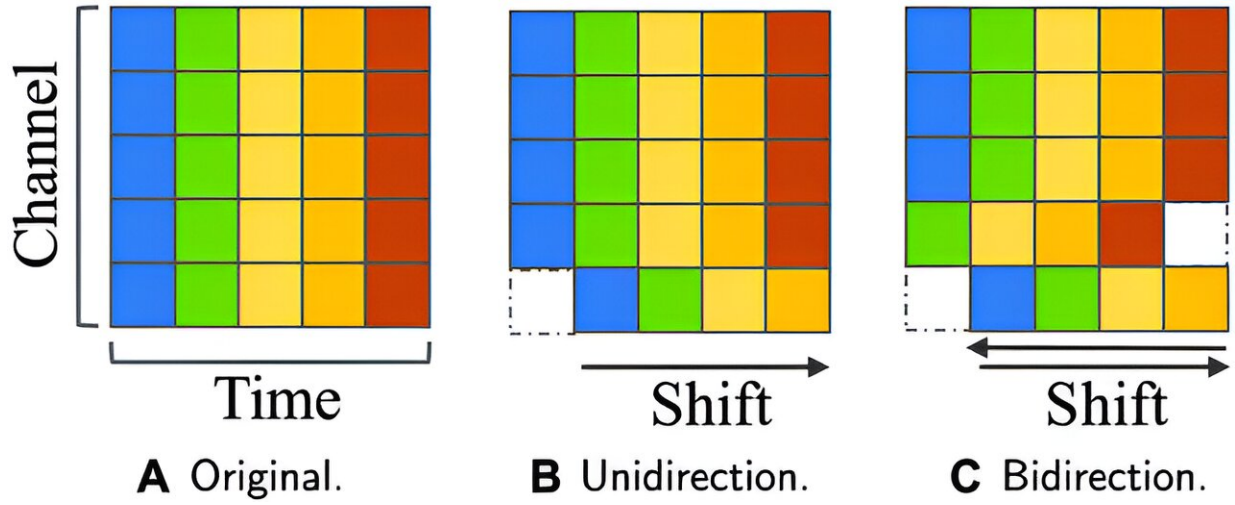
Temporal shift for speech emotion recognition

July 15 2024



Original representation and proposed temporal shift . Credit: *Intelligent Computing* (2024). DOI: 10.34133/icomputing.0073

Humans can guess how someone on the other end of a phone call is feeling based on how they speak as well as what they say. Speech emotion recognition is the artificial intelligence version of this ability. Seeking to address the issue of channel alignment in downstream speech emotion recognition applications, a research group at East China Normal University in Shanghai developed a temporal shift module that outperforms state-of-the-art methods in fine-tuning and feature-extraction scenarios.

The group's research was published Feb 21 in *Intelligent Computing*.

According to the authors, "This architectural enrichment improves performance without imposing computational burdens." They introduced three temporal shift models with different architectures: a convolutional neural network, a transformer and a long short-term memory recurrent neural network.

Experiments pitted these temporal shift models against existing models on the large benchmark IEMOCAP dataset and found them to be generally more accurate, especially in the fine-tuning scenario. The temporal shift models also performed well in feature extraction when using a trainable weighted sum layer.

In addition, the temporal shift models outperformed the baselines on three small datasets, RAVDESS, SAVEE and CASIA. Furthermore, temporal shift, serving as a network module, outperforms the kind of common shift operations that have been used for data augmentation.

The new temporal shift module achieves better performance by allowing the mingling of past, present and future features. Although such mingling benefits accuracy, it can also cause misalignment, which harms accuracy.

The authors employed two strategies to address this trade-off: control of shift proportion and selection of shift placement. Models were tested with one half, one quarter, one eighth and one sixteenth of all channels shifted; a larger proportion allows more mingling but causes more misalignment. Two different placement models were tested: residual shift, in which the temporal shift module is located on a branch of the network and thus preserves unshifted data alongside shifted data, and in-place shift, which shifts all the data.

After investigating shift proportion and shift placement, the authors chose the best-performing variants for each of the three architectures for conducting experiments against the state-of-the-art models in fine-tuning and feature extraction.

Existing speech emotion recognition methods that rely on deep neural network architectures are effective, but they face the challenge of accuracy saturation. That is, their accuracy does not increase with incremental increases in the network size. A key part of the problem is that channel information and temporal information are not processed independently.

Future work can investigate the influence of the scale of the dataset and complexity of the downstream model on accuracy. Additional downstream tasks, such as audio classification, merit quantitative analysis. Moreover, it would be advantageous to make the parameters of future versions of the temporal shift model learnable to enable automatic optimization.

More information: Siyuan Shen et al, Temporal Shift Module with Pretrained Representations for Speech Emotion Recognition, *Intelligent Computing* (2024). DOI: 10.34133/icomputing.0073

Provided by Intelligent Computing