Supplementary Note 1

1 Content Overview

The content in this manuscript is summarized in Fig. 1.

2 Model Architecture and Training Methods

2.1 Classifier

In artificial intelligence, classifiers are essential algorithms for partitioning data into discrete classes. Linear classifiers separate classes using a linear decision boundary, represented as a hyperplane in a multidimensional feature space, dividing the space into distinct regions for each class. Support Vector Machines (SVMs) optimize this hyperplane to maximize the margin between classes, enhancing generalization to unseen data. Naive Bayes classifiers, based on Bayesian probability theory, assume conditional independence among features given the class label, enabling efficient computation of posterior probabilities for classification. In contrast, the k-Nearest Neighbor (k-NN) algorithm is a non-parametric method that classifies query instances based on the majority class among the 'k' most similar training examples, measured by a predefined distance metric. Linear Discriminant Analysis (LDA) reduces dimensionality while preserving class separability by projecting data into a lower-dimensional space that maximizes inter-class variance relative to intra-class variance. Logistic Regression, a probabilistic classifier for binary outcomes, models the log probability of the positive class using a logistic (sigmoid) function, providing interpretable probability estimates. Each method offers unique strengths, making them indispensable for diverse classification tasks.

2.2 Embedding models

Embedding models are essential in artificial intelligence for converting raw data into numerical vectors that capture meaning and context. Word2vec[1] is used primarily with text data, transforming words into vectors based on their usage within sentences. It learns from large amounts of text to place words with similar meanings closer together in a multi-dimensional space. Wav2vec[2] and its advanced version wav2vec 2.0[3] focus on audio data, especially speech. These models learn from raw audio recordings without needing detailed labels by predicting masked parts of the audio. This process helps them understand and represent the important features of spoken language. Wav2vec 2.0 improves upon this by using more sophisticated techniques to

better capture the context and nuances of speech, making it highly effective for tasks like speech recognition.

2.3 Network module

Network modules are key components in artificial intelligence that help process and learn from different types of data. A linear module, or fully connected layer, simply connects every input to every output with adjustable weights, making it useful for basic transformations of data. Convolutional Neural Networks (CNNs) specialize in handling structured data like images. They use filters to scan through the image, detecting important features such as edges or shapes, which makes them very effective for tasks like recognizing objects in pictures. Recurrent Neural Networks (RNNs) are designed for sequences of data, like sentences or time series. They have a "memory" that allows them to consider past information while processing new data, which is great for understanding language or predicting future events based on historical data. Transformers[4], including models like BART[5] and Whisper[6], are powerful tools for handling long sequences. Instead of using memory like RNNs, they pay attention to different parts of the input at once, allowing them to understand context better over longer stretches of text or speech.

2.4 Large Language Models

Large Language Models (LLMs) have transformed natural language processing (NLP) by employing deep learning to understand and generate human-like text. The GPT series introduced transformer architectures with self-attention mechanisms to capture contextual word relationships. These models are pre-trained on extensive text corpora using unsupervised learning to predict subsequent words in a sequence, enabling them to learn rich linguistic patterns that can be fine-tuned for specific NLP tasks. LLaMA, developed by Meta, adheres to a similar approach but focuses on efficiency and scalability. It utilizes an optimized transformer architecture to deliver strong performance across various NLP tasks while minimizing computational demands.

2.5 Contrastive learning

Contrastive learning is a type of self-supervised learning method in artificial intelligence where models are trained to distinguish between similar and dissimilar data points without the need for explicit labels. CLIP (Contrastive Language-Image Pre-training) exemplifies this approach by jointly learning from paired image and text data[7]. Trained on a large dataset of image-caption pairs, CLIP learns to associate images with their corresponding textual descriptions while differentiating them from unrelated ones. This training method allows CLIP to effectively understand and bridge visual and textual information, facilitating applications such as image retrieval, captioning, and cross-modal tasks. Through contrastive learning, CLIP achieves robust performance without the need for explicit labels for each image, showcasing the power of this unsupervised learning paradigm.

2.6 Generative methods

Generative methods in artificial intelligence create new data resembling the training data. Two key approaches mentioned in this review are Generative Adversarial Networks (GANs) and Vector Quantized-Variational Autoencoders (VQ-VAE)[8, 9]. GANs consist of a generator, which produces synthetic data, and a discriminator, which distinguishes real from fake data. Trained adversarially, the generator improves until it creates highly realistic outputs. VQ-VAE, a type of autoencoder, compresses data into a discrete latent space using learned codes and reconstructs the input from these codes. By decoding random points from this space, VQ-VAE generates new samples, enabling applications in image synthesis and audio generation.

References

[1] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[2] Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 (2019)

[3] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[4] Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

[5] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)

[6] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

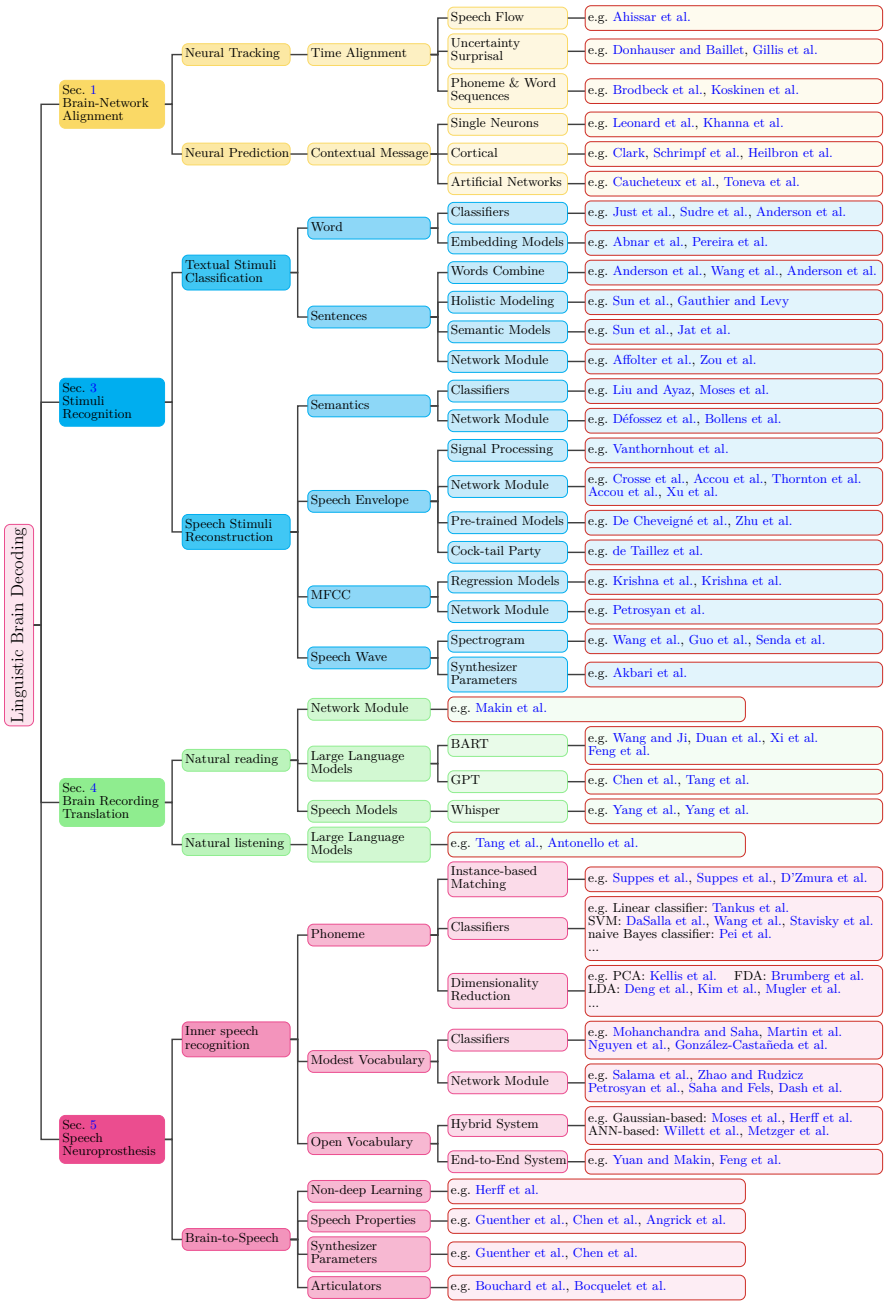
[7] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

[9] Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

Figures

Fig. 1: Content Overview.



The main content flow and categorization of this survey.