Depth-Based Activity Recognition in ICUs using CNNs

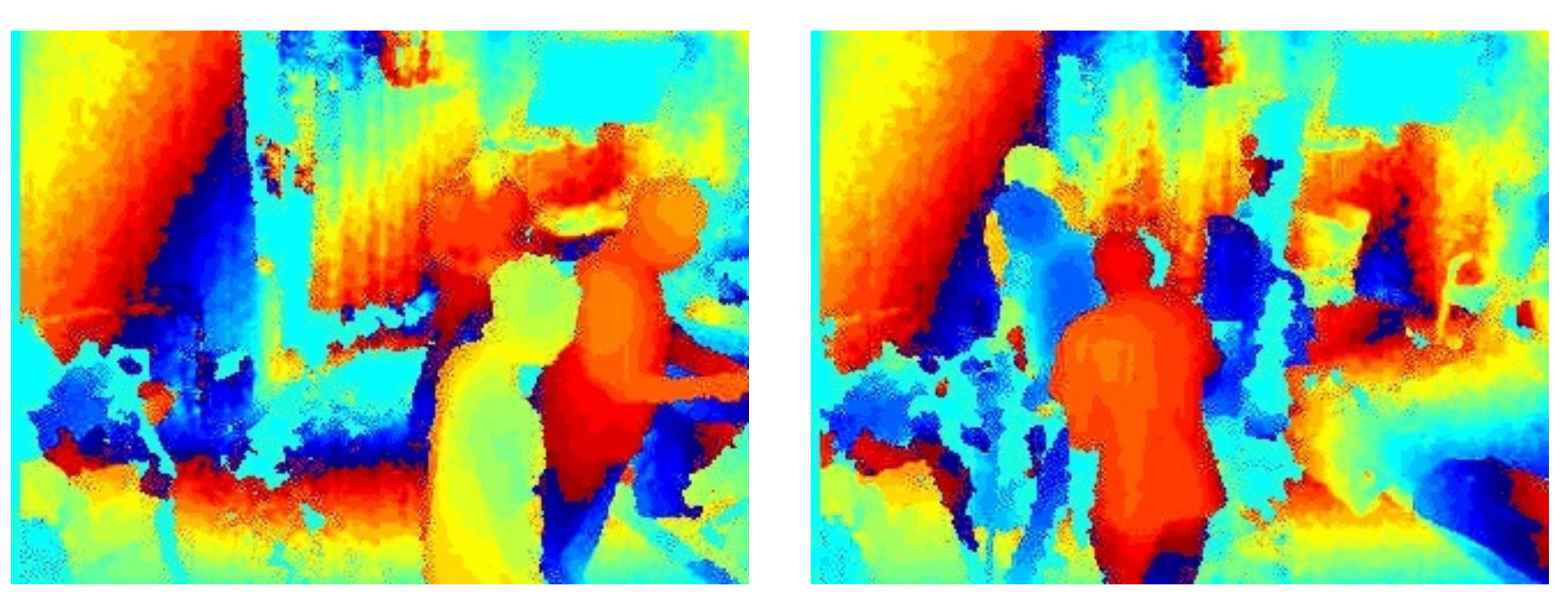
Rishab Mehra & Gabriel Bianconi
Stanford University

Project Overview

- **Problem**
  - Monitoring patients is an expensive and time-consuming task performed manually by nurses
- **Goal**
  - Design a system to automatically create a log of activities that occur in an ICU using Convolutional Neural Networks (CNNs) and Long Short Term Memory Units (LSTMs)

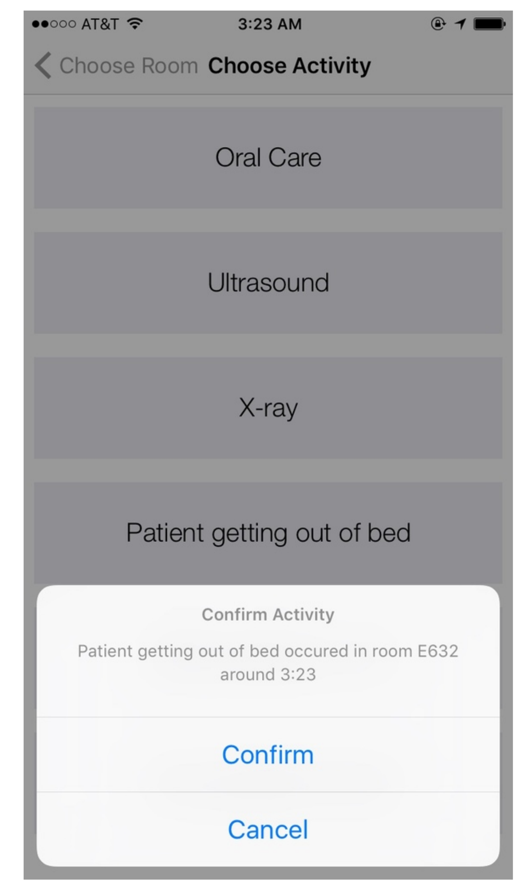
Data Collection

- **Overview**
  - Created a new dataset using depth sensors recording patient and staff activity in an ICU
  - Partnered with Intermountain Healthcare to collect data at the LDS Hospital
  - Recorded data in eight ICU rooms with four sensors each (different viewpoints)
  - RGB cannot be used for privacy reasons
- **Preprocessing Techniques**
  - Trained with random crops and horizontal flips
  - Normalized for ResNet-18 pre-trained on ImageNet
- **Dataset**
  - Sampled 500 64-frame clips from each class for training and validation
  - Created longer videos with an activity surrounded by background frames for testing



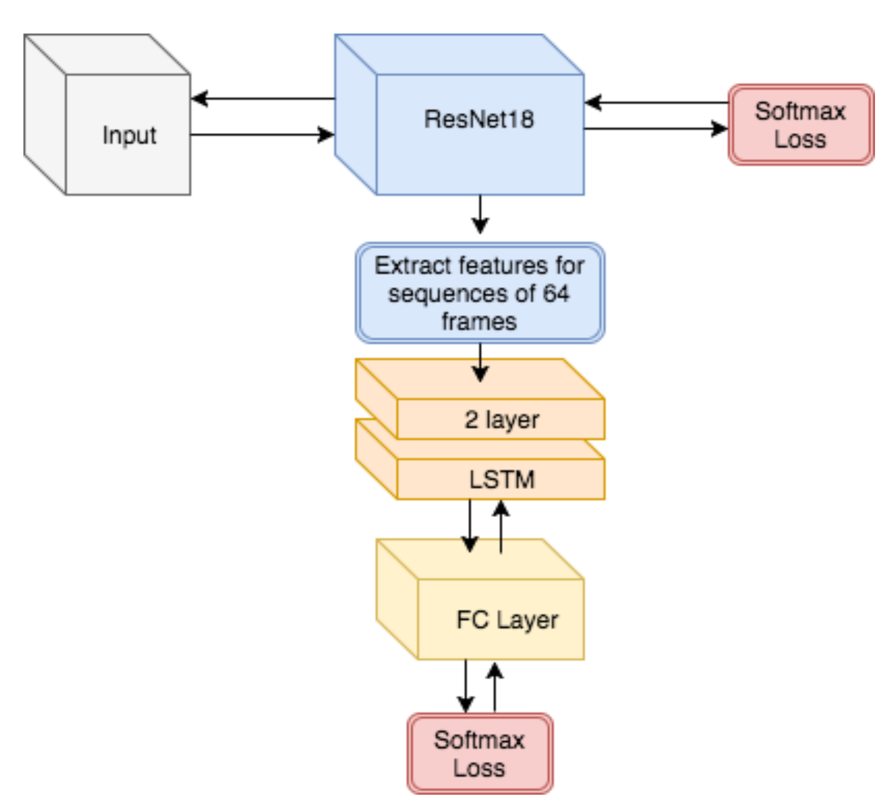
Data Labeling

- Created an iOS app to help nurses annotate the approximate time of activities
- Manually annotated precise time boundaries for activities
- Annotated seven activities: getting into bed, getting out of bed, getting into chair, getting out of chair, turning in bed, oral care, x-ray, and ultrasound
- Due to limited annotations by the nurses, we restricted the final dataset to patient getting out of bed, oral care, and background (no activity) clips



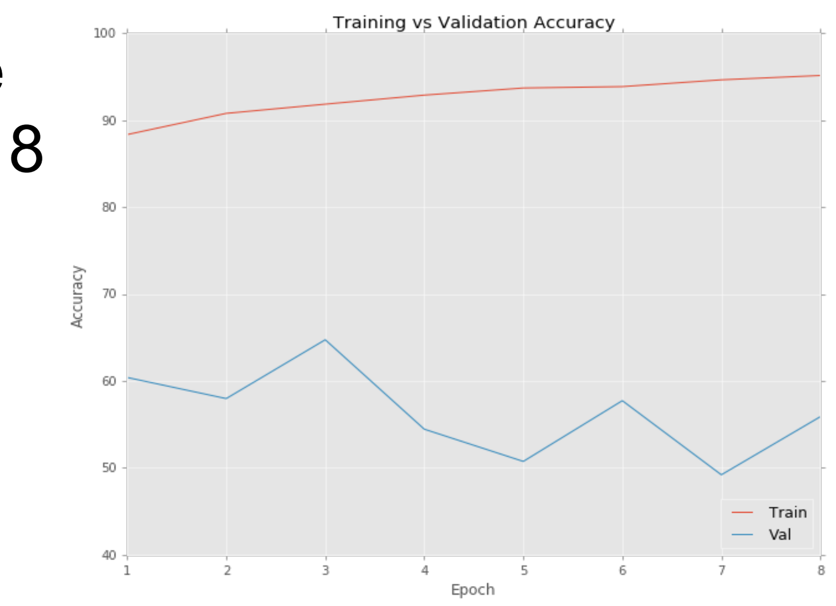
Methodology

- **Single Frame Models**
  - Used CNNs to classify activities frame-by-frame
  - Experimented with combining frames from different viewpoints depth-wise in these single-frame models
- **LSTM Model with Temporal Information**
  - Used pre-trained ResNet-18 (CNN) to extract features from the frames
  - Leveraged temporal information by feeding these features through an LSTM 64 frames at a time

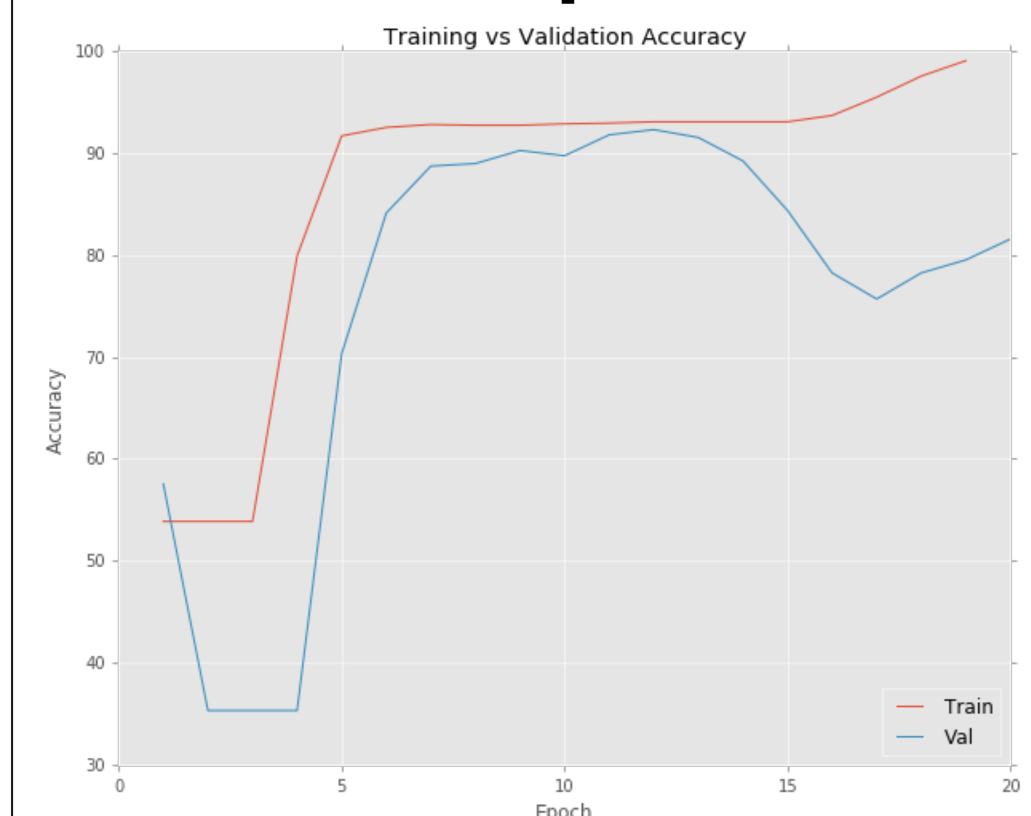


Baseline

- Calculated a baseline for the 3-class model with ResNet-18
- The training set was able to overfit to 94.63% accuracy, while the validation set was only able to 66.21% accuracy



Experimental Results



- Using the LSTM model, we were able to achieve a 99.1% accuracy on the training set and up to 92.3% accuracy on the validation set
- However, the results are very sensitive to initialization
- Tested our model on long unseen clips and successfully recognized activities and their time boundaries. An example is shown below: