Research Topic

Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals Based on LLM

Background and Key Scientific Questions of The Proposed Research

In recent years, computer-mediated systems have increasingly facilitated verbal communication, which is often the primary mode of communication. Platforms like Google Meet, Zoom, and Microsoft Teams have been widely adopted and provide capabilities such as live captioning and noise cancellation to facilitate conversations.

While such tools help people better understand each other, we envision that visual augmentations leveraging the semantics of spoken language could also be helpful, especially for conveying complex, nuanced, and unfamiliar information. People are already using visual aids to provide additional context and clarification in daily conversations. For example, when discussing a recent trip, people may use photos from their album to help their audience follow along. Similarly, when describing a new restaurant to a friend, one might say, "it's a small place with a lot of character," and then search for a picture online to show what the place looks like. Research has shown that people learn more effectively from videos than from audio-only versions of the same material, and prefer podcasts and stories with visuals over those without. The Multi-modal Phenomena and the Principle of Inverse Effectiveness have also proven that the human sensory system has a super additive effect when responding to stimuli from multiple, simultaneous modalities. As the adage goes, "a picture is worth a thousand words."

Previous research has proposed various automated text-to-visual systems that can automatically transform audio-only content into audiovisual content and create realistic images and art from natural language. However, augmenting synchronous human-human verbal communication with visuals presents unique challenges that have not yet been addressed by these systems.

First, the input is a continuous stream of conversation, not a discrete textual description of the visual, requiring the system to go beyond keyword-based approaches, like named entity detection, to understand the implicit intent of what people want to show in the context. **Second**, when people are actively engaged in conversation, they have limited cognitive resources to interact with AI prompts and results, necessitating interactions to be as subtle and minimal as possible to avoid disrupting the conversation. **Third**, without a real-time system deployed, it is difficult to study how people could interact with and benefit from visuals in real conversations, and how such systems would impact their communication.

Related Work

1. **Visual Enhancement Communication Systems**: Advanced technologies, such as real-time captioning systems, already provide textual descriptions during user interactions. These systems use advanced speech recognition technology to convert spoken words into text, which is then displayed in real time, aiding individuals with hearing impairments or those in noisy environments to better understand the conversation.
2. **Text-to-Image Conversion**: Models like OpenAI's DALL-E and CLIP have shown the ability to generate images directly from natural language descriptions. These models understand complex textual descriptions and generate high-quality images that match the content, significantly contributing to multi-modal learning and creative arts generation.
3. **Applications of Augmented Reality Technology**: AR technology has been applied in various scenarios, including medical, educational, and entertainment fields. In the medical field, AR can assist in complex surgical guidance by overlaying 3D anatomical images in the surgeon's field of view. In education, AR technology can virtualize textbook content, enhancing student engagement and understanding through interactive learning.
4. **Multi-modal Interaction Systems**: Recent research highlights the potential of multi-modal interaction systems to enhance user experience by integrating tactile, visual, and auditory inputs. These systems demonstrate how simultaneous transmission of information through multiple sensory channels can enhance information absorption and understanding.
5. **Real-time Captioning and Translation Systems**: Video conferencing platforms (e.g., Zoom, Microsoft Teams) have implemented real-time captioning and translation features to facilitate multilingual conversations. By using speech recognition technology, these systems convert spoken language into text and provide translations in real-time.
6. **Mixed Reality (MR) and Virtual Reality (VR) Technologies**: MR and VR technologies are gradually merging to create more immersive user experiences. These technologies are used not only for entertainment but also for virtual meetings and remote collaboration, enhancing interactive experiences in virtual environments.

Research Objective

The goal of this project is to improve existing technology and propose a real-time integrated image and text display HCI system. By generating relevant images and texts during user dialogues, the system aims to enhance information understanding and communication efficiency. Specific objectives include:

1. Developing a model capable of predicting visual intents in real-time.
2. Proposing a user interface that supports synchronized display and interaction of images and text.
3. Validating the system's effectiveness and user satisfaction across various scenarios through user experiments.

Research Methods to Accomplish the Proposed Project

1. **System Improvement**:
    - **Algorithm Improvement**: Based on existing visual intent prediction models, propose algorithms that can generate images and text simultaneously. Utilize large language models (such as GPT-3) for text generation and combine them with image search and generation technologies to provide relevant image and text content in real-time.
    - **Interface Improvement**: Design a user interface that allows users to use the system in video conferencing and online education scenarios. The interface should support the synchronized display of images and text and provide user interaction functions such as zooming, moving, and deleting displayed content.
2. **User Experiments**:
    - **Experiment Design**: Select participants from different backgrounds and set up various experimental scenarios (e.g., education, business meetings, social chatting) to evaluate the system's performance and user experience. Experiments will include both system-supported and non-system-supported conditions.
    - **Data Collection**: Collect subjective evaluations from participants through questionnaires and interviews, and record objective data such as the number of operations, time, and error rates during system use.
3. **Data Analysis**:
    - **Quantitative Analysis**: Statistically analyze experimental data, comparing communication efficiency, information understanding, and user satisfaction between system-supported and non-system-supported conditions.
    - **Qualitative Analysis**: Analyze user feedback, summarize the advantages and shortcomings of the system, and propose improvement suggestions.

References

1. Liu, X., et al. (2023). In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.
2. Ramesh, A., et al. (2021). DALL·E: Creating Images from Text. OpenAI.
3. Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the International Conference on Machine Learning (ICML).
4. Azuma, R. T. (1997). A Survey of Augmented Reality. Presence: Teleoperators and Virtual Environments.
5. Kim, J., & Ganapathi, V. (2020). Real-time Captions and Translation in Video Conferencing. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).