

Tinkering with Tingets: Using Computer Vision to Improve Image Processing in a Children’s Social and Emotional Learning Toy

Merve Cerit

Graduate School of Education
Learning Sciences and Developmental Psychology
mmervecerit@stanford.edu

Victoria Docherty Delaney

Graduate School of Education
Learning Sciences and Teacher Education
vldocherty@stanford.edu

Abstract: Emotion classification through computing is a task that has interested scientists for over half a century. We echo the enthusiasm for this work using image classification, computer vision, and a convolutional neural network to classify seven emotions in Tingets, a children’s social-emotional learning toy. Using sparse categorical cross-entropy and ADAM optimization, our model achieved over 96% accuracy. However, the close-proximity nature of Tingets photographs made low variability in the data set and model overfitting constant points of contention.

I. Introduction and Problem Statement

Decades of research in children’s abilities to identify and regulate emotion has shown to have positive outcomes on wellness, mental health, and social-emotional awareness (Schweinhart, 2003; Shala, 2013; World Economic Forum, 2016). However, although children with higher emotional intelligence tend to accrue long-term mental health benefits, the degree to which all children are able to label their emotions is highly variable (Tominey et al., 2017). Studies suggest that young children struggle to classify their emotions, specifically negative emotions, which may lead to difficulties in learning to manage and regulate negative emotion. Furthermore, this may yield greater discrepancies between children who have mastered emotional identification and regulate and those who have not. These differences may be exacerbated by COVID-19 as many children’s normal home environments, learning environments, social interactions, and routines have been substantially disrupted.

Tingets, a multimodal emotion identification toy, seeks to aid young children in this problem space. Tingets are plush, friendly, monster-like tangibles that enable children to construct original arrangements of facial features in their physical environments. Children can choose from “happy,” “sad,” or “angry” eyes and/or mouth facial features to stick on the Tinget tangible. Ideally, the child would select facial features for their Tinget that correspond to their current mood. Then, by taking a photo of the Tinget and uploading the photo to a website, a corresponding cartoon picture would appear on the computer screen. An example of a Tingets toy and web representation can be found in **Appendix A**. It is worth noting that Tingets were originally conceived as a master’s thesis for the first author of this paper. The actual Tingets plushes used in the data set are still in prototype phase.

The purpose of this project is to enhance the visual translations of Tingets from the physical world to the digital. Specifically, we aim to use image classification on the data set to label Tinget emotions from seven classes: happy, sad, angry, disgusted, excited, tired, and surprised. Our initial development in this project will build toward an eventual classifier that will label Tingets instantaneously without saving the photo in cloud storage to preserve children’s privacy and identities when using the tool.

II. Related Work

This work builds from advancements in image and emotion classification that have proliferated artificial intelligence over several decades. An eventual goal of this work is to use localization in eye-mouth facial features to more accurately translate Tingets made by children, a task that has been developed in the field long before neural networks. Methods that precede neural networks nonetheless identified that localization of facial features via segmented facial

regions were a key component. Sobottka and Pitas explored facial recognition through feature extraction, and by using localization, classification hypotheses were verified using search for facial features in segmented regions of the human face (Sobottka & Pitas, 1996). Data were augmented by varying color, light conditions, and face shape. Their early work identified that the human eyes and mouth are crucial in the extraction phase. Similarly, Hajati, Faez, and Pakazad confirmed the importance of eye location using localization and principle components analysis (Hajati, Faez, and Pakazad, 2006). These examples of early work confirm that attending to eyes and mouths in faces using localization is productive as an overall optimization strategy.

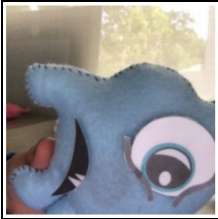

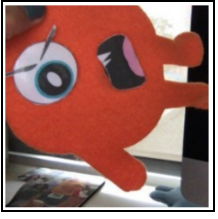

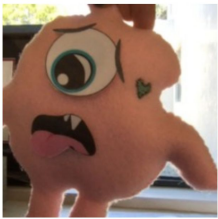


Convolutional neural networks and deep learning offer a faster approach to the facial expression recognition problem, especially recognition in real time. Mollahosseini, Chan, and Mohammad classified six facial features using a deep neural network with two convolutional layers, a max pooling layer, and four inception layers (Mollahosseini, Chan, and Mohammad, 2016). Their approach was tested across seven public repositories of faces data and provided faster, more accurate classifications than comparable methods.

Transfer learning has enabled new applications for emotion classification in novel contexts. Yang and colleagues demonstrated that neural networks could be used to classify six emotional states in students during distance learning (Yang et al., 2018). Using feature extraction, subset features, and emotion classifiers, local facial components were identified then categorized using Sobel edge detection, ultimately producing a characteristic value for the face. This contemporary application permits teachers to assess students’ emotional states while learning at a distance. Like Yang and colleagues, we aim to conduct a similar approach and expand the breadth of computer vision applications devoted to the wellness of children.

III. Data Set and Features

The data used in this work consist of 4,725 photos compressed to (64, 64, 3) pixels. The photos were reduced from (224, 224, 3) dimensions in order to increase tractability between our data and model. Table 1 below summarizes the data and provides an example of each emotion category. Importantly, we also included images with more noise to increase variability and ensure that our model identified images even though they are not cut perfectly.

Table 1: Tinget Image Data Set

Tinget Image Data (n = 4,725)			
Sad Tinget (n = 656) 	Happy Tinget (n = 696) 	Angry Tinget (n = 673) 	Excited Tinget (n = 663) 
Disgusted Tinget (n = 693) 	Tired Tinget (n = 649) 	Surprised Tinget (n = 695) 	70% Training Data (n = 3,413) 15% Validation Data (n = 603) 15% Test Data (n = 709)

Data were collected using Google’s Teachable Machine’s webcam capture feature. This preserved consistency in file format (jpg) and photo size. Initially, we judged the feasibility of our project using a training set of 1908 images and a test set of 180 images. These tests involved only Angry, Sad, and Happy Tingets. Once we were confident that our model was workable, we supplemented initial data with Surprised, Excited, Tired, and Disgusted Tingets as well as noisy images

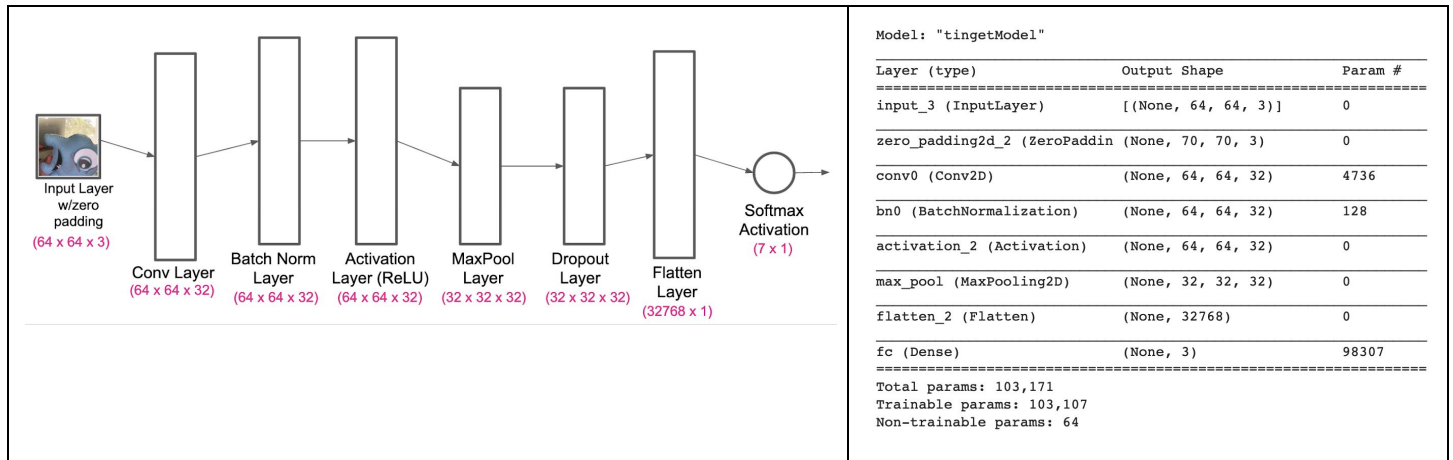
to the Sad and Angry classes. Noisy images contained human subjects, background details, and generally less-cropped photos around the Tinget to improve the classification abilities of the model. All Tinget photos contained eye-mouth pairs and were classified with an emotion label (i.e. there were no Tingets with eyes only, or Tingets of unknown emotional origin). Overall, the final data set contained 4,725 images and all emotional classes were approximately balanced. Data were randomly shuffled into train, development, and test sets for all subsequent testing in accordance with the proportions given in Table 1.

IV. Methods

We constructed a CNN that detects the physical Tinget and its facial features then converts it into the corresponding digital Tinget via image classification, reporting the detected emotion back to the user (child). This allows the model to run on a web application in conjunction with Tinget’s online storytelling platform. Currently, our model classifies the input image into one of the 7 classes: angry (0), happy (1), sad (2), excited (3), surprised (4), tired (5) and disgusted (6).

We used Keras to build our CNN. Our CNN currently contains a zero padding layer, a convolution layer with stride length = 1, a batch normalization layer with size 32, a ReLU activation layer, a max-pooling layer, a flattening layer, and a softmax function. The softmax function calculates the most likely emotion class based on the image. The total number of parameters as well as a visual of the model can be found in Table 2 below.

Table 2: CNN with Parameter Count



ADAM optimization and sparse categorical cross-entropy objective and loss functions were used to model the output and train the classifier. The sparse categorical loss was necessary because we aimed to identify seven unique emotional states rather than a binary output, and the emotional states were coded as integers rather than one-hot encodings of individual categories.

V. Results

Variability within our data set was a continual problem throughout the project’s development. In fact, initial trials of the model yielded 100% accuracy after very few epochs, suggesting that we needed urgent and significant increases to the complexity of our data. We originally desired only images of Tingets close to the camera because it is not ideal for a children’s toy to obtain image data of the children themselves without parental consent. However, this did not reflect the very realistic tendencies of children to take sloppy photos, and it interfered with our model’s ability to train. The remainder of the project was therefore oriented toward strategies to raise variation within and between emotion classes. Table 3 below reflects our efforts to inject variability into the data and train the models using a variety of machine learning principles. Table 4 illustrates outcomes of hyperparameter tuning that led to our final model selection with the default learning rate 0.001.

Table 3: Development Phases of Tingets Model

Version	Summary of Changes	Rationale	Result
0	Baseline model: train and test data set, three emotional states (Angry, Happy, Sad), CNN	Ensure that our model and theory works before expanding our idea.	Working but highly overfitted model.
1	ADAM optimizer with default parameters, sparse categorical cross-entropy loss, addition of validation set	Need validation set as part of our model.	Improved test performance, but may be overfitting that is not visible since the data is still not realistic.
2	Added cross validation, early stopping and model checkpoint to pick the best model	We wanted to see the model's performance over different validation sets.	Model is still highly overfitted. This helped us narrow the problem down to the data itself rather than the validation set.
3	Added realistic/noisy pictures and four additional emotion classes to the data	Our cropped data with low variability produced a highly overfitted model.	Still near-perfect accuracy, which made us more suspicious of data leakage. A random bird picture gets classified as angry with 100% prob from Softmax.
4	Removed cross validation and inserted dropout regularization, changed weight initialization from default (Glorot uniform) to He_Normal.	Needed to reduce model overfitting and induce variability. Softmax probability outputs are problematic.	Lower accuracy, higher loss, improved model overall.
5	Changed ADAM learning rate between 0.1, 0.01, 0.001, and 0.001	We wanted to adjust hyperparameters to see if it might further address the fitting problem.	Default learning rate performed the best: highest train, validation, and test accuracy. Keep default learning rate (0.001).

Please see our github repo (<https://github.com/mmervecerit/tingets>) to see the code of each version included in Table 3.

Table 4: Hyperparameter Tuning, Outcomes

	Learning Rate			
	0.1	0.01	0.001	0.0001
Model Accuracy and Loss: Graph				
Model Accuracy and Loss: Result	Loss = 0.32787 Test Accuracy = 0.94076	Loss = 0.13739 Test Accuracy = 0.96333	Loss = 0.134056 Test Accuracy = 0.96615	Loss = 0.16336 Test Accuracy = 0.94358
Early Stopping	After 6 epochs	After 18 epochs	After 18 epochs	After 20 epochs

Learning Rate for ADAM	Training Accuracy	Validation Accuracy	Test Accuracy
0.1	0.8728	0.9453	0.9408
0.01	0.9414	0.9751	0.9633
0.001	0.9446	0.9751	0.9661
0.0001	0.8995	0.9569	0.9436

Error Analysis

As you can see in Table 4, there is an interesting phenomenon occurring in the models: the training accuracy is lower than validation and test accuracy. These results made us think that our randomly-selected validation and test subsets might include difficult-to-classify images, hence why we applied cross validation and observed that this trend persisted in some of the folds. Another important factor in our accuracy irregularity is the dropout regularization. We used dropout with a keep-probability of 0.5, which made training accuracy lower.

Overall, our results seem to indicate that we have a strong model for classifying the Tingets' emotions. It appears we resolved the overfitting problem, and we might even increase the keep-probability for the dropout. However, we are aware that a next data set should contain more real-world pictures for validation and test sets. We are aiming at developing a demonstrative app to collect real-world data from the users (young children). Training duration will also be a criterion for model selection in our future work, although in this project, timing and efficiency were not bottlenecks.

VI. Discussion

The CNN was successful, perhaps *too* successful, at emotion classification with Tingets images. Our perfectly cut data set included images with a focus on the Tinget. Despite different backgrounds, variation in lighting, and colors of the tangibles (which we hypothesized would increase the variability of the data), our model was highly accurate regardless. Training images were highly similar with the validation and the test images, which created a potential data leakage issue, hence the overfitting in our model. Dropout regularization and inclusion of new data increased variability; however, future iterations of testing will involve more data from the users in variable contexts and settings, different toys, and images without Tingets.

Another important point is that we will continue to train our model with a greater proportion of negative label images, other commercial toys similar but not identical to Tingets to improve the model's ability to detect Tinget-specific emotions. We also want to have an object detection model to detect the Tinget's exact location and cut the bounding box before feeding the image to our classification model.

One advantage of our model involves its speed of convergence. We observed that our loss function stabilizes around 20 epochs, preserving computational efficiency. It is unlikely that our model will remain this efficient as we continue to diversify the data set, but because Tingets are fairly similar to one another, this may lead to faster training and fewer images required in order to train. We will ultimately continue to monitor and develop model performance as the ability to classify Tingets in variable settings is a key goal, particularly if the produce is launched commercially.

VII. Contributions

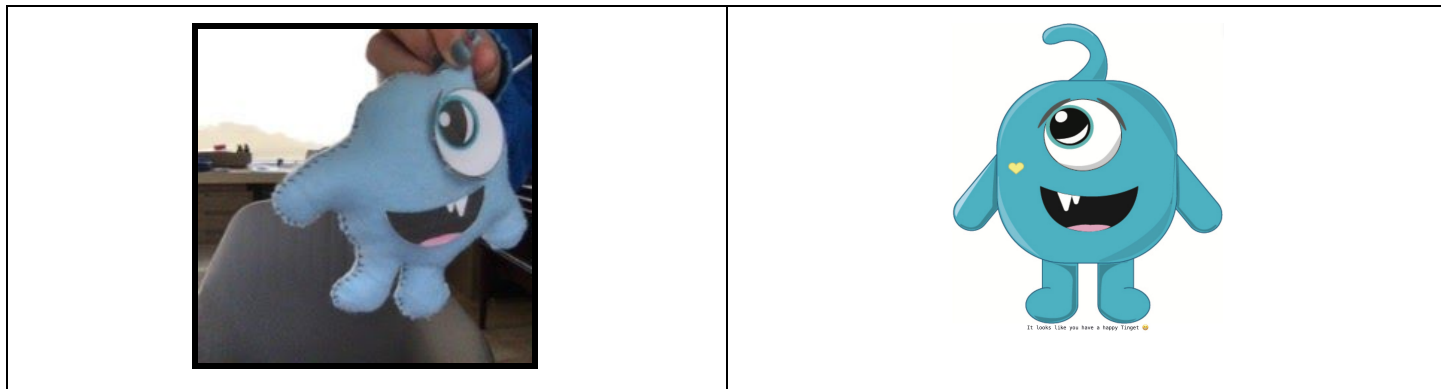
Both authors contributed and built upon ideas from their learning in CS230 to the growth and development of the project. Merve managed the various model versions in Keras and was in charge of running/training. She gathered the data and improved upon its variability when the model proved to have too few imperfections. Tingets were also developed as a primary component of her master's thesis in the Learning Design and Technology (LDT) program in the Graduate School

of Education. Victoria managed the research and technical writing components of the project, error analysis, and hyperparameters. She scoped the project and provided feedback on increasing variability in the data. Finally, she owned coordinating with Teaching Assistants and clarifying expectations.

References

- Hajati, F., Faez, K., and Pakazad, S. (2006) "An Efficient Method for Face Localization and Recognition in Color Images," *2006 IEEE International Conference on Systems, Man and Cybernetics*, Taipei, 2006, pp. 4214-4219, doi: 10.1109/ICSMC.2006.384796.
- Mollahosseini, A., Chan, D., and Mahoor, M. H. (2016). "Going deeper in facial expression recognition using deep neural networks," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY, 2016, pp. 1-10, doi: 10.1109/WACV.2016.7477450.
- Schweinhart, Lawrence J. (2003). "Benefits, Costs, and Explanation of the High/Scope Perry Preschool Program", High/Scope Educational Research Foundation, http://www.highscope.org/file/Research/PerryProject/Perry-SRCD_2003.pdf
- Shala, M. (2013). The Impact of Preschool Social-Emotional Development on Academic Success of Elementary School Students. *Psychology*, 4, 787-791. doi: 10.4236/psych.2013.411112.
- Sobottka, K., and Pitas, I., (1996). "Face localization and facial feature extraction based on shape and color information," *Proceedings of 3rd IEEE International Conference on Image Processing*, Lausanne, Switzerland, 1996, pp. 483-486 vol.3, doi: 10.1109/ICIP.1996.560536.
- Tominey, S. L., O'Bryon, E. C., Rivers, S. E., & Shapses, S. (2017). Teaching emotional intelligence in early childhood. *YC Young Children*, 72(1), 6-14.
- World Economic Forum (2016) "New Vision for Education: Fostering Social and Emotional Learning through Technology".
- Yang, D., Alsadoon, A., Prasad, P. W. C., Singh, A. K., & Elchouemi, A. (2018). An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment. *Procedia Computer Science*, 125, 2–10. <https://doi.org/10.1016/j.procs.2017.12.003>

Appendix A: Tingets Visualizations



Tingets are displayed in the images above. Children place eyes and mouth features onto the plush (left) and upload photos onto the web application, yielding a digitized Tinget (right). Additional examples of Tingets in action can be found at the links below:

<https://www.youtube.com/watch?v=T0z2EkVr3LY>

https://drive.google.com/file/d/1JLdNULhuaBp5aZWhxNBiwNKfu759_rbL/view?usp=sharing