**SIMPLE IMPLEMENTATION OF T5 TRANSFORMER MODEL FOR INDONESIAN NEWS SUMMARY**

**Kisana Adzan Sitorus**

Corresponding author’s Email: [kisanaadzansitorus@gmail.com](mailto:kisanaadzansitorus@gmail.com)

**Executive Summary:** This report presents a comprehensive analysis of a text summarization system built using the T5 (Text-to-Text Transfer Transformer) model. The system is designed to generate concise summaries from longer input texts, with a particular focus on Indonesian language news articles. The implementation uses PyTorch and the Transformers library, with a structured approach to data processing, model training, evaluation, and inference.

The system demonstrates reasonable performance on the test dataset, achieving a ROUGE-1 F1 score of 0.5476, ROUGE-2 F1 score of 0.4390, ROUGE-L F1 score of 0.4762, and a BLEU score of 0.2667. The training process included early stopping to prevent overfitting, which was triggered after 4 epochs when no improvement in validation loss was observed.

1. **System Architecture**
2. **Technologies and Libraries**

The system utilizes the following key technologies and libraries:

* PyTorch: Core deep learning framework
* **Transformers**: Hugging Face library providing pre-trained transformer models
* **T5 Model**: Text-to-Text Transfer Transformer for conditional generation tasks
* **NLTK**: Natural Language Toolkit for text preprocessing and evaluation
* **Rouge Score**: For evaluating summary quality
* **BLEU Score**: Additional metric for summary evaluation
* **Pandas**: For data manipulation and management
* **Matplotlib/Seaborn**: For visualization of training progress and results

1. **System Components**

The system is organized into several components:

* **Configuration Module**: Centralized parameter management through the Config class
* **Data Processing Pipeline**: Functions for loading, preprocessing, and preparing data
* **Dataset and DataLoader Classes**: Custom implementation for handling text summarization data
* **Training Engine**: Functions for model training, validation, and testing
* **Evaluation System**: Metrics calculation and model performance assessment
* **Inference Module**: For generating summaries from new texts
* **Visualization Tools**: For analyzing model performance and results

1. **Data Processing**
2. **Data Loading and Preparation**

The system loads data from CSV files containing the following columns:

* source: Source of the article
* source\_url: URL of the original article
* summary: Reference summary
* paragraphs: Full text content to be summarized

The datasets are structured as:

* Training set: 1 record with 4 columns
* Validation set: 1 record with 4 columns
* Test set: 1 record with 4 columns

This suggests a very small dataset was used, likely for demonstration purposes.

1. **Text Preprocessing**

The preprocessing pipeline includes:

* Converting text to lowercase
* Replacing newlines and tabs with spaces
* Removing URLs and HTML tags
* Normalizing whitespace

Statistical analysis of the data revealed:

* Average input length (training): 1968 characters
* Average summary length (training): 377 characters
* Average input length (validation): 1890 characters
* Average summary length (validation): 492 characters
* Average input length (test): 2645 characters
* Average summary length (test): 398 characters

1. **Dataset Implementation**

A custom NewsSummaryDataset class handles:

* Adding the prefix "summarize: " to input texts for T5
* Tokenization of inputs with padding and truncation
* Tokenization of target summaries
* Special handling of padding tokens in target sequences by replacing them with -100 for loss calculation

1. **Model Architecture**
2. **Base Model**

The system uses the pre-trained t5-base model from Hugging Face's Transformers library, which has:

* 220 million parameters
* 12 encoder and decoder layers
* 12 attention heads
* Hidden size of 768

1. **Configuration Parameters**

Key Configuration parameters include:

* Maximum input length: 512 token
* Maximum output length: 150 tokens
* Training batch size: 4
* Gradient accumation steps: 2
* Number ofepoch: 5
* Learning rate: 3e-4
* Weight decay: 0.01
* Warmup ratio: 0.1
* Maximum gradient norm: 1.0
* Early stopping patience: 3 epochs

1. **Training Process**
2. **Training Strayegy**

The training process included:

* Optimization using AdamW optimizer
* Linear learning rate schedule with warmup
* Gradient accumulation to simulate larger batch sizes
* Gradient clipping to avoid exploding gradients
* Early stopping to prevent overfitting

1. **Training Progress**

The training ran for 4 epochs before early stopping was triggered:

| **Epoch** | **Training Loss** | **Validation Loss** | **ROUGE-1 F1** | **ROUGE-2 F1** | **ROUGE-L F1** | **BLEU** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 5.4500 | 4.9421 | 0.8125 | 0.7419 | 0.8125 | 0.5903 |
| 2 | 5.5589 | 4.9421 | 0.8125 | 0.7419 | 0.8125 | 0.5903 |
| 3 | 5.4818 | 4.9421 | 0.8125 | 0.7419 | 0.8125 | 0.5903 |
| 4 | 5.4302 | 4.9421 | 0.8125 | 0.7419 | 0.8125 | 0.5903 |

**Observations**:

* The validation loss remained constant across all epochs
* Similarly, all evaluation metrics remained unchanged across epochs
* This could indicate that the model reached its optimal performance on the first epoch, or that the validation set is too small to show meaningful variations

1. **Evaluation Results**
2. **Test Set Performance**

The final model, selected based on the best validation performance (epoch 1), achieved the following results on the test set:

* Loss: 5.5267
* ROUGE-1 F1: 0.5476
* ROUGE-2 F1: 0.4390
* ROUGE-L F1: 0.4762
* BLEU: 0.2667

1. **Example Analysis**

Example of generated summary:

**Original Text (truncated):**

jakarta, cnn indonesia - - dilansir afp, seorang warga mesir yang dipercaya sebagai wanita terberat di dunia masuk sebuah rumah sakit di abu dhabi kamis (5/5), dimana ia akan melanjutkan perawatan setelah menjalankan operasi menurunkan badan di india. eman ahmed abd el aty memiliki berat bad...

**Original Summary:**

eman ahmed abd el aty memiliki berat badan mencapai 500 kilogram sebelum menjalankan operasi di mumbai maret lalu dimana ia mengurangi seperlima dari berat badannya. abd el aty diberi diet cairan khusus selama berada di india yang bertujuan menurunkan berat badan. kini, berat badannya telah turun drastis sebanyak 323 kilogram dalam tiga bulan. sekarang berat badannya tinggal 176,6 kilogram.

**Generated Summary**:

ahmed abd el aty diberi diet cairan khusus selama berada di india yang menghadapi masalah obesitas yang kini sebanyak 323 kilogram dalam tiga bulan. sekarang berat badannya tinggal 176,6 kilogram.

**Analysis:**

* The model successfully captured key information about the weight loss and final weight
* Some details about the initial weight (500 kg) and the surgery in Mumbai were omitted
* The generated summary is more concise (26 words) than the reference summary (45 words)
* The summary maintains factual accuracy while focusing on the most important details

1. **Sample Inference**

For the sample text about Bank Mandiri's financial results:

**Input:**

PT Bank Mandiri (Persero) Tbk. mencatat kenaikan laba bersih sebesar 45,2 persen secara tahunan pada kuartal pertama 2023 yaitu menjadi Rp 12,5 triliun. Direktur Utama Bank Mandiri, Darmawan Junaidi, mengatakan pertumbuhan laba tersebut didorong oleh pertumbuhan kredit sebesar 9,5 persen secara tahunan menjadi Rp 1.216,3 triliun. Selain itu, juga didorong oleh kenaikan dana pihak ketiga sebesar 6,2 persen secara tahunan menjadi Rp 1.162,6 triliun. 'Pertumbuhan kredit ini sejalan dengan pembiayaan Bank Mandiri yang tidak hanya fokus pada pemulihan ekonomi, tetapi juga mendukung pembangunan infrastruktur nasional,' kata Darmawan dalam keterangan tertulis yang dikutip Selasa (25/4/2023).

**Generated Summary:**

pt bank mandiri (persero) tbk. mencatat kenaikan laba bersih sebesar 45,2 persen secara tahunan pada kuartal pertama 2023 yaitu menjadi rp 12,5 triliun.

**Analysis:**

* The model extracted the main headline about the profit increase
* The summary omits supporting details about credit growth and third-party funds
* The generated summary is extremely concise, focusing only on the most critical information
* This demonstrates the model's ability to identify the most salient information

1. **System Limitations and Challenges**
2. **Dataset Limitations**

* The extremely small dataset size (only 1 record in each split) is a major limitation
* Such a small dataset is insufficient for proper training and evaluation
* The model is likely memorizing the examples rather than learning generalizable patterns
* This explains the static validation metrics across epochs

1. **Model Constraints**

* The maximum input length of 512 tokens may truncate longer articles
* The T5-base model has 220M parameters, which is moderate compared to larger language models
* Computation seems to be performed on CPU, which significantly slows down training

1. **Evaluation Challenges**

* The small test set doesn't provide a robust evaluation of the model's capabilities
* The metrics show moderate performance, but these cannot be considered representative due to the dataset size

1. **Recommendations for Improvement**
2. **Data Enhancement**

* **Increase Dataset Size**: Collect and annotate a larger corpus of Indonesian news articles with summaries
* **Data Augmentation**: Use techniques like paraphrasing, translation and back-translation to expand the dataset
* **Cross-Domain Variety**: Include texts from different domains to improve generalization

1. **Model Improvements**

* **Fine-tuning Strategy**: Implement progressive unfreezing of layers during fine-tuning
* **Model Size Exploration**: Experiment with different T5 variants (small, large) to find the optimal size
* **Hyperparameter Optimization**: Use grid search or Bayesian optimization to find optimal hyperparameters
* **Accelerated Training**: Implement GPU-based training to speed up the process

1. **Evaluation Enhancement**

* **Human Evaluation**: Supplement automatic metrics with human judgment of summary quality
* **Diverse Test Set**: Create a larger and more diverse test set for robust evaluation
* **Additional Metrics**: Consider factual consistency and information coverage metrics

1. **Conclusion**

The text summarization system demonstrates a well-structured approach to implementing a transformer-based summarization pipeline. The code is organized, includes comprehensive logging, and implements best practices like early stopping and gradient clipping.

However, the system's practical utility is severely limited by the extremely small dataset. The consistency in validation metrics across epochs suggests the model is not truly learning to generalize but is instead memorizing the few examples it has seen.

To develop a production-ready text summarization system, particularly for Indonesian language content, a substantially larger dataset and more rigorous evaluation would be required. The current implementation provides a solid foundation and architecture that could be scaled up with proper data resources.