Stanford ENGINEERING
Computer Science

# ROBO-NANNY: CONVNETS FOR INTELLIGENT BABY MONITORING

nest

## INTRODUCTION

Most baby monitors today are triggered on sound so parents are woken up every time there is some crying. They then need to look at the baby monitor to determine the baby status to see if adult intervention is required.

Often times, babies cry in the middle of the night but if they are still lying down (vs. standing up), chances are that they can drift back to sleep by themselves.

Baby monitors would have greater utility if the visual information can be interpreted by machine to determine whether the adult really needs to be alerted, resulting in less interruptions in sleep.

## THE PROBLEM

***To accurately determine the status of baby in crib using video feed from a baby monitor according to 5 states: Caretaker, Sit, Stand, Sleep and Empty.***

**Some challenges and objectives:**

**Working with Limited & Unbalanced Data:**

- Only have access to 1 baby for a limited number of days. Can we train an accurate deep learning model with ~2500 samples of labelled data data?
- Most of the images collected would be of the baby sleeping. Only a limited number would be in other states (sitting/standing etc.). How do we work with an unbalanced dataset?
- Is it even possible to reuse this model on pictures of other babies in cribs that the model hasn't seen before in the training?

**Motion Detection for Salient Frame Extraction & Localisation/Cropping:** What motion detection techniques can we use to help us with extraction of salient frames from the video feed and also to helps us localise the baby for more effective cropping (potentially making the model less sensitive to camera placement)?

**Model Architecture & Optimisation:** As an edge computing application where the forward pass is likely to be done on a mobile device, different model architectures should be compared for trade-offs between accuracy and compute/memory requirements.

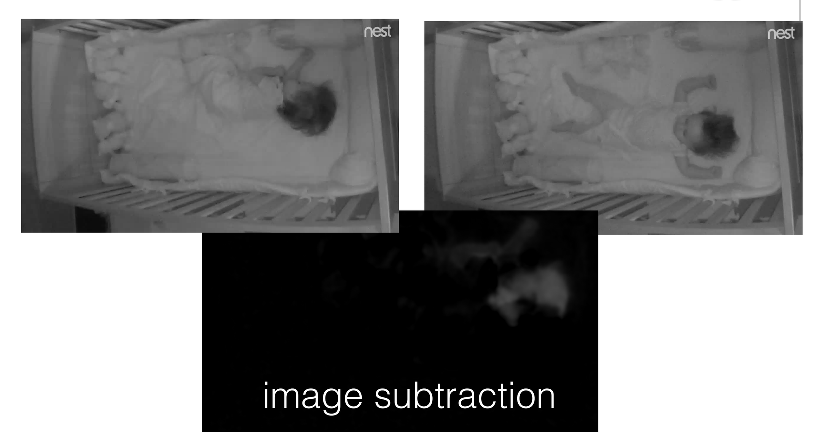
**Network Visualisation:** Each image class is actually a collection of the baby in many different positions in the crib, hence, would we get more recognisable images if we trained pixels to maximise a certain neuron activation in an earlier layer \rather than the final output layer?

## DATA PRE-PROCESSING

Capture Video Stream from Nest Cam (0.1 FPS)

Select Salient Frames using Motion Detection

$\tau_{lower} < |P[F(t-1)] - P[F(t)]| < \tau_{upper}$

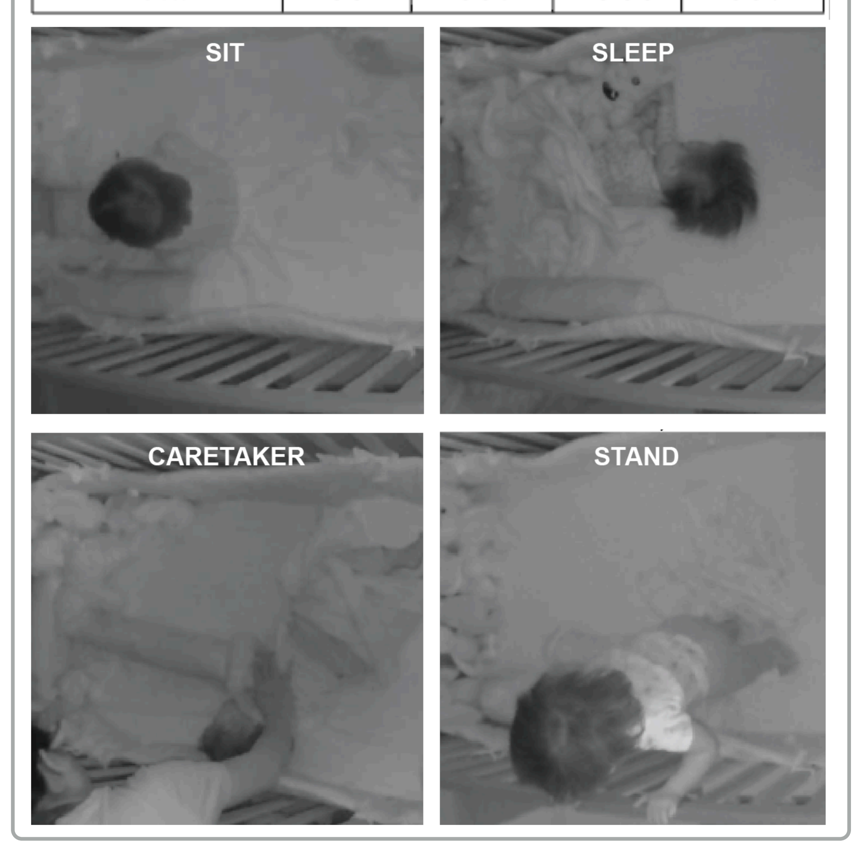



Manually classify images into 5 classes

| Class | A1 | A1% | A2 | A2% |
|---|---|---|---|---|
| Caretaker | 40 | 2% | 51 | 11% |
| Empty Crib | 182 | 9% | 278 | 11% |
| Sitting Up | 209 | 10% | 269 | 11% |
| Lying Down | 1165 | 58% | 1462 | 58% |
| Standing Up | 406 | 20% | 440 | 18% |
| Total | 2002 | 100% | 2500 | 100% |

Split into Train, Validation and Test Sets according to ratio - 68:20:12

Crop image to 224x224 or 227x227




Synthetic Augmentation of Training Set (e.g. flipping etc.)

CNN Model

*Note: In addition, a secondary test dataset was collected of other babies and cribs from Google/Youtube to see if the model predict an unseen dataset*

## METHODS & ALGORITHMS

**Convolutional Networks & Transfer Learning**

- Pre-trained AlexNet on ImageNet - with modified last layer to output 5 classes
- Pre-trained ResNet18 on ImageNet - with modified last layer to output 5 classes. Tested on training all weights after initialising with pre-train weights vs. training only last layer of weights ("ResNet18-Fr")
- Optimized using Stochastic Gradient Descent with Momentum

**Weighted Cross-Entropy Loss Function**

$$Loss(X,C) = W[C] * (-X[C] + log(\sum_j exp(X[j])))$$

Where X are images, C are classes and W are the weights applied to each class

**Multiple Layer Network Visualization**

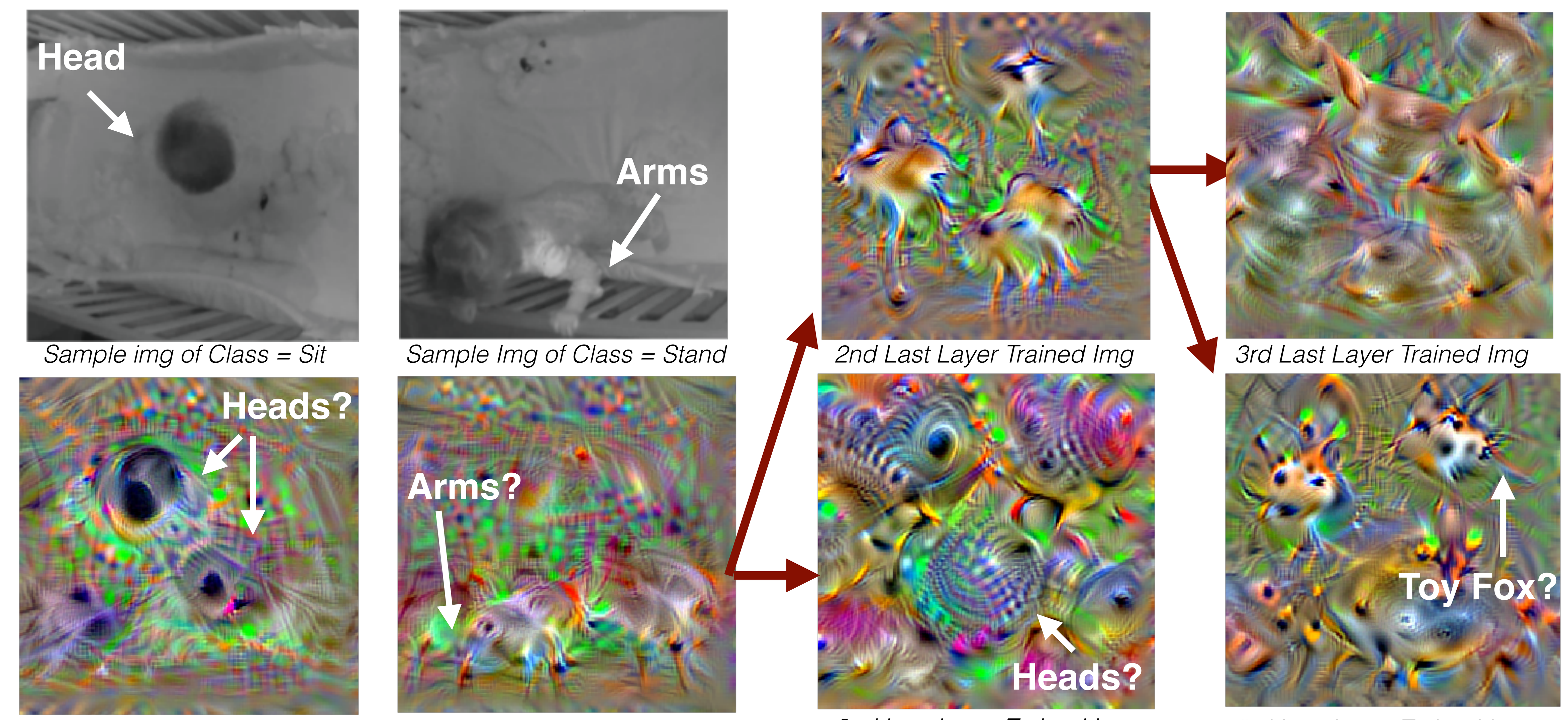
- Gradient Ascent on image pixels using fixed weights of trained AlexNet network to maximise output score for on fully connected layer (example: class = 4 "stand")
- Repeat above but maximising activation of a given neutron in the second last fully-connected layer (FC2). The index of the neuron is selected based on highest value in weight matrix corresponding to the class in question (i.e. argmax( W[4 , :] ) = index of neuron on FC2)

$$I^* = argmaxs_{N_y}(I) - \lambda\|\mathbf{I}\|_2^2$$

where, $N_y = argmax(W_{FC2}[y,:])$

- Repeat above for the third-last fully-connect layer

## "BABYDREAM" NETWORK VISUALIZATIONS




*Sample img of Class = Sit* *Sample Img of Class = Stand* *2nd Last Layer Trained Img* *3rd Last Layer Trained Img*

*Trained Image for Sit* *Trained Image for Stand* *2nd Last Layer Trained Img* *3rd Last Layer Trained Img*

## RESULTS & CONCLUSIONS

**Train, Validation & Test Accuracy**
*Different models and datasets*

| Network | Data | Weighted Accuracy Train % | Val % | Test % |
|---|---|---|---|---|
| ResNet18-Fr | A1 | 86.9% | 85.7% | 81.2% |
| ResNet18 | A1 | 100% | 93.5% | 90.5% |
| ResNet18 | A2 | 99.9% | 94.2% | 96.0% |
| ResNet18 | A2* | 98.2% | 94.7% | 96.7% |
| AlexNet | A1 | 97.5% | 93.7% | 94.2% |
| AlexNet | A2* | 97.5% | 94.3% | 95.0% |

*A2 has more data than A1. A2* is synthetically augmented*
*"ResNet18-Fr" refers to pre-trained with frozen weights other than last year. All other networks do not have frozen weights*

**Confusion Matrix**
*Resnet18 with Augmented A2 Dataset*

| Pred | Ground Truth Care | Empty | Sit | Sleep | Stand |
|---|---|---|---|---|---|
| Care | 5 | 0 | 0 | 0 | 0 |
| Empty | 0 | 31 | 0 | 0 | 1 |
| Sit | 0 | 0 | 24 | 2 | 0 |
| Sleep | 0 | 0 | 0 | 152 | 1 |
| Stand | 1 | 0 | 0 | 0 | 52 |
| Total | 6 | 31 | 24 | 154 | 54 |
| Recall | 83% | 100% | 100% | 99% | 96% |

**Preliminary Findings and Conclusions:**

- Using pre-trained model on Imagenet then retraining weights on custom dataset of 2500 labelled samples was sufficient to get good accuracy
- ResNet and AlexNet did not show significant difference in accuracy (fluctuations likely due to small test and validation data set)
- Weighted cross-entropy loss function was effective to address the unbalanced dataset
- Training pixels on earlier layers of the network created some interesting images but it wasn't very clear that the class image is a composite of these higher level images
- Some preliminary testing has been done with images of other babies and crib configuration taken from Google/YouTube. So far we have not yet been able to prove that data augmentation can make the model general enough to make predictions on an unseen dataset of other babies and crib confirmations

**Lily Cheng - lilcheng@gmail.com**

**CS231n - June 2017**