Already Reviewed Papers:

|  |  |  |
| --- | --- | --- |
| [1] | B. S. S. W. &. N. d. F. Yannis M. Assael, "LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING". | |
| [2] | Owayjan M, Kashour A, Haddad NA, Fadel M, Souki GA (2012) The design and development of a lie detection system using facial micro-expressions. In: 2012 2nd international conference on advances in computational tools for engineering applications | |
| [3] | S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *PAMI*, 35(1):221–231, 2013. | |
| [4] | A. Garg, J. Noyola, S. Bagadia, “Lip reading using CNN and LSTM,” in Technical Report, 2016. | |
| [5] | | S. Hilder *et al.*, “Comparison of human and machine-based lip-reading,” in *INTERSPEECH*, 2009. |
| [6] | | T. Afouras, J. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” *arXiv preprint arXiv: 1809.02108*, 2018. |
| [7] | | Petridis, S., Stafylakis, T., Ma, P., Tzimiropoulos, G., & Pantic, M. (2018). *Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture. 2018 IEEE Spoken Language Technology Workshop (SLT).* DOI:10.1109/slt.2018.8639643 |
| [8] | | Brendan Shillingford*∗* , Yannis Assael*∗* , Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior, Nando de Freitas DeepMind & Google, *LARGE-SCALE VISUAL SPEECH RECOGNITION.* |
| [9] | | Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman, Visual Geometry Group, Department of Engineering Science, University of Oxford, UK, *LRS3-TED: a large-scale dataset for visual speech recognition.* |
| [10] | | T. Afouras, J. S. Chung, and A. Zisserman, “My lips are concealed: Audio-visual speech enhancement through obstructions,” *arXiv preprint arXiv: 1907.04975*, 2019. |

New Paper List:

1. Hearing by eye: The psychology of lip-reading.
2. Lip reading sentences in the wild
3. Lip reading in the wild
4. Extraction of visual features for lipreading
5. Lipreading Using Temporal Convolutional Networks
6. A novel image classification approach via dense-MobileNet models
7. A New Image Classification Approach via Improved MobileNet Models with Local Receptive Field Expansion in Shallow Layers
8. Rethinking the inception architecture for computer vision
9. Inception-v4, inception-resnet and the impact of residual connections on learning
10. Object-part attention model for fine-grained image classification
11. Residual attention network for image classification
12. Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers.
13. Lipreading with DenseNet and resBi-LSTM
14. Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-term Memory
15. SpotFast Networks with Memory Augmented Lateral Transformers for Lipreading
16. Synchronous Bidirectional Learning for Multilingual Lip Reading
17. Lip-Reading Using Pixel-Based and Geometry-Based Features for Multimodal Human–Robot Interfaces
18. A study of the temporal resolution in lipreading using event vision
19. A Transformer-based Model for Sentence-Level Chinese Mandarin Lipreading
20. Lip-reading System based on Bayesian Classifier
21. End-to-End Lip Synchronisation
22. LRRo: a lip reading data set for the under-resourced romanian language
23. Towards practical lipreading with distilled and efficient models
24. Speech Training System for Hearing Impaired Individuals Based on Automatic Lip-Reading Recognition
25. A multimodel keyword spotting system based on lip movement and speech features
26. An Experimental Analysis of Different Approaches to Audio–Visual Speech Recognition and Lip-Reading
27. Spatio-temporal fusion based convolutional sequence learning for lip reading
28. Multi-grained spatio-temporal modeling for lip-reading
29. Lipper: Synthesizing thy speech using multi-view lipreading
30. Combining residual networks with LSTMs for lipreading
31. Multi-view automatic lip-reading using neural network
32. Optimized input for CNN-based hyperspectral image classification using spatial transformer network
33. Accelerating neural transformer via an average attention network
34. A multiscale visualization of attention in the transformer model
35. Deep lip reading: a comparison of models and an online application
36. Multi-channel Transformers for Multi-articulatory Sign Language Translation
37. An improved automatic lipreading system to enhance speech recognition
38. An image transform approach for HMM based automatic lipreading
39. Toward movement-invariant automatic lip-reading and speech recognition
40. Visual words for automatic lip-reading

**1. Lip reading sentences in the wild**

In this paper [11], they have successfully achieved the goal of recognizing phrases and sentences uttered by a person. Moreover, they tried to show that their model can perform better than a professional lip reader using a novel dual-channel attention mechanism that can perform at a time with visual input and audio input, or both separately. Furthermore, they have used a unique curriculum learning strategy which is named as WLS (Watch, Listen, and Spell) model to accelerate training by reducing overfitting. They have gained word error rate on LRW and GRID dataset 23.8% and 3.0% respectively. On the other hand, we have built a separate model using mobile-Net and trained with our dataset and gained a very low error rate.

**2. Lip reading in the wild**

In this paper [2], they generated and developed a pipeline for fully automated large large-scale data collection from TV broadcasts with over a million-word instance, spoken by over a thousand different people. Moreover, they have built a CNN model to train this dataset effectively and recognize hundreds of words. They have separated their test set and named 500-class and 333-class respectively. They achieved the top-10 accuracy of 92.3% on the 333-word test set which exceeds the previous best on multiple datasets having lexicon sizes that were orders of magnitude smaller.

**3. Extraction of visual features for lipreading**

Matthews et al. describe the approach of top-down that fit a model of the inner and outer lip counters and derives lip-reading features from a principal component analysis of shape, or shape and appearance respectively, and another model which is a bottom-up method that uses a nonlinear scale-space analysis to form features directly from the pixel intensity [3]. Later, they applied a comparison on all methods on a multi-talker visual speech recognition task of isolated letters which is built with audio-visual recognition. They mainly addressed the extraction of features of lip-reading related videos by using this statistical approach.

**4. Lipreading Using Temporal Convolutional Networks**

Marinez et al. discuss a new approach of fixing the limitation of BGRU (Bidirectional Gated Recurrent Unit) by giving a solution that replaces the BGRU layers with Temporal Convolutional Networks (TCN) [4]. Moreover, they showed a simplified way of the training procedure, which allows the model to train in a single stage and addressed the variable-length procedure limitations. Their model achieved an improvement of 1.2% and 3.2% respectively on the largest publicly available datasets for isolated word recognition in English and Mandarin, LRW, and LRW1000.

**5. A novel image classification approach via dense-MobileNet models**

In this paper [5], they worked on memory-intensive and highly computational intensive  
features of deep learning which restricts its application in portable devices by compressing and accelerating network models which will reduce the classification accuracy. For this purpose, they proposed the Dense and MobileNet based model having dense blocks for classification purposes while the dense block used to improve the total performance of the mobileNet model structure by reducing parameters by at least half and the amount of calculation by nearly half using hyperparameters growth rate. Having that, they observed the classification accuracy of the whole model has increased to 92.1%.

**6. A New Image Classification Approach via Improved MobileNet Models with Local Receptive Field Expansion in Shallow Layers**

Wei et al. proposed three improvement models based on MobileNet with local receptive field expansion in shallow layers, also called Dilated-MobileNet (Dilated Convolution MobileNet) models, in which dilated convolutions are introduced into a specific convolutional layer of the MobileNet model without increasing the number of parameters, dilated convolutions are used to increase the receptive field of the convolution filters to obtain better classification accuracy [6]. They experimented with their model on the Caltech-101, Caltech-256, and AWA datasets where they gained an improvement of 2% higher classification accuracy from Dilated-MobileNet then the classical MobileNet.

**7. Inception-v4, inception-resnet, and the impact of residual connections on learning**

In this paper [7], they have discussed the better impact of Residual Connection on learning of any benefit in combining the Inception architecture with residual connections. Moreover, they have described the clear empirical evidence on training with residual connections which accelerates the training of Inception networks significantly. Lastly, they built an ensemble of three residual and one Inception-v4 model and achieved an improvement of 3.08% top-5 error on the test set of the ImageNet classification (CLS) challenge. On the other hand, we have implemented the inception-v3 model on our dataset and our customized model performed well on the training and validation set but did not perform well on the test set.

**8. Object-part attention model for fine-grained image classification**

In this paper [8], the authors proposed the object-part attention model (OPAM) for weakly supervised fine-grained image classification where the unique novelties are the object part attention model which integrates two-level attentions on object-level attention localizes objects of images and part-level attention which selects discriminative parts of the object by avoiding the heavy labor consumption of labeling. Hence, they have gained accuracy of 92.19% on CARS-196 datasets, 85.83% CUB-200-2011 datasets, 93.81% of OXFORD-IIIT PET datasets, and 97.10% on OXFORD-FLOWER-102 datasets; whereas, compared withmore than 10 state-of-the-art methods on this 4 widely-used datasets their OPAM approach achieved the best performance in this regard.

**9. Lipreading with DenseNet and resBi-LSTM**

In this paper [9], they work with a simple method to build a dataset for sentence-level Mandarin lip-reading from programs like news, speech, and talk show. They have used Hanyu Pinyin which is a phonetic transcription of Chinese as a label having 349 classes, while the number of Chinese characters is 1705 in the dataset. In their work, they first obtain the Hanyu Pinyin sequence by proposing a model that is composed of a 3D convolutional layer with DenseNet and residual bidirectional long short-term memory [9]. Hence, they have got the final Chinese character's results which are a model with a stack of multi-head attention applied to convert Hanyu Pinyin into Chinese characters. Likewise, we have also created our datasets and preprocessed all videos having 11 classes with a fixed frame rate and fed to our customized model to train the system. On the other hand, they used DenseNet and resBi-LSTM, whereas we have used MobileNet with Dense layers.

**10. Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-term Memory**

In this paper [10], they have proposed a hybrid neural network architecture, which integrates CNN and bidirectional LSTM (BiLSTM) for lip reading. On their work, they have firstly extracted key frames from each isolated video clip and used five key points to locate mouth region. After that, they extracted features from raw mouth images using an eight-layer CNN. The extracted features have the characteristics of stronger robustness and fault-tolerant capability. Finally, they have used BiLSTM to capture the correlation of sequential information among frame features in two directions and the softmax function to predict final recognition result. The proposed method is capable of extracting local features through convolution operations and finding hidden correlation in temporal information from lip image sequences. The evaluation results of lip-reading recognition experiments demonstrate that the proposed method outperforms conventional approaches such as active contour model (ACM) and hidden Markov model (HMM). On the other hand, we have done some pretty similar work to find which is best for our model, and lastly we chose MobileNet rather than using LSTM.

**11. SpotFast Networks with Memory Augmented Lateral Transformers for Lipreading**

# References

1. J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In CVPR, 2017.

2. J. S. Chung and A. Zisserman, “Lip reading in the wild,” in Proceedings of the Asian Conference on Computer Vision, 2016

3. Matthews, I., Cootes, T.F., Bangham, J.A., and Cox, S. Extraction of visual features for lipreading. PAMI, 24: 198 {213, 2002

4. B. Martinez, P. Ma, S. Petridis and M. Pantic, "Lipreading Using Temporal Convolutional Networks," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6319-6323, doi: 10.1109/ICASSP40776.2020.9053841.

5. Wei Wang, Yutao Li, Ting Zou, Xin Wang, Jieyu You, Yanhong Luo, "A Novel Image Classification Approach via Dense-MobileNet Models", *Mobile Information Systems*, vol. 2020, Article ID 7602384, 8 pages, 2020. <https://doi.org/10.1155/2020/7602384>

6. Wei Wang, Yiyang Hu, Ting Zou, Hongmei Liu, Jin Wang, Xin Wang, "A New Image Classification Approach via Improved MobileNet Models with Local Receptive Field Expansion in Shallow Layers", *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 8817849, 10 pages, 2020. <https://doi.org/10.1155/2020/8817849>

7. C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR, abs/1602.07261, 2016

8. Y. Peng, X. He and J. Zhao, "Object-Part Attention Model for Fine-Grained Image Classification," in IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1487-1500, March 2018, doi: 10.1109/TIP.2017.2774041.

9. Chen, X., Du, J. & Zhang, H. Lipreading with DenseNet and resBi-LSTM. *SIViP* **14,** 981–989 (2020). <https://doi.org/10.1007/s11760-019-01630-1>

10. Yuanyao Lu (1School of Electronic and Information Engineering, North China University of Technology, Beijing, P. R. China) and Jie Yan (1School of Electronic and Information Engineering, North China University of Technology, Beijing, P. R. China) International Journal of Pattern Recognition and Artificial Intelligence 2020 34:01

11.