secara praktis arsitektur model kecerdasan buatan berbasis transformer tersebut menjadi dominan dalam perkembangan dan aplikasi penggunaan kecerdasan buatan, khususnya dalam bidang pengolahan bahasa alami [8].

Karena pesatnya perkembangan tersebut, perlu dibuat sebuah tinjauan terhadap penelitian-penelitian yang telah ada dan menyajikannya secara singkat didalam sebuah artikel untuk dapat memberikan gambaran mengenai proses kemajuan yang telah terjadi.

Received March 20, 2023; Revised September 02, 2024; Accepted September 02, 2024



Gambar. 1. Arsitektur Model Transformer [9]

II. METODE PENELITIAN

Tinjauan ini dilakukan dengan melakukan pengumpulan karya tulis yang memiliki kesesuaian topik bahasan dengan tujuan utama artikel ini. Kemudian akan dilakukan perangkuman, dan analisis serta interp

A. Seleksi Artikel Ilmiah

Penelitian ini akan dimulai dengan mengumpulkan beberapa karya tulis yang dianggap signifikan dalam perkembangan kecerdasan buatan, khususnya dalam perkembangan arsitektur berbasis transformer. Karya tulis yang dikumpulkan antara lain berasal dari jurnal, prosiding, ataupun *technical report* yang telah mendapatkan penilaian dari rekan sejawat (*peer-review)*. Artikel dinilai dampaknya berdasarkan dengan kesesuaian bahasan dengan tujuan dari penelitian ini, kemudian jumlah sitasi yang didapatkan, dan *presence* secara umum dari karya tulis tersebut dalam perkembangan teknologi terkait. Dari artikel-artikel ini, akan dipilih lima buah artikel yang dianggap paling mewakili kemajuan di bidang ini dalam lima tahun terakhir.

B. Rangkuman Artikel Ilmiah

Lima buah artikel tersebut kemudian akan dirangkum untuk memberikan gambaran mengenai temuan dan usulan yang dipaparkan dalam masing-masing artikel ilmiah tersebut. Rangkuman akan mencakup tujuan penelitian, metodologi yang digunakan, hasil, implikasi dan kontribusi dari pertemuan tersebut dalam konteks perkembangan teknologi kecerdasan buatan berbasis arsitektur transformer.

C. Analisis dan Interpretasi

Selanjutnya, dari rangkuman tersebut akan dilakukan analisis dan interpretasi terhadap temuan-temuan dan hasil penelitian tersebut, serta untuk melakukan penyelarasan pemahaman terhadap temuan-temuan dari artikel-artikel ilmiah tersebut.

III. HASIL

Terhadap artikel-artikel ilmiah yang beredar, telah dilakukan pengumpulan, kemudian dilakukan seleksi berdasarkan imbasnya terhadap perkembangan teknologi kecerdasan buatan berbasis arsitektur transformer. Dari proses seleksi ini didapatkan tujuh buah referensi yang digunjukkan pada Table 1. Namun, tinjauan ini hanya akan membahas lima referensi saja.

A. Penelitian oleh Vaswani

Penelitian [2] merupakan artikel yang pertama kali mengusulkan penggunaan transformer, untuk menggantikan model-model sekuensial berbasis RNN dan LSTM. Penelitian ini mengusulkan penggunaan mekanisme atensi *self-attention*. Arsitektur ini juga menawarkan suatu metode baru untuk melakukan pengolahan dan memahami data sekuensial dalam ranah NLP.

Tabel 1 Artikel Ilmiah Tentang Arsitektur Transformer

No.	Referensi	Dampak
1.	A. Vaswani et al., "Attention is all you need," in Proc. Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, Dec. 2017, pp. 5998-6008.	Pengusul arsitektur transformer
2.	J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019), Minneapolis, MN, USA, Jun. 2019, pp. 4171-4186.	Inovasi dalam bentuk model BERT
3.	T. B. Brown et al., "Language models are few-shot learners," arXiv:2005.14165, Nov. 2020.	Pemahaman dan menghasilkan teks.
4.	Z. Dai, Z. Yang, Y. Yang, W. W. Cohen, J. Carbonell, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," arXiv:1901.02860, Jan. 2019.	Model transformer XL yang efisien
5.	Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, Jul. 2019.	Model RoBERTa yang telah dioptimasi
6.	A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, Jun. 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf	Dasar model GPT 2
7.	L. Dong et al., "Unified Language Model Pre-training for Natural Language Understanding and Generation," in Proc. Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Vancouver, Canada, Dec. 2020, pp. 13042-13054.	Model UniLM

Penelitian tersebut mencoba mengatasi keterbatasan model sekuensial seperti RNN dan LSTM, dengan menawarkan model arsitektur transformer. Model sekuensial semisal RNN dan LSTM memiliki keterbatasan dari pengolahannya yang sulit untuk dilakukan secara parallel, yang mana hal ini berpengaruh terhadap efisiensi waktu pengolahannya. Transformer juga menggunakan mekanisme *self-attention* untuk memperhitungkan ketergantungan antara kata dalam sekuens, sekaligus meningkatkan kecepatan pelatihan dan inferensi dari model.

Mekanisme *self-attention* ini diintegrasikan kedalam sebuah lapisan *Multi-Head Attention*, yang memungkinkan model untuk mempelajari representasi kata yang berbeda pada tingkat abstraksi yang berbeda. Hasil dari penggunaan ini memungkinkan model transformer untuk mengurangi jumlah parameter, sembari meningkatkan kualitas representasinya.

Selain itu, model arsitektur transformer juga menunjukkan potensi yang lebih baik dalam mengatasi masalah *vanishing gradient* serta potensi dalam mempelajari struktur hierarki dari teks yang lebih kompleks. Evaluasi yang dilakukan oleh peneliti juga menunjukkan bahwa arsitektur transformer lebih unggul dibandingkan RNN dan LSTM utamanya dalam proses pelatihan dan inferensi. Beberapa data empirik dari penelitian ini seperti pencapaian nilai BLEU 28.4 untuk tugas penerjemahan WMT 2014 English-To-German, dan nilai BLEU 41.0 untuk tugas penerjemahan WMT 2014 English-to-French.

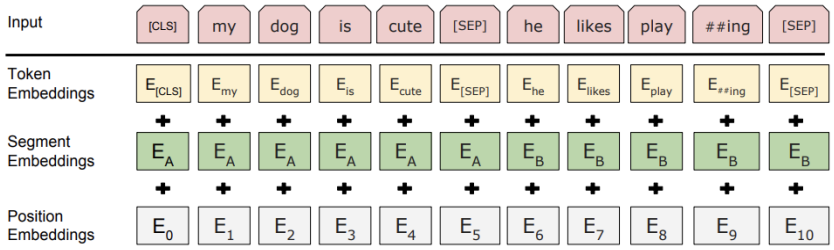
Sejak diusulkan, telah berkembang banyak model-model berbasis transformer utamanya dalam bidang NLP antara lain BERT, GPT-3, dan RoBERTa. Model transformer ini pun telah berkembang ke berbagai bidang termasuk pengolahan citra.

B. Penelitian oleh Devlin

Penelitian [6] mengusulkan model BERT (*Bidirectional Encoder Respresentations from Transformers)* yang merupakan suatu penelitian dengan kontribusi besar di bidang NLP berbasis Transformer. BERT dibuat dalam dua versi dengan jumlah 110 juta dan 340 juta parameter. Pendekatan ini memanfaatkan penggunaan *pre-training* untuk mempelajari representasi kata secara lebih baik dan kaya secara konteks,

untuk menghasilkan model yang lebih baik dalam pengolahan bahasa alami, ditunjukkan pada Fig.2.

Analisis dan interpretasi terhadap penelitian ini menunjukkan bahwa penggunaan *pre-training* dan arsitektur transformer yang bersifat dua arah adalah sinergi yang amat penting untuk model BERT ini. Pelatihan awal secara *unsupervised* untuk prediksi kata hilang dilakukan terlebih dahulu sebelum dilanjutkan untuk pelatihan yang lebih spesifik. Selanjutnya model BERT memanfaatkan arsitektur transformer dua arah untuk mempelajari representasi kata berdasarkan konteks dari kedua arah (kanan dan kiri) pada sebuah sekuen teks. Hal ini sangat penting dalam NLP karena seringkali konteks dari suatu



Gambar. 2. Representasi Input pada BERT [6]

kata dipengaruhi oleh kata-kata yang berada pada sebelah kiri maupun kanan kata-kata awal. Dari sisi empiriknya, BERT mampu mencapai nilai 80.5 untuk benchmark GLUE, dan nilai F1 untuk SQuAD dan SQuAD v2.0 berturut-turut 93.2, dan 83.1.

Selain itu, *pre-training* dalam model ini mengurangi maslaah overfitting dan meningkatkan pemahaman yang lebih baik terhadap bahasa secara umum, sebelum disesuaikan dengan tugas yang spesifik. Model yang dihasilkan menjadi lebih fleksibel dan efisien untuk berbagai macam permasalah NLP. Evaluasi empirik oleh penulis juga menunjukkan keunggulan model BERT dibandingkan model-model lain semisal GLUE, SQuAD, dan SWAG. Model ini banyak digunakan untuk permasalahan NLP terkait klasifikasi teks hingga analisis sentiment. Model ini juga menjadi dasar dari model lanjutan seperti RoBERTa.

C. Penelitian oleh Brown

Dalam penelitian [10] berjudul "Language Models are Few-Shot Learners" oleh Brown et al., penulis memaparkan model Generative Pre-trained Transformer 3 (GPT-3), model kecerdasan buatan berskala besar yang dirancang untuk melakukan tugas umum dalam ranah pemrosesan bahasa alami (NLP) dengan hanya menggunakan sedikit data pelatihan. GPT-3 diimplementasikan menggunakan arsitektur transformer dan teknik pre-training pada skala yang lebih besar dibandingkan penelitian-penelitian sebelumnya, yang mana skala ini meliputi penggunaan 175 miliar parameter, untuk meningkatkan kemampuan model dalam memahami teks.

Penelitian tersebut menitikberatkan potensi GPT-3 sebagai *few-shot learner*, yaitu bahwa model ini dapat dilatih ulang untuk suatu permasalah yang lebih spesifik dengan menggunakan data pelatihan yang relatif sedikit. Hal ini mengindikasikan potensi keunggulan dari model ini dilihat dari segi biaya dan waktu pelatihan model.

GPT-3 memanfaatkan teknik pre-training skala besar, dengan mempelajar teks dalam jumlah besar untuk meningkatkan kemampuannya memahami representasi teks. Berkat pelatihan pada berbagai jenis teks, GPT-3 memiliki pemahaman yang lebih baik mengenai struktur bahasa, topik, dan nuansa. Untuk menyesuaikan model terhadap suatu masalah baru dengan hanya menggunakan sedikit pelatihan baru.

Evaluasi empiris yang dilakukan dalam artikel menunjukkan bahwa GPT-3 memiliki kinerja yang baik dalam berbagai tugas NLP, seperti klasifikasi teks, pertanyaan dan jawaban, serta generasi teks, seringkali hanya dengan sedikit atau tanpa data pelatihan tambahan. GPT memiliki nilai *benchmark* LAMBADA 66.5 untuk Perplexity, skor SuperGLUE 71.8, SQuAD dengan nilai F1 82.6. Namun, peneliti juga mencatat bahwa GPT-3 memiliki beberapa keterbatasan, seperti kecenderungan untuk menghasilkan jawaban yang panjang dan tidak konsisten atau kurangnya pemahaman tentang fakta-fakta spesifik.

D. Penelitian oleh Dai

Dalam penelitian [11], berjudul "Transformer-XL: Attentive Language Models Beyond a Fixed-Length

Context" oleh Dai et al., penulis mengusulkan model Transformer-XL yang dirancang untuk mengatasi keterbatasan arsitektur transformer dalam menangani konteks berpanjang variabel. Transformer-XL menggabungkan mekanisme rekurensi dan pengindeksan relatif posisi untuk meningkatkan efisiensi dan fleksibilitas model.

Transformer-XL merupakan pengembangan penting dalam bidang pemrosesan bahasa alami, khususnya dalam mengatasi masalah konteks panjang yang tidak efisien pada model transformer sebelumnya. Dengan memperkenalkan segment rekurensi dan pengindeksan relatif posisi, Transformer-XL mampu mengurangi kompleksitas waktu dan memori secara signifikan, serta meningkatkan kualitas pemahaman konteks panjang.

Fakta-fakta empirik dalam penelitian ini mencakup:

- Waktu Pelatihan: Transformer-XL dilaporkan 1.800 kali lebih cepat daripada model Transformer standar dalam hal waktu pelatihan.
- Performa pada Tugas Pemodelan Bahasa: Transformer-XL mencapai skor perplexity sebesar 20,5 pada dataset WikiText-103 dan 18,3 pada dataset One Billion Word. Kedua skor ini lebih baik daripada model Transformer standar dan model LSTM yang telah ada sebelumnya.
- Skor pada Tugas Pemahaman Teks: Transformer-XL mencapai skor state-of-the-art pada tugas LAMBADA, dengan skor akurasi sebesar 86,6, mengungguli model Transformer standar dan LSTM.
- Skalabilitas: Transformer-XL berhasil menunjukkan peningkatan performa yang signifikan pada tugas yang melibatkan konteks panjang, seperti tugas pemahaman teks LAMBADA.
- Jumlah Layer dan Parameter: Transformer-XL dibangun dengan berbagai konfigurasi, termasuk model dengan 12-layer dan 6-layer, serta jumlah parameter yang bervariasi.

Artikel ini menunjukkan bahwa Transformer-XL merupakan langkah maju penting dalam mengatasi keterbatasan model transformer sebelumnya, dengan potensi penggunaannya dalam berbagai aplikasi pemrosesan bahasa alami yang melibatkan konteks panjang.

E. Penelitian oleh Liu

Dalam penelitian [12] berjudul "RoBERTa: A Robustly Optimized BERT Pretraining Approach" oleh Liu et al., penulis mengusulkan pendekatan pelatihan ulang (pretraining) yang dioptimalkan untuk model BERT, yang disebut RoBERTa. Pendekatan ini didasarkan pada pengamatan bahwa kinerja BERT sangat dipengaruhi oleh metode pelatihan dan hiperparameter yang digunakan.

Penulis menganggap RoBERTa sebagai suatu kontribusi yang penting untuk optimasi kinerja arsitektur BERT. Penulis mempertimbangkan berbagai faktor antara lain ukuran dataset, jumlah iterasi pelatihan, serta hiperparameter model, untuk mencoba menemukan suatu kombinasi optimal dari faktor-faktor tersebut. Secara umum, RoBERTa diklaim mampu meningkatkan kinerja BERT pada berbagai tugas NLP. Peningkatan kinerja ini digarisbawahi dalam beberapa fakta empirik penelitian ini. Capaian RoBERTa dalam beberapa *benchmark* untuk NLP adalah sebagai berikut:

- GLUE (General Language Understanding Evaluation) benchmark: RoBERTa-Large mencapai skor rata-rata 88.5, sementara RoBERTa-Base mencapai skor 86.6.
- SQuAD (Stanford Question Answering Dataset) v1.1: RoBERTa-Large mencapai skor F1 sebesar 94.6, mengungguli BERT-Large.
- SQuAD v2.0: RoBERTa-Large mencapai skor F1 sebesar 89.4 dan skor EM (Exact Match) sebesar 86.5.

Penelitian ini secara umum menggarisbawahiu bahwa optimalisasi terhadap proses pelatihan dan hyperparameter terhadap RoBERTa mampu meningkatkan kinerja BERT secara signifikan pada berbagai tugas NLP, dan memiliki peranan sangat penting dalam pengembangan model-model untuk permasalahan NLP.

IV. KESIMPULAN

Dari lima referensi yang telah dianalisis, dapat kami simpulkan bahwa perkembangan kecerdasan buatan khususnya yang berbasis arsitektur transformer telah terjadi dengan sangat pesat terutama pada bidang NLP. Perkembangan ini dimulai oleh usulan dari Vaswani et al. [2], yang kemudian dilanjutkan dengan berbagai variasinya semisal BERT [6], GPT-3 [10], Transformer-XL [11], dan RoBERTa [12]. Perkembangan ini diikuti dengan peningkatan kinerja yang diukur secara empirik berdasarkan beberapa

benchmark NLP, dan menjadikannya sebagai suatu model yang merupakan *state-of-the-art* di bidang NLP.

V. SARAN

Dari tinjauan dan kesimpulan yang ditarik terhadap hasil penelitian tersebut, penulis mengemukanan beberapa saran terkait potensi pengembangan arsitektur berbasis transformer, khususnya dalam bidang NLP.

- Perlu dilakukan penelitian lebih lanjut untuk mengembangkan teknik-teknik pre-training, dan *fine-tuning* untuk meningkatkan efisiensi dan efektivitas pelatihan model, serta mengurangi konsumsi sumber daya yang diperlukan baik secara konsumsi maupun kapasitas pada saat pelatihan model.
- Dapat dilakukan penelitian lebih lanjut untuk mencari arsitektur transformer yang lebih hemat sumber daya, untuk dapat diletakkan pada perangkat-perangkat dengan kemampuan komputasi terbatas namun dengan cakupan besar, semisal untuk perangkat-perangkat IoT.
- Pengembangan suatu metoda untuk dapat menginterpretasi dan menjelaskan cara kerja dari model transformer, untuk memberikan gambaran mengenai alasan dari model transformer membuat suatu prediksi, dan meningkatkan kepercayaan pengguna untuk penggunaan pada aplikasi-aplikasi kritis.
- Penggunaan transformer untuk bidang-bidang di luar NLP, semisal pengolahan citra, pengenalan audio, dan masalah-masalah lainnya.
- Penelitian mengenai sisi keamanan, etika, dan privasi penggunaan dan pengembangan model transformer, untuk memastikan bahwa teknologi tersebut memberikan manfaat yang positif dan meminimalisir dampak negatif.

VI. REFERENSI

[1] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2020.

[2] A. Vaswani et al., "Attention is all you need," in Proc. Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, Dec. 2017, pp. 5998-6008.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.

[4] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," Neural Computation, vol. 12, no. 10, pp. 2451-2471, Oct. 2000.

[5] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), Vancouver, BC, Canada, May 2013, pp. 6645-6649.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019), Minneapolis, MN, USA, Jun. 2019, pp. 4171-4186.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, Jun. 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[8] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv:1609.08144, Sep. 2016.

[9] Wikipedia. (2022, March 16). The Transformer Model Architecture. [Online]. Available: https://commons.wikimedia.org/wiki/File:The-Transformer-model-architecture.png

[10] T. B. Brown et al., "Language models are few-shot learners," arXiv:2005.14165, Nov. 2020.

[11] Z. Dai, Z. Yang, Y. Yang, W. W. Cohen, J. Carbonell, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," arXiv:1901.02860, Jan. 2019.

[12] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, Jul. 2019.