Classification in Math Class: Using Convolutional Neural Networks to Categorize Student Cognitive Demand

Victoria Delaney [1] and Jai Bhatia [2]

[1] *Stanford University, 485 Lasuen Mall, Stanford, CA 94305, United States*
[2] *Fremont High School, 575 W. Fremont Avenue, Sunnyvale, CA 94087, United States*

Abstract
Maintaining cognitively demanding instruction is a primary goal of classroom teachers. Yet students' cognitive demand is difficult to measure and track during the enactment of a rigorous task. This in-progress research addresses this problem space by predicting and modeling students' cognitive demand with computer vision and convolutional neural networks, providing an in-the-moment analysis of cognitive demand during an eighth grade mathematics task enactment. The findings suggest that models which leveraged behavior-based visual proxies for cognitive demand (e.g., gesturing, using a computer) achieved substantially higher accuracy than the baseline model. Taken together, the results of this work build toward a classroom analytic tool for teachers and have implications for the contributions of computer vision in real-world classroom studies.

Keywords
Cognitive Demand, Mathematics Education, Convolutional Neural Networks, Transfer Learning

1. Introduction

There has been much interest in applying artificial intelligence analytic tools to classroom settings in the past decade. Although many educational applications that leverage AI examine speech data with natural language processing [26, 29], there exists a growing enthusiasm for computer vision-based research in classrooms to analyze and improve teachers' instructional practices [2, 7, 10]. This study explores the extent to which students' cognitive demand, one aspect of classroom instruction, can be modeled with computer vision via the analysis of classroom video recordings in eighth grade mathematics.

The maintenance of students' *cognitive demand*, the amount of intellectual work required to create meaning for a mathematical task and solve it [14], is crucial for teachers to measure and track from students because of its direct relationship to learning outcomes [27]. When students exhibit high cognitive demand, they develop deeper understandings and connect concepts across the discipline [28]. However, cognitive demand is not a static construct and can be influenced by a number of instructional factors, including the initial presentation of the task to students [14], resources provided to the students while solving the task [19], and teacher-student and student-student interactions during enactment [15]. Because measuring cognitive demand in-the-moment is difficult, yet potentially beneficial for teachers, we are curious to explore the extent to which computer vision may be used to provide cognitive demand measurements as students solve a mathematical task in small groups. Such data may assist teachers by providing indicators for which students continue to exhibit high cognitive demand throughout the task's enactment, and conversely, which students struggle to uphold high demand after the task is launched.

Since cognitive demand is not a purely visual construct, our model draws upon five proxy student behaviors to identify *potentially cognitively demanding activity* while solving a

EMAIL: vdoch@stanford.edu, jbhatia187@student.fuhsd.org

mathematics task, then uses the presence of the five behaviors to predict the level of cognitive demand. Though this approach omits cues from students' speech, we hypothesized that modeling cognitively demanding visual behaviors may yield additional contributions toward predicting overall demand. We therefore ask: to what extent can computer vision model changes in students' cognitive demand during mathematical problem solving?

2. Literature Review

Modeling cognitive demand with computer vision is a novel task in classroom analytics research. Our exploration of relevant literature investigates the extent to which other computer vision-based methods have demonstrated success in tasks with adjacent features. By incorporating these features: transfer learning, multiclass binary classification, and use of pre-trained networks, into the present study, we aim to utilize the affordances of computing toward a classroom setting. We discuss each feature in detail.

Transfer Learning. This research leverages transfer learning using ImageNet pre-trained weights; an approach that is not uncommon for developing novel applications in image classification. Since its onset, ImageNet has been established as a reliable, general-purpose benchmark for transfer learning on a variety of learning tasks, number of classes, and amounts of trainable data [17]. Numerous past studies have investigated the relationship between factors that impact transfer learning and fine-tuning of convolutional neural network (CNN) models, including the perils of model overfitting [6, 30] and the layers of ImageNet that should be optimized for transfer learning [16]. We drew upon this research when considering the duration of hyperparameter tuning and the overall fit of the training data to each binary classification model, as it suggests that model overfitting may be perilous to transferring learning to validation and testing sets.

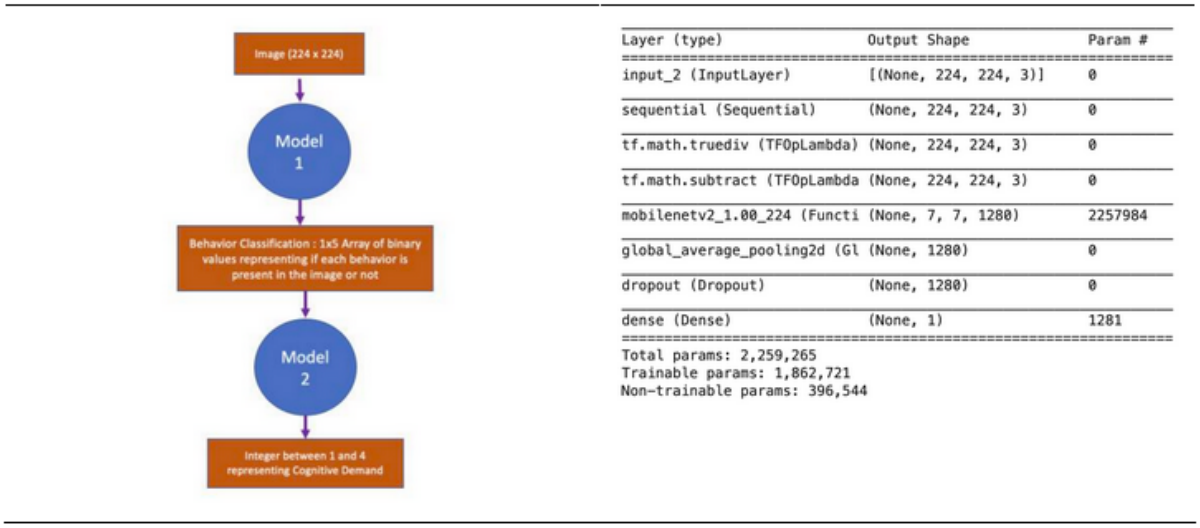
Pre-Trained Networks for Multiclass Classification. Additionally, we relied on MobileNet V2, a neural network specifically constructed for classification tasks, for binary and categorical classifications of cognitive demand. MobileNet V2 was developed for "lightweight classification tasks" [11] in transfer learning, image classification, and localization. It is commonly used in object recognition and classification tasks, such as detecting human tissue abnormalities in medical research [3]. As our investigation involves the classification of certain objects in order to detect human behavior (e.g., detecting the presence of a computer in the "using a computer" proxy behavior class), MobileNet V2 served as a reasonable choice for a first-pass exploration of the data.

One drawback experienced was the amount of labeled training data needed to optimize transfer learning using ImageNet and MobileNet V2. Past studies and experiments suggest that a large quantity of labeled training data is required for transfer learning, particularly in tasks that involve feature localization [1, 31] and modification of architectures that improve transfer learning [4]. Using unlabeled data has been an appealing area to explore in this research space [20] and some self-supervised methods have attempted to improve feature generalization in auxiliary tasks [5], although none have outperformed ImageNet's performance on purely supervised learning tasks. Weak supervision, which applies noisy labels from non-expert users [18], is now seen as a plausible middle-ground for large-scale ImageNet transfer learning tasks. We utilized weak supervision when applying hand labels to binary classes in the training data, as one member of the research team was unfamiliar with coding student and teacher behaviors in mathematics education research.

3. Methods

Our approach to modeling cognitive demand through convolutional neural networks consisted of three primary steps. First, we constructed the baseline cognitive demand model for comparison, which predicted demand from still images alone. Next, we devised the experimental model, which utilized binary classification for students' cognitive demand proxy activities (computer use, leaning in, pointing to the task, talking to the teacher, and writing on the task) to predict cognitive demand. Finally, we compared performance between the two models.

Both models applied transfer learning from ImageNet weights. Although the MobileNet V2 network, which relies upon ImageNet weights, contains approximately 2.3 million parameters, our method applies transfer learning to the bottom Dense bottleneck trainable parameters (approximately 1,300 layers). These layers solely focus upon the localized features of the five binary classes. Figure 1 shows a depiction of our model as well as a schematic of the trainable network architecture that was applied.



Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 224, 224, 3)]	0
sequential (Sequential)	(None, 224, 224, 3)	0
tf.math.truediv (TFOpLambda)	(None, 224, 224, 3)	0
tf.math.subtract (TFOpLambda	(None, 224, 224, 3)	0
mobilenetv2_1.00_224 (Functi	(None, 7, 7, 1280)	2257984
global_average_pooling2d (Gl	(None, 1280)	0
dropout (Dropout)	(None, 1280)	0
dense (Dense)	(None, 1)	1281

Total params: 2,259,265
Trainable params: 1,862,721
Non-trainable params: 396,544

Figure 1: Architecture of Trainable Neural Network Layers

Both the experimental and baseline models freeze a majority of MobileNet V2's layers to preserve the classification architecture and build upon the network's ability to detect edges, objects, and groups of objects. This model consists of 16 repeated blocks which contain a 2-dimension convolutional layer, batch normalization, and a RELU activation layer to contend with nonlinearity.

Categorical cross-entropy loss was used in the baseline model to classify levels of students' cognitive demand, where level 1 indicated the least demanding activity and level 4 indicated the most demanding activity. Binary cross-entropy loss was used in the experimental model to categorize each of the five feature classes, as we aimed to assess whether each student behavior was present. Finally, we implemented a support vector machine classifier to transform intermediate binary feature predictions to cognitive demand scores on testing data. SVMs potentially work well with smaller datasets, such as ours, and are ideal for categorizing data into linear classes [24].

4. Data and Preprocessing

The data were collected from two eighth grade mathematics classrooms that focused on building students' capacities for cognitively demanding work through engagement with mathematical tasks. Four 30-minute video recordings were taken in Spring 2017 that featured students solving "The Washing Machine Problem" with Desmos, a dynamic graphing calculator application. The video recordings were rated for cognitive demand on a 1-4 scale called the Instructional Quality of Assessment Rubric [8, 9], a research-backed tool for rating cognitive demand of students' mathematical activity. Demand was rated at the level of the entire student group, and importantly, cognitive demand ratings were not uniformly distributed between the 1-4 scale. This is to be expected, because students were more likely to achieve moderate cognitive demand throughout the task (level 2 or 3) than extreme ratings (level 1 or 4). Initial ratings were assigned in Winter 2021 by two mathematics education experts (Delaney & Kinsey) who reached 87.9% inter-rater agreement. 87.9%, classified as "very good agreement" [21] serves as the upper accuracy threshold for human performance on this task.

The video recordings were spliced into still images taken at 1-second intervals, and we assigned 1-4 cognitive demand labels to each image from Delaney and Kinsey's ratings. We then hand-labeled each image in the five binary classes according to the following schematic:

- **Computer class:** the image received a "1" if students were using the computer to solve the task, and a "0" otherwise.
- **Leaning class*:*** the image received a "1" if more than one student was leaning into the center of the table to collaborate with the group, and a "0" otherwise.
- **Pointing class*:*** the image received a "1" if one or more students were visibly pointing to gesturing to the task or computer, and a "0" otherwise.
- **Teacher class:** the image received a "1" if the teacher and students were conversing with one another at the same table, and a "0" otherwise.
- **Writing class:** the image received a "1" if one or more students were writing on the task card, and a "0" otherwise.

The five binary classes were generated based on our hypothesized relationship of each indicator to students' cognitive demand. Prior research has demonstrated that students' use of computational tools to assist with problem solving can either raise or lower cognitive demand, contingent upon how students use it [13]. Similarly, conferring with a teacher should increase cognitive demand, as teachers may draw students' attention to cognitively demanding features of the task during small-group interactions [15]. Finally, the ways in which students work collaboratively and use one another as resources may increase cognitive demand, as visually indicated through pointing, collective writing, and leaning in toward the "middle space" [22].

In total, the data set contained 2000 images distributed uniformly across the four classroom video recordings. Each image was rescaled to 224 by 224 pixels to accommodate the maximum weight size of MobileNet V2.

5. Results

5.1. Experiment 1: Training the Baseline Model

The first model classifies cognitive demand from images alone. We expected the accuracy of this model to be relatively low because, in comparison with the five binary class indicators in the experimental model, the baseline model's feature space was high-dimensional. We experimented with various combinations of hyperparameters: learning rates, batch sizes,

epochs, and optimizers to investigate the training accuracy of the baseline. We aimed to achieve accuracy of around 25%, the expected accuracy that would be generated from a balanced cognitive demand distribution over the four levels. Table 1 shows the training and validation accuracy as we tuned hyperparameters over 20 epochs.

Table 1
Hyperparameter Tuning with Baseline Cognitive Demand Model

	Learn Rate: .001 Batch Size: 32	Learn Rate: .0001 Batch Size: 32	Learn Rate: .0001 Batch Size: 16	Learn Rate: .00001 Batch Size: 16
Categorical Cross-Entropy	Train_Acc: **7.2%** Val_Acc: **4.8%**	Train_Acc: **8.4%** Val_Acc: **4.8%**	Train_Acc: **8.8%** Val_Acc: **3.2%**	Train_Acc: **24.6%** Val_Acc: **29.8%**
Sparse Categorical Cross-Entropy	Train_Acc: **3.5%** Val_Acc: **2.1%**	Train_Acc: **3.8%** Val_Acc: **2.1%**	Train_Acc: **5.4%** Val_Acc: **3.3%**	Train_Acc: **9.1%** Val_Acc: **6.7%**

Categorical Cross-Entropy Loss was selected to model the data with a batch size of 16 and a learning rate of 0.0001. Conceptually, we anticipated that Sparse Categorical Cross-Entropy Loss would have been a better fit because it is designed for integer input; however, this was not the case during training. The final combination of hyperparameters caused the training accuracy to increase quickly, then level off after approximately 10 epochs.

5.2. Experiment 2: Training the Experimental Model using Binary Behavioral Proxies

5.2.1. Phase 1: Binary classification using MobileNet V2

The experimental model sought to improve cognitive demand predictions by first identifying five binary student behaviors that might impact demand, then applying predicted binary class labels for the behaviors to testing data for prediction. Each of the five binary class sub-models were trained using MobileNet V2 with ImageNet weights. Data were split into 80% training, 10% validation, and 10% testing. We ensured that both the validation and testing sets contained all four cognitive demand levels.

Hyperparameters were tuned for each class separately, although many classes showed optimal training accuracy using similar inputs. Similar to Experiment 1, all binary classes were tuned for learning rate, number of epochs, loss optimization function, and batch size. The ADAM optimizer was used in all classes because it handled the noisy classroom data well, an important consideration for localization of class features.

We anticipated that the Teacher and Computer classes would achieve high training accuracy faster, because there was less ambiguity in labeling those classes compared to Writing, Leaning, and Pointing. We hypothesized that the latter classes would take longer to converge because they were based on pose estimation, and were more likely to vary per student. For example, we associated students' elbows on the table with the "leaning" class, but since not every student in the group need to have exhibited the "leaning" action in order for the image to be classified as "leaning," this nuance may have been difficult for the model to detect. Table 2 shows the training and validation accuracies as we tuned hyperparameters for all five student behavioral proxies trained over 50 epochs.

Table 2
Training and Validation History of Five Binary Behavior Proxy Classes

	Learn Rate = .0005 Batch Size = 32	Learn Rate = .001 Batch Size = 16	Learn Rate = .0001 Batch Size = 16	Learn Rate = .0005 Batch Size = 16	Learn Rate = .00065 Batch Size = 16
Computer Class	Train_Acc: **96.4%** Val_Acc: **94.8%**	Train_Acc: **55.7%** Val_Acc: **53.0%**	Train_Acc: **96%** Val_Acc: **97.5%**	Train_Acc: **99%** Val_Acc: **98.5%**	Not tested
Leaning Class	Train_Acc: **77.2%** Val_Acc: **68.3%**	Train_Acc: **72.3%** Val_Acc: **69.4%**	Train_Acc: **74.3%** Val_Acc: **68.2%**	Train_Acc: **85.7%** Val_Acc: **70.9%**	Train_Acc: **85.6%** Val_Acc: **76%**
Pointing Class	Train_Acc: **76.3%** Val_Acc: **60.2%**	Train_Acc: **74.1%** Val_Acc: **63.9%**	Train_Acc: **78.7%** Val_Acc: **64.2%**	Train_Acc: **90.2%** Val_Acc: **67.8%**	Not tested
Teacher Class	Train_Acc: **81.8%** Val_Acc: **66.5%**	Train_Acc: **79.3%** Val_Acc: **64.2%**	Train_Acc: **84.1%** Val_Acc: **69.4%**	Train_Acc: **88.5%** Val_Acc: **75.2%**	Train_Acc: **97.5%** Val_Acc: **81.6%**
Writing Class	Train_Acc: **78.3%** Val_Acc: **76.7%**	Train_Acc: **68.4%** Val_Acc: **66.6%**	Train_Acc: **76.3%** Val_Acc: **71.2%**	Train_Acc: **75.2%** Val_Acc: **73.4%**	Train_Acc: **88.0%** Val_Acc: **77.1%**

The models performed best given low learning rates, smaller batch sizes, and longer training duration to achieve high training and validation accuracy. This is not surprising, given the localization required for the network to learn and classify each of the five feature behaviors. Models were trained until each obtained a training accuracy over 85%, a value similar to human accuracy applied for the original cognitive demand labels. In the event that multiple models fit this criterion, the model whose parameters the highest validation accuracy was selected. The final selected hyperparameters are highlighted in yellow in Table 2.

Figure 2 illustrates one example of our error analysis per each binary class. As we interrogated the nuances these errors, it appeared that some class models learned to identify subtleties in the data better than others. For example, the highly-accurate Computer class differentiated between closed computers and open computers after 50 epochs of training. In a majority of images, the Teacher class teased apart differences between the teacher's presence at the table versus the teacher performing other actions in the image background. Classification errors occurred when the teacher was only partially visible in the image, which made sense, as teachers were not actively monitoring their body position and placement during the original video recordings. Errors in the Pointing, Writing, and Leaning classes occurred when the students did not clearly demonstrate the intended action; for example, when the point was blurry or incomplete, when only one student was writing or leaning, or when the leaning action was subtle.

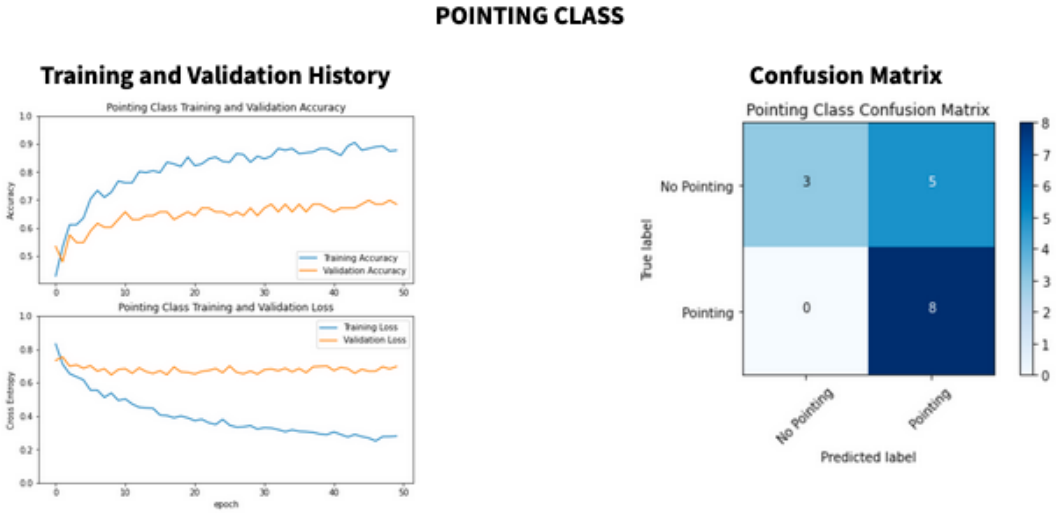


Figure 2: One example of error analysis in the Pointing Class's training data

5.2.2. Phase 2: Labeling Cognitive Demand using Trained Multiclass Models and a Support Vector Machine

Once the binary multiclass models were established, we utilized a small test set of data (n = 40 images, 10 per cognitive demand class) to examine the Computer, Leaning, Pointing, Teacher, and Writing models' abilities to (1) correctly predict the five binary classes of students' behaviors in the test set and (2) calculate cognitive demand based on labels generated by the five models. We tested both a linear and a generalized support vector machine to predict final cognitive demand labels. Regularization parameters were tuned in both models (e.g., the kernel and gamma parameters in the generalized SVM, and the loss function in the linear SVM). Figure 3 summarizes the results for both classifiers and provides a confusion matrix to summarize classification errors.

	Reg_rate: 1.0	Reg_rate: 10.0	Reg_rate: 40.0	Reg_rate: 50.0	Reg_rate: 60.0
General SVM	Accuracy: **0.60**	Accuracy: **0.575**	Accuracy: **0.625**	Accuracy: **0.65**	Accuracy: **0.65**
Linear SVM	Accuracy: **0.50**	Accuracy: **0.525**	Accuracy: **0.525**	Accuracy: **0.525**	Accuracy: **0.525**

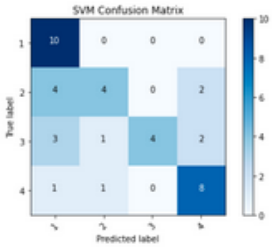


Figure 3: Testing accuracy for cognitive demand classification with a support vector machine

After tuning the regularization rate and aforementioned hyperparameters, it did not appear that the models' predictive accuracies for cognitive demand varied substantially. The general SVM classifier was the better overall choice because it improved cognitive demand classification from the baseline model (55.7%), although it does not surpass human performance (87.9%). This result is not surprising, because cognitive demand is an abstract concept that was previously rated by human experts using both speech and visual cues. However, the drastic improvements in cognitive demand classification from the baseline model validate our current approach despite the relatively small size of the data set.

6. Discussion

The experimental two-phase model did not approach human-level performance, but showed both improvements from the baseline model and promise for future work. Compared to the baseline model, the SVM classifier performed better than unsupervised classification. This result presents a case for weak supervision to be used when training data are identified, sorted, and labeled, which could draw upon the expertise of more teachers in future iterations of this work. We hypothesize that teachers' involvement during data labeling would offer improvements to the model due to their developed practices in interpreting students' behaviors in their day-to-day experiences.

Future developments in this study will increase the sample sizes and apply data augmentation to re-examine outcomes. Increasing the sample size will improve predictive stability in the binary models, particularly the Pointing class, which contained a smaller proportion of positive cases with respect to the others. Future data augmentations to be tested include varying the brightness in classroom photos, rotating classroom images, and including more classroom images with noisy features (for example, the presence of additional individuals in the image frame who are not the teacher). Although MobileNet V2 appeared to be a suitable

classifier for binary class inputs, it was likely not the best choice for the baseline categorical model. Other neural networks, such as VGG net, may have produced better transfer learning [12], and will be tested in future iterations of this work.

A key long-term goal of this project is to build toward a cognitive demand classification tool that can be used to support and empower teachers' professional learning. By analyzing their students' variations in cognitive demand throughout a mathematical task, teachers may better understand the range and variation in students' enacted demand, and adjust their future instructional practices accordingly. Such a tool may be useful in teachers' video clubs [25], a form professional development activity designed to hone teachers' noticing and inquiry of student behavior. By supplying teachers with a cognitive demand classifier, teachers may attend to student behavioral features that impact cognitive demand more frequently, and adjust their practices in response. We aim to test this theory in future iterations of this work.

7. Acknowledgements

We thank the Amir Lopatin Fellowship committee, which supplied funding to this project in support of its potential contributions to the learning sciences. This study emerged under the mentorship of Dr. Nick Haber and Dr. Ranjay Krishna during Stanford's Spring 2021 academic quarter. It was originally submitted as the project component of their CS 432 and CS 231n courses, respectively. We thank Gina Kinsey for her work in hand-labeling the original cognitive demand data and Jagriti Agrawal for her contributions during the initial conception and modeling in this study.

8. References

[1] Agrawal, P., Girshick, R., & Malik, J. (2014, September). Analyzing the performance of multilayer neural networks for object recognition. In *European conference on computer vision* (pp. 329-344). Springer, Cham.

[2] Ngoc Anh, B., Tung Son, N., Truong Lam, P., Le Chi, P., Huu Tuan, N., Cong Dat, N., ... & Van Dinh, T. (2019). A computer-vision based application for student behavior monitoring in classroom. *Applied Sciences*, *9*(22), 4729.

[3] Ansar, W., Shahid, A. R., Raza, B., & Dar, A. H. (2020, March). Breast cancer detection and localization using MobileNet based transfer learning for mammograms. In *International symposium on intelligent computing systems* (pp. 11-21). Springer, Cham.

[4] Azizpour, H., Sharif Razavian, A., Sullivan, J., Maki, A., & Carlsson, S. (2015). From generic to specific deep representations for visual recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 36-45).

[5] Bengio, Y., Courville, A. C., & Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, *1*, 2012.

[6] Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., & Oord, A. V. D. (2020). Are we done with ImageNet?. *arXiv preprint arXiv:2006.07159*.

[7] Bosch, N., Mills, C., Wammes, J. D., & Smilek, D. (2018, June). Quantifying classroom instructor dynamics with computer vision. In *International Conference on Artificial Intelligence in Education* (pp. 30-42). Springer, Cham.

[8] Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal*, *113*(1), 76-104.

[9] Boston, M., & Wolf, M. K. (2006). Assessing Academic Rigor in Mathematics Instruction: The Development of the Instructional Quality Assessment Toolkit. CSE Technical Report 672. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

[10] Canedo, D., Trifan, A., & Neves, A. J. (2018, June). Monitoring students' attention in a classroom through computer vision. In *International Conference on Practical Applications of Agents and Multi-Agent Systems* (pp. 371-378). Springer, Cham.
[11] Chen, J., Zhang, D., Suzauddola, M., Nanehkaran, Y. A., & Sun, Y. (2021). Identification of plant disease images via a squeeze- and- excitation MobileNet model and twice transfer learning. *IET Image Processing*, *15*(5), 1115-1127.
[12] Cheng, P. M., & Malhi, H. S. (2017). Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *Journal of digital imaging*, *30*(2), 234-243.
[13] Common Core State Standards Initiative. (2010). Common Core State Standards for mathematics. Retrieved from http://www.corestandards .org/assets/CCSSI_Math%20Standards.pdf

[13 →14] Doyle, W. (1983). Academic work. *Review of educational research*, *53*(2), 159-199.
[15] Franke, M. L., Turrou, A. C., Webb, N. M., Ing, M., Wong, J., Shin, N., & Fernandez, C. (2015). Student engagement with others' mathematical ideas: The role of teacher invitation and support moves. *The Elementary School Journal*, *116*(1), 126-148.

[14 → 16] Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., & Feris, R. (2019). Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4805-4814).
[15 → 17] Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes ImageNet good for transfer learning?. *arXiv preprint arXiv:1608.08614*.
[16 → 18] Joulin, A., Van Der Maaten, L., Jabri, A., & Vasilache, N. (2016, October). Learning visual features from large weakly supervised data. In *European Conference on Computer Vision* (pp. 67-84). Springer, Cham.
[17 → 19] Kaput, J. J. (1992). Technology and mathematics education. *Handbook of research on mathematics teaching and learning*, *515*, 556.
[18 → 20] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
[19 → 21] Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.
[22] Lotan, R. A. (1994). Talking and Working Together: Conditions for Learning in Complex Instruction.
[20 → 23] Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International journal of remote sensing*, *26*(5), 1007-1011.
[21 → 24] Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019, May). Do ImageNet classifiers generalize to ImageNet?. In *International Conference on Machine Learning* (pp. 5389-5400). PMLR.
[22 → 25] Gamoran Sherin, M., & Van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of teacher education*, *60*(1), 20-37.
[23 → 26] Suresh, A., Sumner, T., Huang, I., Jacobs, J., Foland, B., & Ward, W. (2018, December). Using deep learning to automatically detect talk moves in teachers' mathematics lessons. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 5445-5447). IEEE.
[24 → 27] Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching

and learning in a reform mathematics project. *Educational Research and Evaluation*, *2*(1), 50-80.

[25 →28] Tekkumru Kisa, M., & Stein, M. K. (2015). "Learning to see teaching in new ways: A foundation for maintaining cognitive demand." *American Educational Research Journal* 52.1:105-136.

[26 →29] Thille, C., & Zimmaro, D. (2017). Incorporating learning analytics in the classroom. *New Directions for Higher Education*, *2017*(179), 19-31.

[22 →30] Xiang, Q., Wang, X., Li, R., Zhang, G., Lai, J., & Hu, Q. (2019, October). Fruit image classification based on Mobilenetv2 with transfer learning technique. In *Proceedings of the 3rd International Conference on Computer Science and Application Engineering* (pp. 1-7).

[27→31] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *arXiv preprint arXiv:1411.1792*.