Peer Review: Gesture Recognition Based on Deep Learning

December 7, 2020

Summary

Hou et al. proposes two types of convolutional neural networks to be used for gesture recognition. Primitive methods for gesture recognition uses extra mechanical devices and old text user interfaces that are unappealing. Rather, training neural networks for this task makes gesture recognition more feasible and accurate. The two types of architectures investigated in this paper are `LeNet` and `ResNet`. The datasets used to train these models are the Sign Language MNIST and Kinect Leap Dataset. The authors report that the `ResNet` architecture generally works better when the dimensions of the image samples are large, and the `LeNet` method works better otherwise. To cope for the difference in accuracies, the authors also propose a thresholding mechanism to determine which model to use for prediction given a gesture sample.

High-level Discussion

Pros

The paper is relatively easy to understand and should appeal to a large audience as gesture recognition is something widely researched. Additionally, since PyTorch is a readily available resource, reproducing these results should be straightforward.

Cons

Though easily reproducible, I fail to see the novelty in this project. In fact, the type of architectures used (`ResNet`, `LeNet`) are ones that are widely used and employed. For example, a simple Google search on using convolutional neural networks for gesture recognition returns several published papers. From these papers, a new addition that the authors proposed was the thresholding mechanism to determine which model to use for prediction. However, this seems like a really small tweak, which I think can be easily done through thresholding the dimensions of the image. I am not sure if this is a significant contribution to the gesture recognition research community, as `ResNet` is already known to be accurate for images with larger dimensions, as the network is built deeper. The latter portion of the paper mentions an ”App”, but I am not sure if this refers to a mobile application. If so, this would be a greater contribution, and should be made clearer in the paper.

Low-level Discussion & Nitpicks

The abstract mentions that the dataset consists of ”gestures of numbers”. Although the author specifically mentions the Sign Language MNIST dataset later in the paper, the authors should add pictures of some samples of this dataset. For example, ”gestures of numbers” alludes to people making gestures of different numbers, but I believe the Sign Language MNIST dataset consists of gestures of the alphabet. In fact, the abstract should be made clearer and samples of the dataset should be added in.

In the introduction, the author motivates this work by discussing the applications of gesture recognition. It is only clear that the authors try to use this for sign language AFTER the dataset is introduced much later in the paper. The authors should relate their work to how it can be readily used.

As for the results, an analysis following the results is much needed. The authors simply list the training and testing accuracies in the form of paragraphs. For example, the authors state that the training accuracy for the Kinect Leap Dataset only reaches 55%. Why do you think this happens? Have you tried optimizing the parameters of the architecture to improve the performance? If not, can you suggest how other people might improve this performance? It is also unclear when the authors say "the accuracy can **reach** 100%". I believe what the authors were trying to convey was that the accuracy can further increase if there were more epochs for training. I believe small things like this should be made clearer. If the authors have time, I believe you can also try to answer questions like: what is the test error of my algorithm as a function of the sample size? Answering questions like this make the paper stronger.

Some nitpicks:

- There are lots of grammar and spelling mistakes made throughout this paper. Here is a link to a good resource to follow when writing papers that I am also sometimes guilty of: [**Resource 1**].
- Here are some of the major mistakes:
  - Section 3.1: "... picture whose size been 28 × 28". Not sure what this means. Was the input size kept at 28 × 28 and then changed?
  - Section 3.1: 'convulational' is misspelled – 'convolutional'.
  - Section 4.3: 'Training' is misspelled. In addition, "With all tools we needed, which are models and datasets" is not a sentence. Try something like "For training, the tools we need are models and datasets."
- **The paper needs citations!!!** I think this is actually a major point than a nitpick – you need to cite your resources. For example, add resources when you talk about `LeNet` and `ResNet`.
- I am not sure what "x identity" means in Figure 3.
- Instead of having blocks of code throughout the paper, it might be neater to have an Appendix section at the end and consolidate all of the code there.

Summary of Review

Hou et al. shows that two different types of architectures of convolutional neural networks can be used for gesture recognition. Though interesting, the authors lack novelty. To resolve this, I propose that the authors possibly try to do more analysis with the models they currently have. For example, you can ask questions that limit the accuracies (or increase) of the architectures that show robustness. Future researchers can look at this and see that using Algorithm x is better than Algorithm y by answering such questions. Overall, the paper also lacks clarity. Fixing some grammar and spelling mistakes may significantly improve the quality of this paper.