Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading

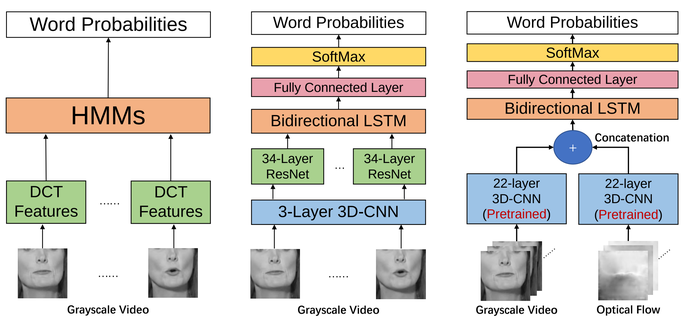
Xinshuo Weng, Kris Kitani
Carnegie Mellon University
{xinshuow, kkitani}@cs.cmu.edu

Background

- **Goal:** recognize the word being spoken from visual information alone, *i.e.*, a sequence of images.
- **Dataset:** LRW, the most challenging word-level lipreading dataset.
  - 500 target words;
  - 488766 training videos.

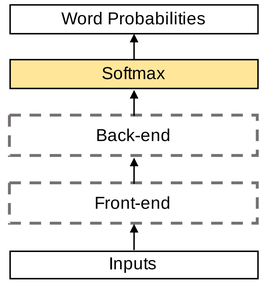
Motivation

- Using ***deep*** *(more than 3 layers)* spatio-temporal visual features extracted from 3D CNNs for lipreading is under-explored in traditional methods (a) and recent deep learning based methods (b)[2].
- ***Pre-training*** of deep 3D CNNs on large-scale video datasets can provide a good initialization for fine-tuning on the target video dataset.
- Using ***optical flow*** in addition to the video data has been proved to be useful in many video tasks, which is yet unexplored in lipreading literature.
- Therefore, we propose (c) for lipreading, which uses a deep 3D CNNs with video and optical flow as inputs, pre-trained on the Kinetics [1] dataset.

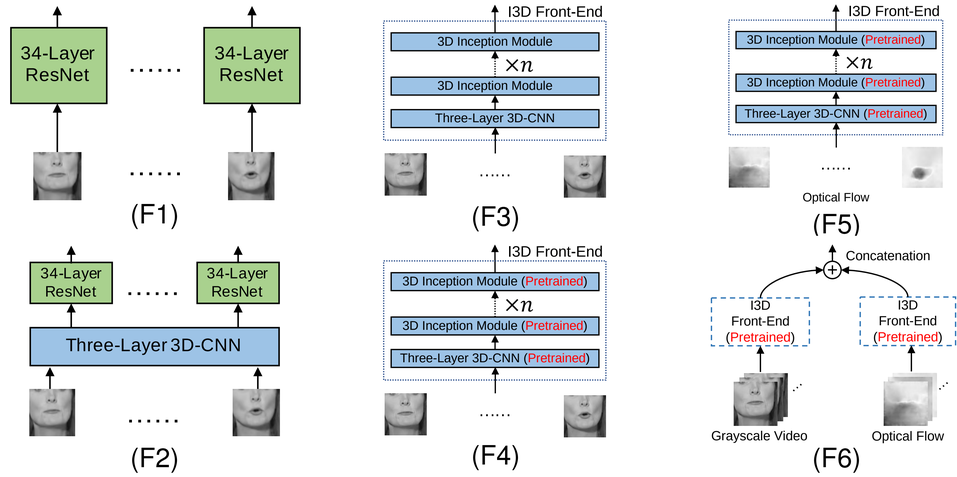


Lipreading Framework

- Inputs: which can be a sequence of grayscale images, optical flow data, or the combination of them;
- Front-end module: to extract the features from the input data;
- Back-end module: to model the temporal dependency and summarize the features into a raw score for each word;
- Softmax layer: to map to the word probabilities.



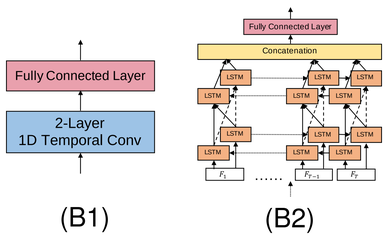
Front-End Modules



- (F1) and (F2): the commonly used front-ends are deep 2D CNNs in (F1) or deep 2D CNNs plus shallow (*e.g.*, three layers) 3D CNNs in (F2).
- (F3) and (F4): I3D front-end (deep 3D CNNs). Also, a pre-training on the Kinetics dataset is performed in (F4) while (F3) is randomly initialized.
- (F5) and (F6): the effect of using optical flow as inputs is explored. In (F5), only optical flow is fed into the front-end while in (F6) both video and optical flow are used. Also, both (F5) and (F6) perform pre-training.

Back-End Modules

- (B1): 1D convolution is operated across channels to aggregate the temporal information and a linear layer is used to map down the dimension to number of words.
- (B2): replace the 1D convolution in (B1) with a bidirectional LSTM.



Experiments

Data Preprocessing

- Grayscale video: a bounding box centered at the mouth is used to crop the images.
- Optical flow: generated by PWC-Net [3].

Results

Method	Val	Test
F2+B1	71.19	70.85
F2+B2 [2]	78.95	78.77

(T1)

Method	Val	Test
F1+B2	75.37	75.23
F2+B2 [2]	78.95	78.77

(T2)

- Conclusions: (T1) Bi-LSTM back-end is more powerful than the 1D convolution; (T2) Using a three layer shallow 3D CNN in the front-end help improve the accuracy up to ∼3%.

Method	Val	Test
F2+B2 [2]	78.95	78.77
F3+B2	59.11	59.45
F4+B2	81.73	81.52

(T3)

Method	Val	Test
F4+B2	81.73	81.52
F5+B2	82.17	82.93
F6+B2 (Ours)	**84.11**	**84.07**

(T4)

- Conclusions: (T3) Pre-training of the deep 3D CNNs on the large-scale video datasets leads to significant improvement after fine-tuning while training randomly initialized deep 3D CNNs often leads to a poor local optimum; (T4) using optical flow alone is effective for lipreading and the combination of optical flow and video data can further improve the accuracy.

Take-Home Message

- Pre-training of the deep 3D CNNs on large-scale video datasets is effective for good convergence.
- The motion information represented in the optical flow is often complementary to the raw video data for video tasks.

[1] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017.
[2] T. Stafylakis and G. Tzimiropoulos. Combining Residual Networks with LSTMs for Lipreading. In *Interspeech*, 2017.
[3] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *CVPR*, 2018.

British Machine Vision Conference (BMVC), 2019