Facial Emotion Recognition

Ling Li (lingli6@stanford.edu), Priyanka Rao (prao96@stanford.edu)
CS 231N

Motivation

What CNN models work best for facial emotion recognition (FER) and why?

Applications

Robotics
- Allow robots to better understand facial emotions
- Supplement spoken cues

Security & Surveillance
- Identify suspicious behavior
- Prevent danger

Advertising
- Tailor ads based on moods and reactions

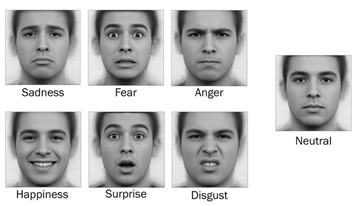
Social Media
- Automatically filter out images
- Better news feeds

Dataset

- Kaggle FER Challenge: 28,709 48 × 48 pixel grayscale images
- Pre-processed and centered

Methods

- Existing architectures
  - Alexnet
  - VGG-16
  - Inception
  - Inception-Resnet



- 3-layer CNN
  - [conv - relu - 2x2 max pool] – [affine – relu] – [affine]
  - 48 filters in the first conv layer, each 7 x 7 x 1

Results & Analysis

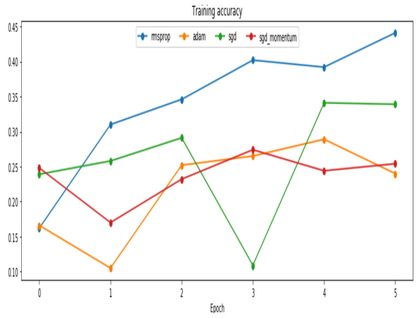


Fig 1. Training accuracy for different update rules for a shallow 3-layer CNN

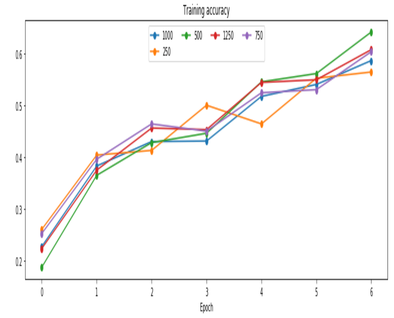


Fig 2. Training accuracy for different numbers of hidden dimensions for a shallow 3-layer CNN

- RMSProp converges faster
- Adaptive update rules overall better

- Optimal number of hidden dimensions is ~500
- Higher stability for higher number of hidden dimensions

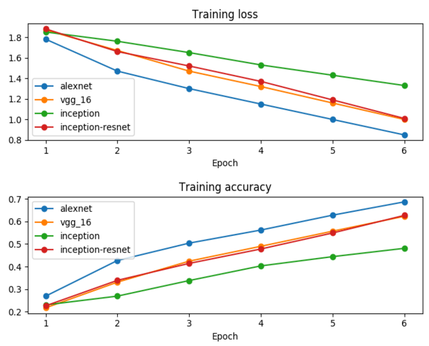


Fig 3. Training loss and training accuracy for existing architectures

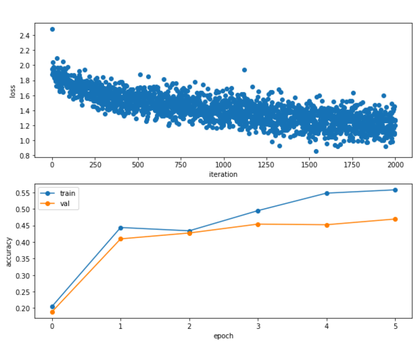


Fig 4. Training loss and training accuracy for shallow 3-layer CNN

- Shallow CNN trains faster due to fewer hyper-parameters
- For FER and small dataset, shallow CNN has comparable performance to deeper models

Model	FER Accuracy	ImageNet Accuracy
AlexNet	51.0%	57.0%
Pre-VGG	34.0%	-
VGG-16	49.4%	71.5%
Inception	46.3%	80.2%
Inception-Resnet	45.7%	80.4%
Shallow 3-layer CNN	52.1%	-

Conclusion

- Faster convergence and higher accuracy with simpler networks
  - Likely due to use of centered grayscale images for training
- Potential to greatly increase accuracy of Inception and Inception-Resnet for FER to match ImageNet accuracies
- For FER, rate of convergence approximately same for existing architectures
- Tendency for overfitting requires high regularization

Future Work

- Test on wild images
- Test different network and filter sizes
- Continue to improve models

References

[1] S. E. Kahou, C. Pali, et al. Combining modality specific deep neural networks for emotion recognition in video. *In Proceedings of the 15th ACM on International conference on multimodal interaction, pages 543–550. ACM,* 2013.

[2] Mollahosseini, Ali, et Al. "Going deeper in facial expression recognition using deep neural networks." *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE,* 2016.

[3] H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. *In Systems, Man, and Cybernetics, Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, vol. 4, pages 3732–3737. IEEE, 1997