Gesture Recognition with Recurrent Neural Networks

Justin Fu
University of California, Berkeley
`justinfu@berkeley.edu`

Siddhartho Bhattacharya
University of California, Berkeley
`siddartho_b@hotmail.com`

Abstract

Gesture recognition is typically done through images, but it requires extensive effort to process image data into a form useful for gesture recognition. Our project uses accelerometer data from smartphones, so no special hardware is required. Recurrent neural networks provide a natural way to model time sequences, and the LSTM architecture in particular provides a way to model long-range dependencies that are prevalent in the gestures we tested. We imlemented X gestures, and achieve an accuracy of X percent on the test set we generated.

1 Introduction

Intro goes here.

2 Data

We collected accelerometer data streamed from a smartphone. Specifically, we only considered values in the X and Y directions as seen from the perspective of the phone.

¡insert picture of phone with axes labeled¿

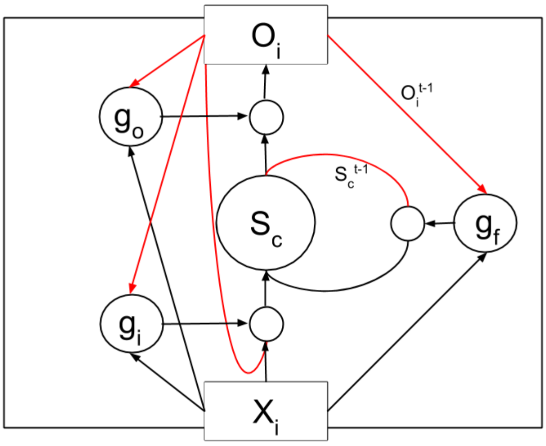
We implemented X different gestures, with Y examples of each.

3 Model

We implemented an LSTM (long short term memory) network based on the architectures proposed by Graves [1]. Specifically, our network contains the forget, input, and output gates, but lacks the peephole connections that connect the internal cell state to these gates (as done in Vinyals et al. [2]).

We then interleave layers of LSTM units with normal network layers to create a multilayer network.

Figure 1: A single LSTM cell. Red arrows are connections forward in time.



3.1 Equations

We used squared error as our objective function. O is the network output and Y is the training label. n sums over each training example, t sums over time, and i sums over the dimensions of each training example at a given timestep. Note that t varies with each training example.

$$L(Y, O) = \sum_{n}^{N} \sum_{t}^{len(O^n)} \sum_{i}^{I} \frac{1}{2}(o_i^{n,t} - y_i^{n,t})^2$$

Most of the equations on the forward and backward passes are the same as those in Graves [1], but some were missing from his paper and we dropped the peepholes so we specify all for completeness.

3.2 Forward Dynamics

The forward dynamics of the network are specified as follows. Equations are specified for a whole layer (rather than per unit). Bolded values represent vectors. Note that $*$ represents a pointwise multiply, and the nonlinearities are applied element-wise to the input vector.

Gates

$$\mathbf{a}_i^t = W_{i,x}\mathbf{x}^t + W_{i,o}\mathbf{o}^{t-1}$$

$$\mathbf{g}_i^t = \sigma(\mathbf{a}_i^t)$$

$$\mathbf{a}_f^t = W_{f,x}\mathbf{x}^t + W_{f,o}\mathbf{o}^{t-1}$$

$$\mathbf{g}_f^t = \sigma(\mathbf{a}_f^t)$$

$$\mathbf{a}_o^t = W_{o,x}\mathbf{x}^t + W_{o,o}\mathbf{o}^{t-1}$$

$$\mathbf{g}_o^t = \sigma(\mathbf{a}_o^t)$$

Cell State

$$\mathbf{a}_c^t = W_{c,x}\mathbf{x}^t + W_{c,o}\mathbf{o}^{t-1}$$

$$\mathbf{s}_c^t = \mathbf{g}_i^t * \sigma(\mathbf{a}_c^t) + \mathbf{g}_f^t * \mathbf{s}_c^{t-1}$$

Cell Output (Hidden State)

$$\mathbf{o}^t = \mathbf{g}_i^t * \tanh(\mathbf{s}_c^t)$$

Normal Layer Output

$$\mathbf{a}_n^t = W_n\mathbf{o}^t$$

$$\mathbf{n}^t = \sigma(\mathbf{a}_n^t)$$

3.3 Backpropagation

$\delta_k^{t,l+1}$ represents the error backpropogating from the above layer in the current time step. As with before, $*$ represents a pointwise multiplication. The δ variables represent vectors for the whole layer.

Cell Output

$$\delta_n^t = \delta_k^{t,l+1} * \tanh'(\mathbf{a}_n^t)$$

$$\delta_h^t = W_{i,o}^T\delta_i^{t+1} + W_{f,o}^T\delta_f^{t+1} + W_{o,o}^T\delta_o^{t+1} + W_{c,o}^T\delta_s^{t+1}$$

$$\frac{\partial L}{\partial \mathbf{o}^t} = W_n^T\delta_n^t + \delta_h^t$$

Cell State

$$\delta_c^t = \mathbf{g}_o^t * \tanh'(\mathbf{s}_c^t) * \delta_h^t * \delta_c^{t+1} * \mathbf{g}_f^{t+1}$$

$$\delta_s^t = \mathbf{g}_i^t * \sigma'(\mathbf{a}_c^t) * \delta_c^t$$

Gates

$$\delta_o^t = \sigma'(\mathbf{a}_o^t) * \tanh(\mathbf{s}_c^t) * \frac{\partial L}{\partial \mathbf{o}^t}$$

$$\delta_f^t = \sigma'(\mathbf{a}_f^t) * \mathbf{s}_c^{t-1} * \delta_c^t$$

$$\delta_i^t = \sigma'(\mathbf{a}_i^t) * \sigma'(\mathbf{a}_c^t) * \delta_c^t$$

Input

$$\delta_k^{t,l} = W_{i,x}^T\delta_i^t + W_{f,x}^T\delta_f^t + W_{o,x}^T\delta_o^t + W_{c,x}^T\delta_s^t$$

3.4 Training

We trained our network using BPTT (backpropogation through time) and gradient descent with momentum. All weights are randomly initialized in the range [-0.1, 0.1].

Training inputs (acceleration values) are directly presented to the network on the lowest layer. We present training labels to the network as a one-hot vector representing the gesture to the network at all time steps.

3.5 Decoding

To retrieve a single gesture prediction from the network, we run a forward pass and collect the network outputs. Then, we average the network outputs across time and select the class with the maximum average response.

$$class = argmax_i(\sum_t o_i^t)$$

A problem with this approach is that our network does not explicitly model this prediction rule.

An alternative (and something to try in the future) would be to build in a softmax into the network architecture as done in Vinyals et al. [2], and change our loss function to reflect this prediction rule and re-derive the backpropogation error.

4 Results

As we collected data ourselves, there were no avilable datasets online to compare our architecture against.

5 Example LaTeXstuff

- Item Example 1
- Item Example 2

Column	Col	Col
row blah	blah	blah
row blah	blah	blah

Table 1: Example table!

Noindent

5.1 subsection

```
\usepackage{times}
\usepackage{latexsym}
```

"(Graves, 2008) Quote"

5.2 Sections

Citations: Citations within the text appear in parentheses as (**?**) or, if the author's name appears in the text itself, as Gusfield (**?**). Append lowercase letters to the year in cases of ambiguity. Treat double authors as in (**?**), but write as in (**?**) when more than two authors are involved. Collapse multiple citations as in (**?**; **?**). Also refrain from using full citations as sentence constituents. We suggest that instead of

you use

> "Gusfield (**?**) showed that ..."

If you are using the provided LaTeX and BibTeX style files, you can use the command `\newcite` to get "author (year)" citations.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, e.g.,

> "We previously showed (**?**) ..."

should be avoided. Instead, use citations such as

> "Gusfield (**?**) previously showed ... "

5.3 Footnotes

Footnotes: Put footnotes at the bottom of the page and use 9 points text. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes should be separated from the text by a line.[2]

Acknowledgments

Brian is our savior.

References

[1] Alex Graves. 2008. *Supervised Sequence Labelling with Recurrent Neural Networks*. PhD Thesis.

[2] Orial Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. 2014. *Show and Tell: A Neural Image Caption Generator*.

[1]This is how a footnote should appear.

[2]Note the line separating the footnotes from the text.