Towards Embedded Emotion Recognition

(A Battery-Powered Friend)

Paul Shved [cs231n, Spring 2019]

Problem

What does it take to build an embedded device that "smiles back" at you?

Challenges:
- Embedded devices have limited hardware
- Neural networks require lots of computation

There's hope:
- Reading sensor data (cameras) requires little power [1]
- Training can be done offline
- TensorFlow Lite proven to run on Arduino [1]

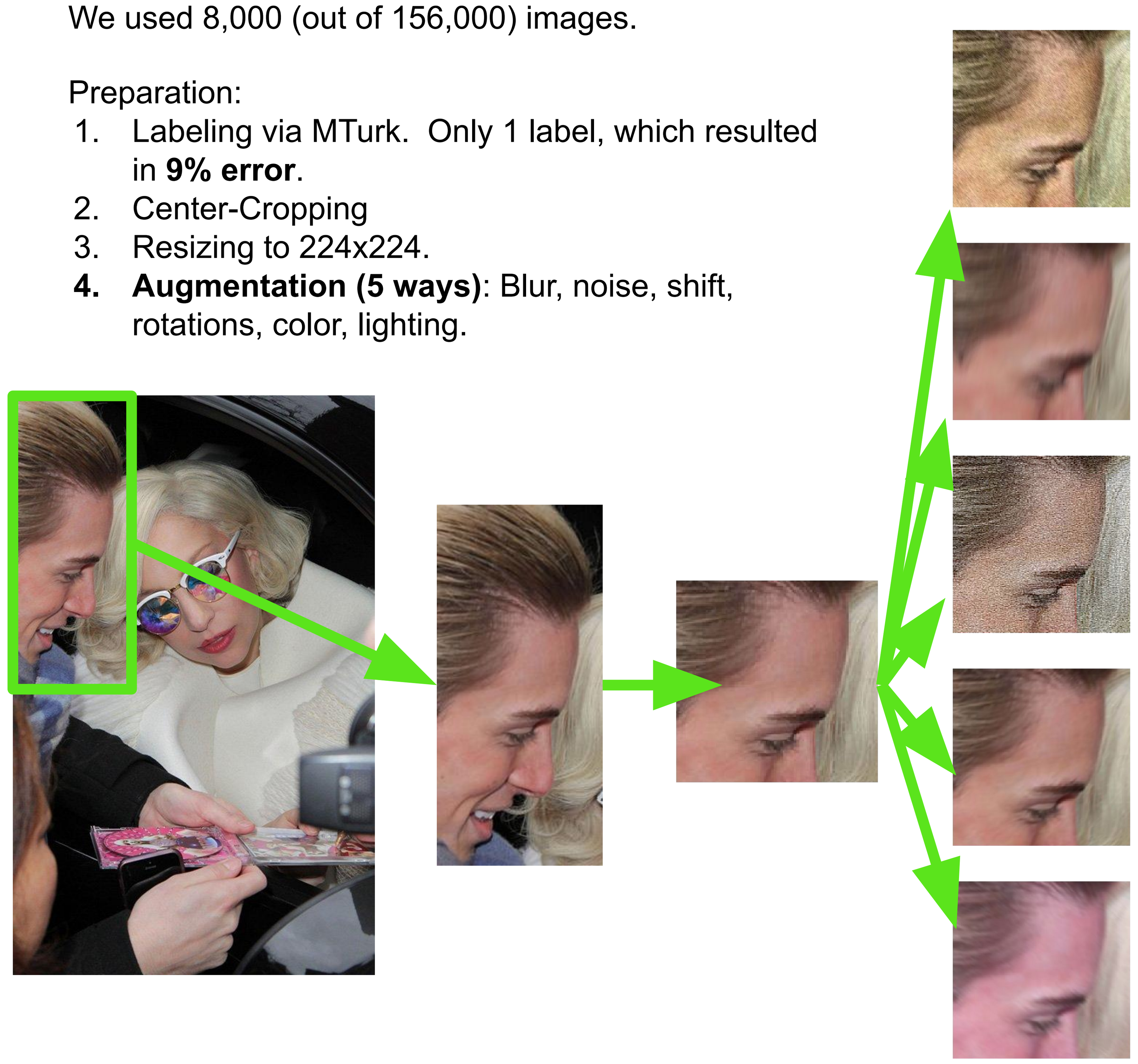
Assumption:
- An embedded device Apollo3 can perform 0.46 mln FLOPS [benchmarks]

Data

Google Facial Expression Comparison Dataset

We used 8,000 (out of 156,000) images.

Preparation:
1. Labeling via MTurk. Only 1 label, which resulted in **9% error**.
2. Center-Cropping
3. Resizing to 224x224.
4. **Augmentation (5 ways)**: Blur, noise, shift, rotations, color, lighting.



Approach

Overall Approach

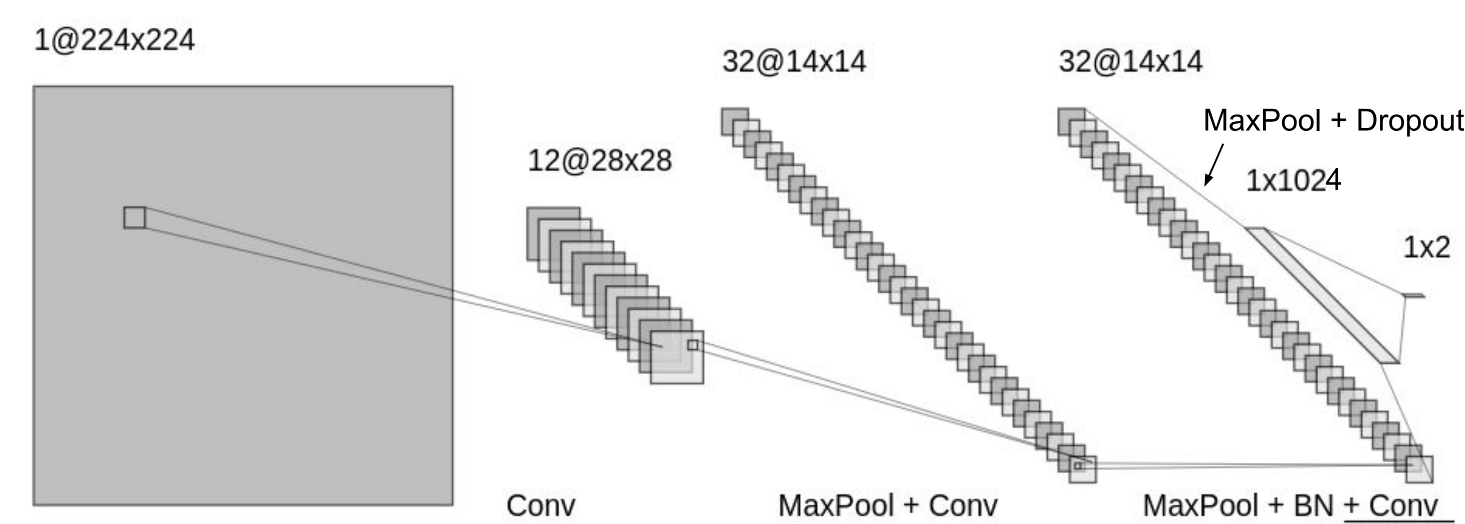
1. Define speed targets (0,46m FLOPS on Apollo 3)
2. Start with vanilla AlexNet
3. Define quality targets (-20% of AlexNet)
4. **Meta-learning**: reduce the net to meet the quality targets (AUC)
   a. Use validation set to evaluate quality

Using AlexNet for Emotion Recognition done in [2].

> We ran the meta-learning algorithm manually; automating it is left for future work.

Final Architecture

3 Convolution layers (the 1st layer large stride), 3 MaxPool Layers, 1 Batch Norm layer, 1 fully connected layer. Dropout on the FCN.

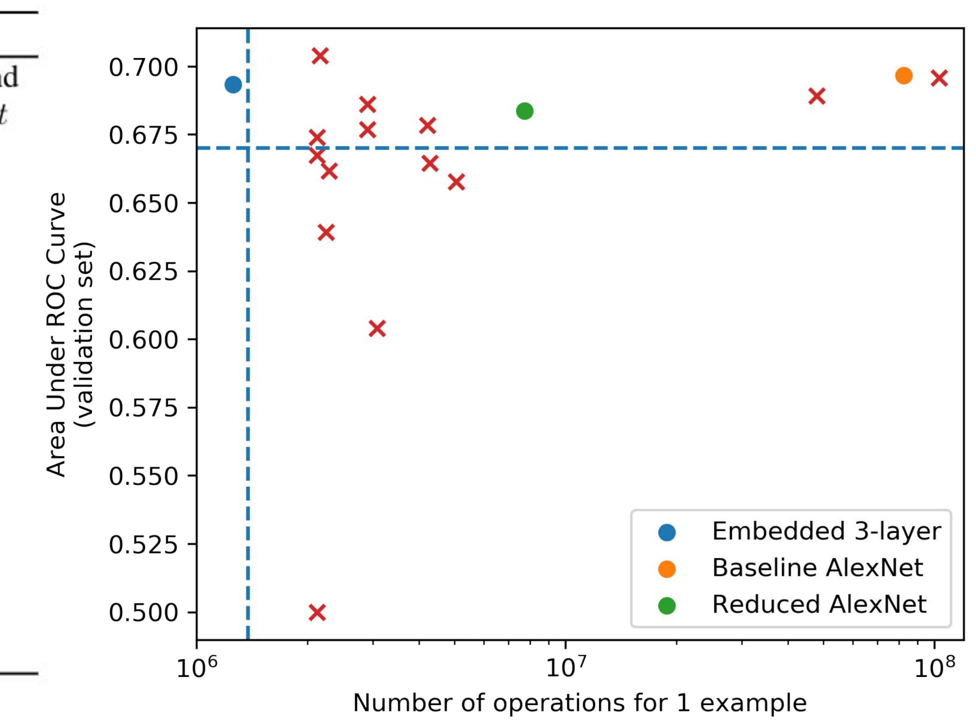


Meta-learning

Use A* search on model space (not hyperparameters!) with objective to minimize FLOPS while maintaining quality β_0

Algorithm 1: An A*-like meta-learning algorithm

Data: A baseline model m_0; function T that trains and evaluates the AUC on validation set $T : m \to t$
Result: Final model m
Function `Recurse` (m, β, s) **is**
  Data: Model m, AUC target β, FLOPS target s
  Result: Model m' or `nil`
  if $T(m) < \beta$
    return `nil`
  elif $Size(m_{best}) < s$
    return m
  for $m' \in Reductions(m)$
    $m_{best} \leftarrow$ `Recurse`(m',β);
    if $m_{best} \neq$ `nil`
      return m_{best}
  return `nil`
$\beta_0 \leftarrow 80\% \cdot T(m_0)$;
return `Recurse` (m_0, β_0)



References

[1] Pete Warden. ***Scaling machine learning models to embedded devices***, 2019. https://petewarden.com/2019/03/27/scaling-machine-learning-models-to-embedded-devices/

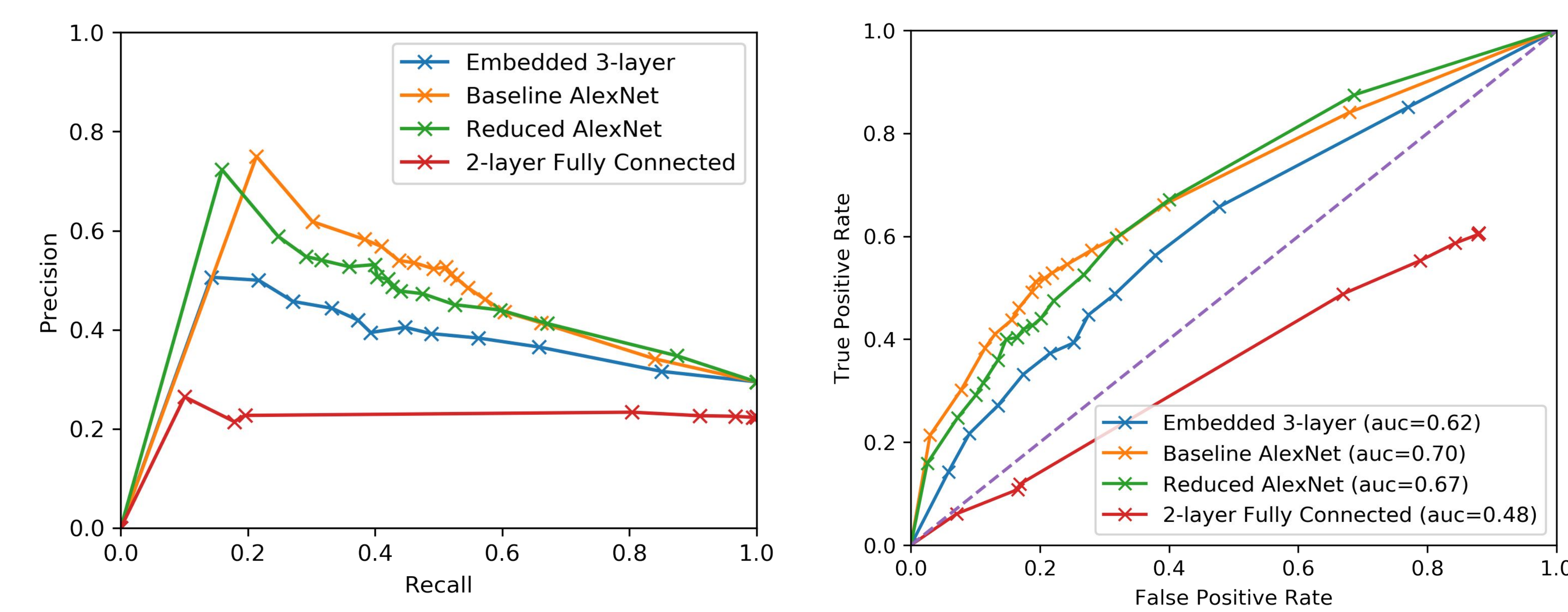
[2]: W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li. ***Audio and face video emotion recognition in the wild using deep neural networks and small datasets***. In proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016

Results

1. Found an intermediate model **16 times faster** with -20% AUC loss
2. Final model did not perform to expectations.
3. Did not actually build a robot ☹

Model	Set	AUC	Δ AUC	Accur.	Prec	Recall	FLOPS	Est runtime
Baseline AlexNet	test	0.7	0	0.72	0.53	0.52	82.7 mln	3min 1s
Reduced AlexNet	val	0.67	-15%	0.68	0.45	0.41	5.1 mln	
Reduced AlexNet	test	0.66	-20%	0.69	0.47	0.44	5.1 mln	11s
Embedded 3-layer	val	0.69	-5%	0.7	0.48	0.49	1.3 mln	
Embedded 3-layer	test	0.62	-40%	0.64	0.39	0.39	1.3 mln	2.9s

Analysis



Saliency maps are consistent with psychology studies [Duchenne, 1862]. The model attends most to the cheeks and eyes; rarely to the mouth.