Tell Me More: Digital Eyes to the Physical World for Early Childhood Learning

Vijay Ekambaram*, Ruhi Sharma Mittal*, Prasenjit Dey,
Ravi Kokku, Aditya K Sinha, Satya V Nitta
IBM Research
vijaye12@in.ibm.com, ruhisharma@in.ibm.com, prasenjit.dey@in.ibm.com
rkokku@us.ibm.com, adksinha@in.ibm.com, svn@us.ibm.com

ABSTRACT

Children are inherently curious and rapidly learn a number of things from the physical environments they live in, including rich vocabulary. An effective way of building vocabulary is for the child to actually interact with physical objects in their surroundings and learn in their context [17]. Enabling effective learning from the physical world with digital technologies is, however, challenging. Specifically, a critical technology component for physical-digital interaction is visual recognition. The recognition accuracy provided by state-of-the-art computer vision services is not sufficient for use in Early Childhood Learning (ECL); without high (near 100%) recognition accuracy of objects in context, learners may be presented with wrongly contextualized content and concepts, thereby making the learning solutions ineffective and un-adoptable. In this paper, we present a holistic visual recognition system for ECL physical-digital interaction that improves recognition accuracy levels using (a) domain restriction, (b) multi-modal fusion of contextual information, and (c) semi-automated feedback with different gaming scenarios for right object-tag identification & classifier re-training. We evaluate the system with a group of 12 children in the age group of 3-5 years and show how these new systems can combine existing APIs and techniques in interesting ways to greatly improve accuracies, and hence make such new learning experiences possible.

1. INTRODUCTION

Children learn a lot from the physical environment they live in. One of the important aspects of early childhood learning is vocabulary building, which happens to a substantial extent in the physical environment they grow up in [17]. Studies have shown that failure to develop sufficient vocabulary at an early age affects a child's reading comprehension and hence their ability to understand other important concepts that may define their academic success in the future. It is also evident from a study that failure to expose a child to sufficient number of words by the age of three years leads to a 30 million word gap between kids who have been exposed to a lot of quality conversations, versus the ones that have not been exposed as much [13].

Vocabulary building has been a theme for early childhood learning and is closely associated with its context in the physical world. The exploration of physical surroundings of the child triggers new vocabulary and vice-versa. Relating physical world objects and concepts, to digital world content requires seamless flow of information. Increasingly, availability of cheap sensors such as camera, microphone etc. on connected devices enable capture of physical world information and context, and translate them to personalized digital learning.

An envisioned system uses mobile devices to take pictures of the child's physical surroundings and make the best sense out of the picture. This is then translated to a learning session where the child is taught about the object in focus, its relation to other objects, its pronunciation, it's multiple representations, etc. Recognition of pictures for teaching a child requires high recognition accuracy. In-the-wild image recognition accuracies are in general low, especially for images taken with mobile devices. Moreover, pictures taken by a child is even more challenging given the shake, blur, lighting issues, pose etc. that come with it.

To this end, in this paper, we take a holistic approach of recognition-in-context using a combination of (a) domain restriction, (b) multi-modal fusion of contextual information, and (c) gamified disambiguation and classifier re-training using child-in-the-loop. Specifically, we use object recognition results from a custom-trained (with images from restricted domains) vision classifier, and combine them with information from the domain knowledge that is available whenever a new domain of words is taught to a child in the classroom or at home. We use a new voting based multimodal classifier fusion algorithm to disambiguate the results of vision classifier, with results from multiple NLP classifiers, for better accuracy. We show that using such a framework, we can attain levels of accuracy that can make a large majority of the physical-digital interaction experiences fruitful to the child, and also get useful feedback from the child at a low cognitive load to enable the system to retrain the classifier and improve accuracy. We tested our system with a group of 12 children in the age group of 3-5 years and show that children can play an image disambiguation game (that allows the child to verify what class label has actually been identified by the system) very easily with graceful degradation of performance on difficult images. In most cases, multi-modal context disambiguation improves object recognition accuracy significantly, and hence the human disambiguation step remains limited to one or two rounds, which ensures the child's continuing interest in the games and learning activities. The system learns from the child feedback, and the child in turn feels engaged to enable the system to learn over time. The nuggets of information made available about the object in focus at the end of playing a game were also found to be very engaging by the child.

In summary, this paper makes the following contributions:

- We take a holistic approach to address the challenges with

*these authors contributed equally to this work

automatic visual recognition for physical digital interaction to enable early childhood learning in context. Our three-stage approach includes (a) domain restriction, (b) contextual disambiguation and (c) gamified human disambiguation, which enables a platform for building a variety of early childhood learning applications with physical-digital interaction.

- We propose a novel re-ranking algorithm that uses the notion of strong vouching to re-order the output labels of a vision classifier based on strong supporting evidence provided by the additional context from semantic representation models in NLP, namely GloVe [6], Word2vec [11] and ConceptNet [5] (which can be textual cues in the form of classroom and curriculum context, domain focus, conversational input and clues, etc.). Note that, we use the terms "re-order" and "re-rank" interchangeably throughout this paper.

- We evaluate a simple disambiguation game for children to choose the right label from the Top-K labels given out by the system. Through an usability study with 12 children, we make the case that engaging user experiences can indeed be developed to bridge the gap between automatic visual recognition accuracies and the requirement of high accuracy for meaningful learning activities.

2. MOTIVATION AND RELATED WORK

Early childhood learning applications with physical-digital interaction fall into two categories: (i) *Application-initiated activities:* In this category, the child is given a context by the application and is required to find relevant physical object and take a picture [2]. For example, the application may prompt the child to take a picture of "something that we sit on", "a fruit", "something that can be used to cut paper", etc. (ii) *Child-initiated activities:* In this category, the child takes a picture of an object and intends to know what it is, where it comes from, other examples of the same type of objects, etc. For example, the child may take a picture of a new gadget or machine found in school, a plant or a leaf or a flower, etc. and wants to know more about them.

In each of these categories, the application is required to identify what the object is with Top-1 accuracy (i.e. a vision recognition solution should emit the right label at the top with high confidence). While a lot of advancement has been made in the improvement of accuracy of vision classifiers, Top-1 accuracy levels are still relatively low, although Top-5 accuracy levels (i.e. the right label is one of the top 5 labels emitted) are more reasonable. Nevertheless, the goal is to be able to work with the Top-5 list, and using the techniques described earlier, push the Top-1 accuracy to acceptable levels for a better interaction.

2.1 Vision Recognition Accuracy

To understand the efficacy of state-of-the-art solutions quantitatively, we experimented with two deep convolution neural networks (Baseline Model 1: VGGNet [18] and Baseline Model 2: Inception V3 [19]). Inception V3 has been found to have 21.2% top-1 error rate for ILSVRC 2012 classification challenge validation set [8]. Even in experiments where baseline models were custom trained with 300 training images per class and tested with images taken from iPad, we observed low Top-1 accuracy (of 72.6% in Baseline Model 1 and 79.1% in Baseline Model 2); i.e. one in about four images will be wrongly labeled. Even the Top-5 accuracy is 88.05% in Baseline Model 1 and 89.3% in Baseline Model 2. We also trained the Baseline models with the complete Imagenet[8] images for the considered classes and we observed <1% improvement. Further, when multiple objects are present in the image frame, the Top-1 accuracy degrades further (38.2% in Baseline Model 1 and 44.5% in Baseline Model 2 for 2 objects in a frame), and so does Top-5 accuracy (of 77.9% in Baseline Model 1 and 85.6% in Baseline Model 2). Note that this could be a common scenario with children taking pictures, in which multiple objects get captured in a single image frame. Observe that recent Augment Reality (AR) Applications such as Blippar [4], Layer [9], Aurasma [3] rely on similar vision recognition task, and hence run into similar inaccuracies in uncontrolled settings. While adult users of such applications may be tolerant to inaccuracies of the application, children may get disengaged when the system detects something wrongly or is unable to detect at all.

2.2 Multi-modal Information Fusion

Using additional information to identify the objects holds promise in imporving the accuracy of vision recognition. For instance, several past works ([22], [14], [15]) improve the image classification output based on the text features derived from the image. Specifically, authors in [20] propose techniques that train the model specifically with images that contain text, for efficient extraction of text and image features from the image. They also propose fusion techniques to merge these features for improving image recognition accuracies. While this may be possible in some scenarios, the application's accuracy will remain a challenge when such textual information embedded in the image is not present. Several works in literature propose indexing of images based on text annotations for efficient image search. [12] surveys and consolidates various approaches related to efficient image retrieval system based on text annotations. Likewise, [21] proposes techniques to label images based on image similarity concepts. These works are complementary, and do not address the problem of correctly determining the labels right when a picture is taken based on a context.

In summary, the early childhood learning scanarios require a holistic solution that leverages the state-of-the-art vision recognition solutions, but goes beyond in improving the detection accuracy of the image captured to make engaging applications for children. We describe one such holistic solution next.

3. PROPOSED APPROACH

Our goal is to enable a holistic solution for applications to provide as input an image taken by a child, and emit as output the final label that should be used as an *index* into the relevant learning content. A high level overview of our solution is depicted in Figure. 1. In one of the envisioned applications built for physical-digital interaction, a child takes a picture that is sent as input to the proposed ECL Image Recognition (ECL-IR) Module that emits the correct label of the image by applying the following three stages: (i) Stage 1: Domain Specific Customized Training (which improves Top-K accuracy), (ii) Stage 2: Domain Knowledge (DK) based disambiguation and reordering (which improves Top-1 accuracy) and (iii) Stage 3: Human Disambiguation game (confirmation step). We now discuss each of these stages in detail.

3.1 Stage 1: Domain Specific Customized Training of Baseline Models

The first stage of our solution strives to improve the Top-K accuracy of the vision classifiers by constraining the domain of child learning in which they are applied. In order to achieve this, we perform custom training of the baseline models with domain-specific data sets. This step is very commonly applied in most of the vision recognition use-cases for improving the Top-K accuracy and several reported statistics indicate good Top-K accuracy improvements through custom training. For example current state-of-art vision classifier [19] reports 94.6% Top-5 accuracy on ILSVRC 2012 classification challenge validation set. However, even this state-of-art vision classifier reports 21.2% Top-1 error rate on the same validation set. In the next section, we discuss how ECL-IR module improves Top-1 accuracy through contextualized reordering (Stage 2).

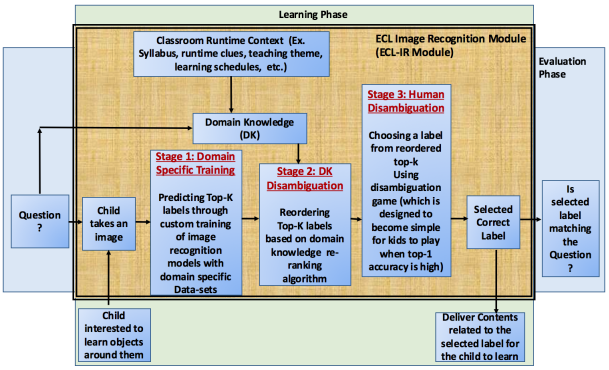


Figure 1: High Level Solution Overview

Algorithm 1: Algorithm to calculate DK Correlation Score

Input: Label, Domain Knowledge text (DK)
Output: DK Correlation Score (i.e. Semantic correlation between DK and Label)

1 For every word in DK, fetch its corresponding N-dimensional semantic vector from the semantic representation model.
2 Representation(DK) <– Compose N-dimensional vector for the complete DK by combining word level vectors to a phrase level vector using linear average technique
3 Representation(Label) <– Fetch N-dimensional vector for the label from the semantic representation model
4 DK Correlation Score = Cosine Distance between Representation(DK) and Representation(Label)
5 **return** *DK Correlation Score*

3.2 Stage 2: Domain Knowledge based Disambiguation and Reordering

In this section, we propose to improve the Top-1 accuracy through intelligent reordering of the Top-5 labels from the vision classifier. In-order to achieve this, we leverage the domain knowledge associated with the teaching activity as a second source of information to re-order the Top-5 output labels. Domain Knowledge refers to the classroom learning context (derived from teacher's current syllabus, teaching themes, object related clues, collaborative clues) based on which the learning activity is conducted. Note that the Domain Knowledge could be a *word* or a *phrase* too. We now discuss various important aspects of this stage in detail.

Enabling Semantic Capability. Domain Knowledge is a text representation of the intent or activity derived from the classroom context. However, same intent or information could be conveyed through different keywords, and hence traditional bag-of-word approaches [23] will not solve the problem in our use-cases. We leverage the support of semantic representations (i.e. distributed word representation [16]) of words for enabling keyword independent re-ranking algorithm. In distributed word representation, words are represented as N-dimensional vectors such that distance between them capture semantic information. There are various pre-trained semantic representation models (also called word embedding models such as Word2Vec [11], GloVe [6]) available which enable semantic comparison of words. Likewise, there is also ConceptNet [5] which is a multilingual knowledge base, representing words and phrases that people use and the common-sense relationships between them. This paper leverages these existing works to achieve an effective re-ranking of the output label-set with semantic capability.

Existing Approach Results. One naive way to approach the problem of re-ranking is to find the DK Correlation Score (DK-CS) using Algorithm. 1 and re-rank the Top-5 labels in descending order of their DK-CS. However, this approach has strong bias towards the semantic representation output and completely ignores the ranking that is produced by the vision classifier.

Other fusion approaches that have been tried are combining one or more of the classifier outputs (i) Word2Vec (S1), (ii) GloVe (S2), (iii) ConceptNet (S3), (iv) Vision (S4) in different ways. The most common are the product rule and the weighted average rule where the confidence scores are combined by computing either a product of them or a weighted sum of them. The improvement in Top-1 accuracy of such combinations varies from -11% to 6%. We observe that the Top-1 accuracy of the system did not increase significantly and in many cases Top-1 accuracy of the system dropped after re-ranking as compared to the original list. The reason being the need for proper and more efficient resolution of conflicts between DK-CS wins vs. vision confidence score wins. In the next section, we explain the proposed novel re-ranking algorithm which highly improves the Top-1 accuracy of the system by effectively resolving the conflicts between DK-CS and vision rankings.

Proposed Re-Ranking Approach. In our proposed approach, we fuse the inferences from various semantic models and vision model using *Majority-Win Strong Vouching algorithm* for re-ordering the Top-5 output list. There are two important aspects of this approach: (i) Strong Vouching of Semantic Models, (ii) Majority Voting across Semantic Models.

Strong Vouching of Semantic Models: As discussed earlier, the reason for failure of the traditional fusion approaches is the need for efficient resolution of conflicts between the semantic model ranks and the vision model ranks. Let us understand this problem through 2 example scenarios. (i) Scenario 1: Top-1 prediction is "orange", Top-2 prediction is "apple", domain Knowledge is "fruits"; (ii) Scenario 2: Top-1 prediction is "orange", Top-2 prediction is "apple", domain knowledge is "red fruits". In the first scenario, since the domain knowledge is semantically correlated towards both Top-1 and Top-2 predicted labels, system should maintain the same order as predicted by the vision model. However, in the second scenario, since the domain knowledge (i.e. "red fruits") is highly correlated towards Top-2 (i.e. "apple")as compared to Top-1(i.e. "orange"), system should swap the order of Top-1 and Top-2 labels. It turns out that just having a higher DK-CS to swap the labels is not enough. We show that DK-CS of one label (label-1) should override the other label (label-2) by a specific threshold value to indicate that label-1 is semantically more correlated with as compared to label-2 and hence effect a swap against the vision rank. Through empirical analysis in Section. 4.2, we show that, in the context of reordering Top-K labels, if normalized DK-CS of a label is greater than the other label by a value equal to $1/k$ (threshold value), then the former label is more semantically correlated with domain knowledge as compared to the latter.

Majority Voting across Semantic Models: As mentioned before, many semantic models exist in the literature and each of them are trained on various data-sets. Therefore, it is not necessary that the strong vouching behavior of all these semantic models to be same. In order to resolve this, our approach considers multiple semantic models together (such as GloVe, Word2Vec and ConceptNet) and enables swapping of i-th label with j-th label (i<j) in the Top-K output list only when majority of semantic models are strongly vouching that j-th label is more correlated with DK as compared to the i-th label. This makes the system more intelligent in resolving across

semantic models as well as resolving conflicts across DK correlation score wins vs. vision confidence score wins. Algorithm. 2 explains the overall flow of the proposed re-ranking algorithm.

Algorithm 2: Fusion based on Majority Win Strong Vouching Concept

Input: Top-K output label from image recognition model, Domain Knowledge(DK)
Output: Reordered Top-K output label list

```
Sort Top-K labels based on vision confidence score
Re-rank the Top-K label by sorting using the following compare logic
Compare logic (i-th label, j-th label, DK): begin
    [Note] i-th label precedes j-th label in the ranked Top-K list.
    X1 = Total number of semantic models strongly-vouching for j-th label as compared to i-th label
    X2 = Total number of semantic models strongly-vouching for i-th label as compared to j-th label
    if X1>X2 then
        swap i-th label and j-th label in the Ranked Top-K list
    else
        Maintain the same order of i-th label and j-th label
return Re-Ranked Top-K List
```

3.3 Stage 3: Human Disambiguation Game

It is important to note that, due to limitation of existing state-of-art vision models, though we achieve effective improvements, we never reach an accuracy of 100%. Even after effective custom training and DK based Top-K re-ranking, accuracy of the system is not 100% (though high improvements are observed). So, there has to be a confirmation step involving human-in-loop to confirm whether the predicted label is the right label to prevent teaching wrong objectives. Since we are dealing with Kids, this step has to be extremely light, simple, and also engaging for the Kids so that, they do not feel any extra cognitive load. In this section, we propose a simple disambiguation game which is designed in a way that, (i) Kids easily play with it correctly, (ii) Kids interaction with the game highly reduces when Top-1 accuracy of the system is high. Through enhancements as explained in previous sections, we make vision model to reach high Top-1 accuracy which in-turn reduces the Kids interactions in the disambiguation game, thereby reducing the overall cognitive overload to the Kids.

Our system leverages image matching for the disambiguation game. Re-ranked Top-K list (which is the output from Stage 2) is fed as input to the disambiguation game. This game is depicted in Figure. 2 renders reference images of the label (with possible variants of a same object) one by one in the order of the re-ranked list and asks the Kid to select the image, if it looks similar to the object clicked (through camera). If not, system show the next reference image and continues till all K labels are rendered. Since the input to the game is a re-ranked Top-K list (which has high Top-1 accuracy), Kid has high chances of encountering the right image in the first or second step itself, thus reducing the cognitive load of the kid to traverse till the end. Usability Guidelines [10] [1] for Child based Apps suggest large on-screen elements which are well spatially separated for Kids to easily interact with them. So, based on the display size of the form-factor, system could configure the no of images to be rendered in one step/cycle. Through usability study with 15 Kids, we show that Kids are able to easily play image similarity based disambiguation games. In scenarios when the right label is not in the predicted Top-K labels, system executes the exit scenarios as configured. Few possible exit scenarios could be: (i) Continue the game with other labels in the learning vocabulary set in the sorted order of DK, (ii) Request for teacher intervention, etc.

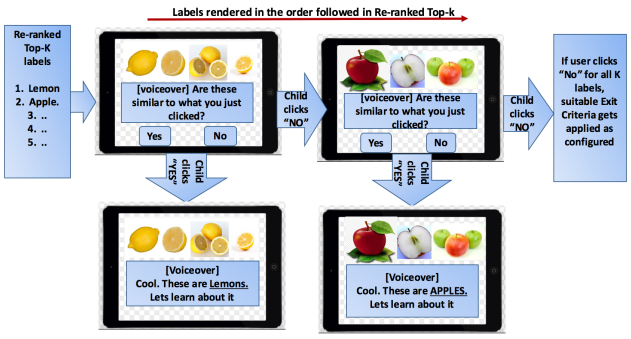


Figure 2: Basic Disambiguation Game

4. EVALUATION

We present here the experimental setup and results of improvement in the vision classifier results achieved by the re-ordering approach. We then explain and present the results of the empirical analysis to determine the value of threshold for strong vouching of the semantic models. To show that our approach is independent of domain knowledge, test set, training class set, and baseline image classification models (generality of approach), we performed various experiments as explained in following subsections. Later in this section, we present the usability study and inferences from the study conducted with a group of 12 children in the age group of 3-5 years.

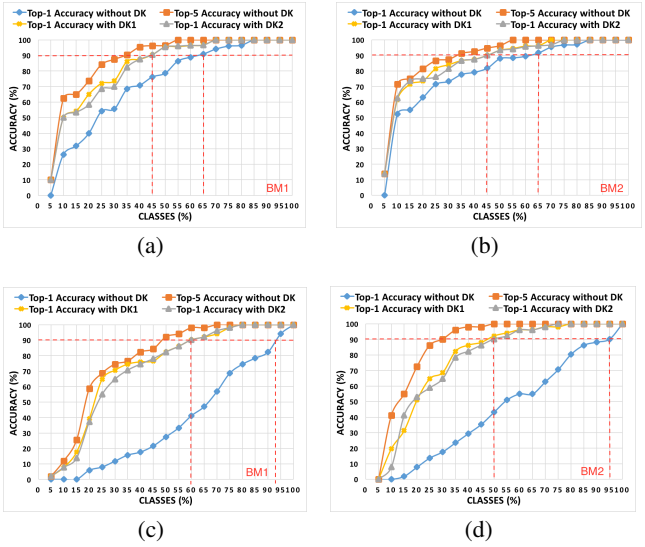
Datasets: The training dataset includes images from Imagenet [8]. We used 52 classes and approximately 400 images per class for training. These 52 selected classes are objects commonly used in early childhood learning, for example, apple, car, book, and violin, etc. The test datasets include real images taken from mobile phones and tablets. The test dataset I includes 1K images where single object (from training set) is present in an image frame. The test dataset II includes 2.6K images where two objects (from training set) are present in an image frame. All the experiments were performed using two baseline image classification models: (i) Baseline Model 1 (BM1): Model based on VGGNet architecture [18], (ii) Baseline Model 2 (BM2): Model based on Inception-V3 architecture [19].

Domain Knowledge: During all the experiments, we used two different domain knowledge (DK): Domain Knowledge 1 (DK1), which is the google dictionary definition [7] of each object class; Domain Knowledge 2 (DK2), which is the merged description of each object class collected from three different annotators (crowdsourced approach). By this way, we make sure that the domain knowledge is not keyword dependent and re-ordering happens at semantic level rather than at any specific keyword matching level.

Evaluation Metrics: In order to illustrate the performance of the proposed approach, evaluation parameters such as Top-1 accuracy, Top-5 accuracy, and improvements in Top-1 accuracy are used. The Top-1 accuracy is computed as the proportion of images such that the ground-truth label is the Top-1 predicted label. Similarly, the Top-5 accuracy is computed as the proportion of images such that the ground truth label is one of the Top-5 predicted labels.

4.1 Experimental Results

The cumulative accuracy distribution of Baseline Model 1 (BM1) and Baseline Model 2 (BM2) on test dataset I and II is shown in Figure. 3. Figures 3(a), 3(b) shows the improvement in the Top-1 accuracy after re-ordering on dataset I which has one object in an image frame. As shown in Figure. 3, for BM1, without re-ordering only 35% of object classes have Top-1 accuracy more than 90%, whereas with re-ordering using DK1 or DK2 around 55% of classes



(a) (b)

(c) (d)

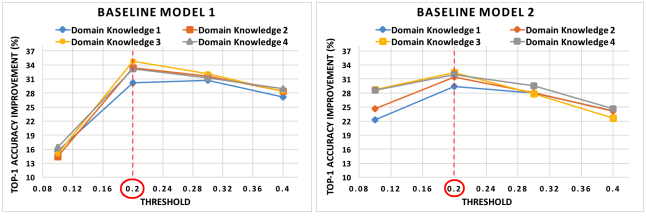
Figure 3: Cumulative accuracy distribution of Baseline Model 1 & Baseline Model 2 on the data set. I(a-b), II(c-d)

.

have more than 90% Top-1 accuracy. Similarly, for BM2 our approach shows 20% improvement in number of classes for 90% or above Top-1 accuracy on dataset I as shown in Figure 3(b).
When a child takes an image, it is common that multiple objects get captured in that image. If more than one object is present in an image, then the confusion of the classifier highly increases which leads to low Top-1 accuracy. Figure. 3(c), 3(d) show the improvement in Top-1 accuracy on data set II, where two objects (from training set) are present in an image frame. As shown in Figure. 3(c), for BM1, without re-ordering only 7% of object classes have Top-1 accuracy more than 90% whereas with re-ordering using DK1 or DK2 around 40% of classes have more than 90% Top-1 accuracy. Similarly, for BM2, our approach shows improvement of 45% in number of classes for 90% or more Top-1 accuracy on dataset II as shown in Figure. 3(d).

4.2 Empirical analysis to determine threshold for strong vouching of semantic models

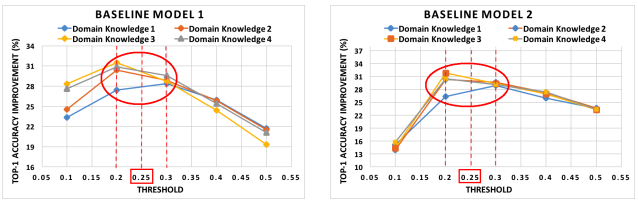
In this section, we explain the empirical analysis which determines the threshold value required by semantic models for strong vouching as discussed in Section. 3.2. In comparing two elements with respect to their semantic correlation with domain knowledge (i.e. DK-CS), the threshold stands for the minimum value by which DK-CS of one element should be higher than the other to confidently say that the element is semantically more correlated with the domain knowledge as compared to the other element. Choice of correct threshold value is very cruicial for the proposed approach. *The threshold value should be as high so as to avoid wrong swapping of labels, and as low to allow correct swapping of labels for better Top-1 accuracy improvements.*
For the empirical analysis of threshold value, we conducted experiments on dataset II with the following combinations (i) four different domain knowledges collected through crowd sourcing, (ii) four different threshold values, and (iii) for both baseline models (BM1&BM2) to make it independent of any local data-behavior. The results are shown in Figure. 4. From the results, we noticed that the correct threshold value is 0.2 for reordering Top-5 predicted labels. As observed in Figure. 4, Top-1 accuracy reaches the peak value when the threshold value is 0.2. We now discuss the reason behind this magical number.



(a) (b)

Figure 4: Improvement in Top-1 accuracy while reordering predicted Top-5 labels for different domain knowledge, threshold values and baseline models

In our approach, we use normalized DK-CS, which means if we consider equal distribution of labels while reordering Top-5 predicted labels, then the DK-CS for each label is 0.2 (i.e. 1/5). We propose that, if DK-CS of one label overrides the semantic score of another label by a value near or equal to the $1/k$ (i.e. individual DK-CS of the labels considering equal distribution of each label), then it is considered as **strong vouching** by semantic model for the former label.
In order to confirm the above proposed claim, we performed ex-



(a) (b)

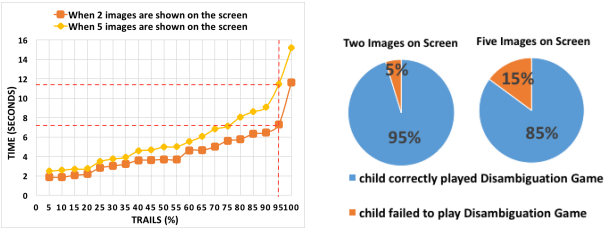
Figure 5: Improvement in Top-1 accuracy while reordering predicted Top-4 labels for different Domain Knowledge, Threshold Values and Baseline Models

periments to reorder Top-4 predicted labels (results are shown in Figure. 5). From the results, we can see that the performance is at peak for threshold between values 0.2 and 0.3, which is near to 0.25 (1/k where k is number 4). There is very noticeable degradation in performance when threshold is below 0.2 or above 0.3. Similar trends were also observed when experimenting with Top-3 re-ordering.
Therefore, the correct choice of threshold while re-ordering Top-k predicted labels is $1/k$. When system is tuned to vouch strongly using this threshold value, we observe high improvements in Top-1 accuracy.

4.3 Usability Study

The main purpose of this usability study is to observe the following key points in children of ages between 3-5 years: (i) whether they can take images using the camera of a phone or tablet, (ii) whether they can perform visual comparison between the physical object for which picture was taken, and its reference image provided by the classifier in the disambiguation game, (iii) comparison of cognitive load on children when they see less vs. more number of images on a device screen during the game. To conduct this study, we asked the child to play with our app installed on iPads, which logged the complete click stream data of the app for tracking various quantitative parameters. We also noted down the feedback from parents/observer during the activity play.

We conducted this usability study on 12 children with a total of 29 trials. In each trail, a child was allowed to play with the app as long



(a) (b)

Figure 6: (a) Cumulative distribution of time taken by children to play disambiguation game when two and five objects are shown on the device screen (b) % of times children correctly played disambiguation game when two and five objects are shown on the device screen

s(he) wanted. We observed that some children played only one time in a trial and some played upto 8 times in a trail. There were no limitations on the number of trails per child. We did not observe even a single instance where a child was asked to capture an image of a relevant object using the camera and s(he) failed to do it. This shows that children of that age group can easily take pictures. The average time taken by a child to search for an object in the environment and take picture was 20 seconds. From the collected data, we observed that around 90% of times, children are able play the disambiguation game correctly. The common feedback which we got from parents/observers is that children liked this app and wanted to play it again and again.

The comparison of cognitive load on a child when (s)he got 2 object images (from Top-2) vs. 5 object images (from Top-5) on a screen during disambiguation game is shown in Figure.6 (a). Around 95% of children took upto 7 and 11 seconds for disambiguation game when they got 2 and 5 options on a screen, respectively (as shown in figure 6a). Similarly, Figure. 6 (b) shows that on an average a child failed to make a visual comparison only 5% of time (when there were 2 images on a screen) and 15% of time (when there were 5 images on a screen). These results indicate that, child is able to easily play the disambiguation game but the cognitive load reduces when less number of images were rendered in each turn of the game. Since the proposed re-ranking algorithm increases the Top-1 accuracy of the system, the child could reach the right object in initial rounds of the disambiguation game with high chance, thereby providing a good user experience.

5. CONCLUSION

We present a holistic visual recognition system for Early Childhood Learning through physical-digital interaction that improves recognition accuracy levels using (a) domain restriction and custom training of vision classifiers, (b) a novel re-ranking algorithm for multi-modal fusion of contextual information, and (c) semi-automated feedback with different gaming scenarios for right object-tag identification & classifier re-training. Through a usability study with 12 children, we make the case that engaging user experiences can indeed be developed to bridge the gap between automatic visual recognition accuracies and the requirement of high accuracy for meaningful learning activities.

Extensive evaluations on large datasets brought forth the deficiency of existing multimodal fusion techniques in combining the domain knowledge context with the vision classification results. Using a data driven approach we show the efficacy of our proposed re-ranking algorithm based on strong vouching, and also show that the swapping threshold (derived from data) is also anchored in a physical meaning.

For future work we would like to conduct extensive pilot study with children to demonstrate evidence-of-learning for vocabulary acquisition using physical-digital interaction. We would also like to use other implicit contexts such as location, speech cues, wearable sensors etc. to derive domain knowledge for better multimodal disambiguation.

6. REFERENCES

[1] Usability guidelines for kids. *http://rosenfeldmedia.com/wp-content/uploads/2014/11/DesignforKids-excerpt.pdf*.

[2] Alien assignment app. *http://my.kindertown.com/apps/alien-assignment*, 2017.

[3] Aurasma. *https://www.aurasma.com/*, 2017.

[4] Blippar. *https://blippar.com/en/*, 2017.

[5] Conceptnet. *https://github.com/commonsense/conceptnet5/wiki*, 2017.

[6] Glove. *http://nlp.stanford.edu/projects/glove/*, 2017.

[7] Google dictionary. *http://www.dictionary.com/browse/google*, 2017.

[8] Imagenet. *http://www.image-net.org/*, 2017.

[9] Layar - augmented reality. *http://appcrawlr.com/ios/layar-reality-browser-augmented*, 2017.

[10] Usability guidelines for kids. *http://hci.usask.ca/publications/2005/HCI_TR_2005_02_-Design.pdf*, 2017.

[11] Word2vec. *https://github.com/dav/word2vec*, 2017.

[12] A. N. Bhute and B. Meshram. Text based approach for indexing and retrieval of image and video: A review. *arXiv preprint arXiv:1404.1514*, 2014.

[13] B. Hart et al. The early catastrophe: The 30 million word gap by age 3. *American educator*, 27(1):4–9, 2003.

[14] Y. Lin et al. Text-aided image classification: Using labeled text from web to help image classification. In *Web Conference (APWEB), 2010 12th International Asia-Pacific*, pages 267–273. IEEE, 2010.

[15] H. Ma et al. Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia*, 12(5):462–473, 2010.

[16] T. Mikolov et al. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[17] B. C. Roy et al. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668, 2015.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] C. Szegedy et al. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[20] L. Tian et al. Image classification based on the combination of text features and visual features. *International Journal of Intelligent Systems*, 28(3):242–256, 2013.

[21] G. Wang et al. Building text features for object image classification. In *Computer Vision and Pattern Recognition*, pages 1367–1374. IEEE, 2009.

[22] C. Xu et al. Fusion of text and image features: A new approach to image spam filtering. In *Practical Applications of Intelligent Systems*, pages 129–140. Springer, 2011.

[23] Y. Zhang et al. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.