ANTALYA BİLİM ÜNİVERSİTESİ

2024-2025 SPRING SEMESTER

Introduction to Robotics

-EE422/CS421-

VISION LANGUAGE MODELS FOR ROBOTIC PERCEPTION

WEEK 5

AGAH KUTAY FASTING

EMIN HAN CELIK

TIMUR GULMEZ

MELİH EGREK

Introduction

Mel-Frequency Cepstral Coefficients (MFCCs) are one of the most widely used features in speech and audio processing. They provide a compact representation of the short-term power spectrum of sound, mimicking the human ear's perception of frequency. When visualized over time—typically as a cepstrogram—they reveal patterns that can be analyzed to infer the physical and articulatory characteristics of speech. This report explores how the temporal dynamics of MFCCs correspond to physical phenomena in speech production.

Understanding MFCCs

MFCCs are extracted by applying a Fourier Transform to short overlapping frames of a signal, filtering them through a Mel-scale filterbank, taking the logarithm of the energies, and finally applying a Discrete Cosine Transform (DCT). The resulting coefficients capture the spectral envelope of the audio in a reduced-dimensional form. The first few coefficients encode general spectral shape, while higher ones relate to fine spectral details.

Time-Varying Behavior and Physical Interpretation

When MFCCs are plotted over time, they form a two-dimensional representation where one axis represents time and the other the MFCC index. Changes in the MFCC pattern over time provide insights into the physical properties of speech. For instance:

Stable MFCC patterns over several frames often correspond to a sustained vowel sound, indicating minimal change in the vocal tract.

Abrupt shifts in MFCCs typically mark transitions between phonemes, such as between a vowel and a stop consonant, which involves significant articulator movement (e.g., tongue or lips).

Rapid oscillations or irregularities can indicate fricatives or noisy speech components, where airflow is turbulent.

Low MFCC index variation often relates to changes in pitch and overall energy, while higher-index variation reflects subtle changes in resonance and timbre.

These observations can be mapped to physical actions in the vocal tract, such as jaw movement, tongue position, and voicing behavior.

Application Scenarios

MFCC dynamics have been used to:

Distinguish phonemes based on how spectral features change over time.

Identify emotions in speech, where intonation and stress patterns vary.

Segment speech into syllables or words using MFCC transition points.Drive machine learning models such as Capsule Networks, where MFCC frames act as low-level input capsules representing instantaneous acoustic states.

Conclusion

Although MFCCs are mathematical abstractions, their temporal behavior reflects the physical realities of speech production. By analyzing MFCC changes over time, one can gain insights into the dynamic articulatory processes behind human speech. This capability is essential for applications in speech recognition, speaker identification, and audio-based diagnostics.

Summary Report on CLIP Model and Prompt Usage

This report is based on a series of questions on the factors affecting prompt usage, sensitivities, and performance of the CLIP (Contrastive Language-Image Pretraining) model. The questions focus on understanding CLIP's working mechanism, prompt engineering, language effects, and potential vulnerabilities. The following is a summary of the key points gleaned from the answers to the questions and the practical use of CLIP.

Source of CLIP's Prompt Sensitivity

CLIP's Transformer-based text encoder is highly responsive to natural language context. The model is trained on 400 million image-to-text pairs, often using "a photo of..." It works with descriptive structures such as. Therefore, prompts that are similar to structures in the training data provide higher success. However, this dependency may partially limit the model's generalization power; Prompts that are far from the distribution of training lead to a decrease in performance.

Prompt Optimization and Variety

Methods such as "Prompt Tuning" and "Prompt Ensembling" are used to automatically optimize prompts. These techniques test a large number of prompts to find the phrases that work best, providing robustness in zero-shot classification. A variety of meaningful prompts can increase the stability of the model, while an irrelevant or large number of prompts can cause confusion. Grammatical structure differences also affect the score; For example, structures that are common in education, such as "A Photo of a Dog", produce more effective results.

Short Prompts and Language Effects

Short prompts often perform poorly due to a lack of context, but can be successful if they are clear and contain words that are mentioned in the training data. English prompts give the highest performance because the standard CLIP is trained with English data. Low scores are expected in different languages, such as Turkish; however, Multilingual CLIP (mCLIP, XCLIP) models can be used in other languages, but are less effective than English.

Classification Mechanism

In CLIP, the visual encoder remains stationary and the classification is done by comparing the text prompts to the visual vector. Therefore, the choice of prompt is the main determinant of performance. Similarity scores are often used for direct classification, but in the case of very close scores, additional thresholding methods such as softmax may be required.

Prompt Engineering etc. Fine-Tuning

Prompt engineering allows for fast and flexible optimization without touching the weights of the model. Fine-tuning, on the other hand, requires more data and calculations, but offers higher task-specific achievement. Prompt engineering is preferred for speed, fine-tuning is preferred for customization.

Reliability and Vulnerabilities

While prompt substitution is effective in zero-shot scenarios, semantic mismatch between text and image can lead to incorrect results. CLIP can be manipulated with adversarial prompts (e.g., misleading phrases) or minimal noise added to visuals. These vulnerabilities indicate that the model should be tested in controlled environments.

Self-Attention Example

This week, I studied the weight matrices Wq (Query), Wk (Key), and Wv (Value), which are crucial components of the self-attention mechanism in Transformer models. I learned how these matrices are used to understand the relationships between words (tokens) within a given context. To demonstrate this knowledge with an example, I applied the self-attention process step by step using the phrase “bank money grab.” In this example, I showed how the initial representation of the word “bank” is transformed into Query, Key, and Value vectors using the Wq, Wk, and Wv matrices. I then calculated the attention scores using these vectors and finally used the scores to generate a new context-aware representation for the word “bank.” This exercise helped me better understand how weight matrices and self-attention contribute to capturing the contextual meaning of words.

Output Details About Convolutional Layer

1-Why convolutional layer’s output may negative what it represents ?

Convolutional layers output is dependent on the kernel. Negative output generally shows opposite direction to the kernel. For example if we have horizontal edge detection for kernel, output 50 shows there is a horizontal feature that matches our kernel but output is -50 this shows this is the opposite direction of the kernel.

2-Higher output in Conv Layer is better ?

This one also depends on the kernel. If outputs coming from same kernel answer could be yes. But if it comes from different kernel answer is no. Because for example if we have 2 kernels and first one is detecting horizontal lines and second kernel detecting vertical lines their outputs could be different from each other and higher number is not a comparable thing in that case.