Towards Personalized AI Personas: Leveraging Transfer Learning and Self-Supervised Techniques

Elvis Kimara, Jian Sun, Kunle Oguntoye

February 8, 2024

Abstract

In this paper, we propose a novel approach to develop personalized AI personas, distinct from traditional personal assistants, by leveraging transfer learning and self-supervised techniques. Our methodology involves building an iOS application for data collection, integrating with backend infrastructure for model training, and periodic refinement. The key novelty lies in the acquisition and utilization of personalized datasets, along with innovative methods for model training and evaluation. We discuss the technical details, challenges, and future directions of this research, aiming to advance the field of human-AI interaction.

1 Introduction

Recent advancements in artificial intelligence have led to the proliferation of virtual assistants capable of general tasks. However, the concept of personalized AI personas, tailored to mimic an individual's characteristics, remains relatively unexplored. Our research seeks to bridge this gap by developing a framework for creating AI replicas of individuals, capable of engaging in conversations and tasks reflective of their unique personalities and preferences.

2 Related Work

While traditional personal assistants rely on predefined responses and rules, recent efforts have explored the development of more personalized AI agents. For instance, ChatGPT has demonstrated the capability to generate context-aware responses based on large-scale language models. However, our approach differs in its focus on individual-level personalization, leveraging transfer learning techniques to adapt models to specific users.

3 Methodology

3.1 Data Collection

We propose the development of an iOS application equipped with voice-to-text functionality to capture users' natural speech patterns and conversations. This data will be anonymized and securely transmitted to a backend server database for storage and processing.

3.2 Model Training

The backend infrastructure will host a pre-trained BERT model, serving as the baseline for further training. Utilizing the collected data, we aim to fine-tune the BERT model through transfer learning techniques, adapting it to the user's specific linguistic style and preferences.

Figure 1: Model Training Pipeline

3.3 Prompt-Based Interaction

To enhance user engagement, the AI persona will offer prompt-based interaction capabilities, similar to ChatGPT. These prompts will be derived from the user's historical data, ensuring relevance and accuracy in responses.

3.4 Model Refinement

Periodic retraining sessions will be conducted to incorporate new data and refine the AI persona's understanding of the user's persona. Self-supervised and reinforcement learning techniques may be explored to further improve model performance.

4 Testing and Evaluation

4.1 In-Sample Evaluation

The AI model's performance will be evaluated through "in-sample evaluation," where it is tested on ground truth data. This entails assessing the model's ability to accurately respond to queries or prompts based on the user's historical interactions.

4.2 Hallucination Detection

Mechanisms will be implemented to detect instances where the AI persona generates responses inconsistent with the user's persona or historical interactions, ensuring the fidelity of the model.

4.3 Prompt Accuracy

Users will be informed in advance of the prompts the model can answer with high accuracy, fostering transparency and trust in the AI persona's capabilities.

5 Future Directions

5.1 Individualized Latent Space Representation

Future research will focus on developing fair and accurate representations of an individual's mind, encompassing emotions, feelings, and nuanced aspects of human experience.

5.2 Emotion Recognition and Response

Integration of emotion recognition algorithms will enable the AI persona to respond empathetically and adaptively to the user's emotional state, enhancing the quality of interactions.

5.3 Application Expansion

Beyond personal avatars, the technology could be extended to applications such as educational assistants, content curation algorithms, and personalized recommendation systems, opening new avenues for exploration.

6 Considerations

6.1 Model Narrowing

To ensure model accuracy and relevance, we have narrowed our focus to evaluating performance based on previously seen training data. This approach minimizes the need for extensive generalization and allows for prompt-based interaction tailored to the user's specific interests and preferences. **This decision was made to accommodate the project timeline of one semester. All the considerations were given by Professor Alexander Stoytchev.**

6.2 Capture Emotion and give Audio Output

In response to considerations from experts, our application will not only capture text but also emotions and other data points in voice-to-text conversions. Additionally, we plan to implement audio output functionality to provide a more immersive and interactive user experience.

7 Initial Dataset

Our initial dataset comprises a diverse range of texts, including Shakespearean works, scripts from popular TV shows like SpongeBob SquarePants, and personalized data collected during the application usage.

8 Formulas

8.1 Transformer Model

The attention mechanism in the transformer model is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where:

- Q is the query matrix,
- K is the key matrix,
- V is the value matrix, and
- d_k is the dimensionality of the key vectors.

8.2 Autoencoder-Decoder

The objective function for training an autoencoder-decoder is given by:

$$\mathcal{L}(x, \hat{x}) = \sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \qquad (2)$$

where:

- x is the input data, $\hat{x}$ is the output of the autoencoder, and, N is the number of data points.

9 References

References

[1] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

[2] 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.