Abstract: This work examines the differences between a human and a machine in object recognition tasks. The machine is useful as much as the output classification labels are correct and match the dataset-provided labels. However, very often a discrepancy occurs because the dataset label is different than the one expected by a human. To correct this, the concept of the target user population is introduced. The paper presents a complete methodology for either adapting the output of a pre-trained, state-of-the-art object classification algorithm to the target population or inferring a proper, user-friendly categorization from the target population. The process is called ‘user population re-targeting’. The methodology includes a set of specially designed population tests, which provide crucial data about the categorization that the target population prefers. The transformation between the dataset-bound categorization and the new, population-specific categorization is called the ‘Cognitive Relevance Transform’. The results of the experiments on the well-known datasets have shown that the target population preferred such a transformed categorization by a large margin, that the performance of human observers is probably better than previously thought, and that the outcome of re-targeting may be difficult to predict without actual tests on the target population.

Keywords: cognitive relevance; deep learning; crowd-sourcing; target user population; categorization; classification

1. Introduction

Humans have a different perception of categories than Convolutional Neural Networks (CNNs). For one, CNNs use exclusively visual features to perform the classification. Aside from the low-level visual features, CNNs also take advantage of the spatial structure of these features, i.e., how these features are located [1]. However, humans think differently and explore different, high-level features on images to perform a classification task [2–8]. Thus, for a human, the images of a bird <dunlin> and a reptile <alligator lizard> (Figure 1) fall into different categories, as humans take into the account the high-level concept of affordances—perceivable action possibilities (i.e., only actions that depend on users’ physical capabilities, their goals and past experiences) [9]. The algorithm, however, focuses only on visual information (color, shape, and texture) [10]. Because both types of animals are similar in color and have a similar texture (see Figure 1), they fall into the same category despite being semantically different. From the human perception point of view, such a categorization is wrong and should be punished more severely (with greater penalty) in the training or evaluation process.
In the context of a CNN as a human assistant, the question is which pre-trained object classification algorithm better imitate target-users? The answer is the one that will separate semantically different images as much as possible, but how can such an algorithm be found? Images from the evaluation set could be labeled by human subjects and then the answers and algorithm results could be compared. However, the new problem immediately arises as people label the same image differently. Following the theory of cognitive relevance [12], the labels are based on the context and processing effort which is a matter of human experience and cultural background. In view of the unchanged context, people will be able to label the image with synonyms (words that denote the same concept and are interchangeable in many contexts), hypernyms (words that denote the more broad meaning of a concept), and hyponyms (words that denote a more specific concept). Considering the hierarchical structure of the lexical database, different human labels are positioned vertically. Following the concept of cognitive relevance [12], it may happen that people change the context and label the image completely differently. In this case, they move horizontally across the hierarchical structure.

A large population can be used to solve the problem of image multi-labeling. Horizontal movements in the lexical database can be considered outliers and are removed with the appropriate analysis. After that, the vertical point in the lexical hierarchy can be defined, which is consistent with the principle of maximizing relevance [12]. The located point in the lexical hierarchy is a lexical representation of a concept to which the target population gets the association when it sees a specific image. However, a lexical representation of the image in a categorization dataset, used for the evaluation of deep learning algorithms, can be different. For example, approximately half of 1000 ILSVRC2012 (ImageNet Large Scale Visual Recognition Challenge 2012) object detection dataset categories are so-called ‘leaf’ categories, i.e., they represent very fine narrow concepts (e.g., breeds of dogs, types of mushrooms, exotic species of birds). A general population that would naturally prefer such categorization is probably very difficult to find. Using such a dataset to evaluate human classification performance is not appropriate.

CNNs are often tested on problems that contain hundreds of categories. ILSVRC2012’s 1000 categories are not, by definition, an open-world problem (all possible answers are specified). In practice, the 1000 category dataset-provided labels that cannot be effectively parsed by humans. For humans, a 1000 class classification is effectively an open-world problem. Therefore, people may provide a good answer which is not correct. It does not correspond to the label, as defined by the dataset creators.
However, an evaluation makes sense only if the categories of the dataset are transformed into the human perception of the target population (user population re-targeting).

Essentially, by adapting the dataset for the target user population the bias is introduced into the dataset. Contrary to the biases already contained in the categorization datasets [13–17] that negatively affect the evaluation of algorithms, this kind of bias helps in comparing humans to machines and finding the algorithm that resembles human cognitive processes as closely as possible. To compare humans to machines and find out which images cannot be distinguished by the algorithms, it makes sense to also adapt the algorithms’ responses vertically in the lexical hierarchy.

**Scope of the Research**

There are differing views on categorization in cognitive science [18]. The view, related to the concept of affordances and described above, is considered to be the modern view. In contrast, the classical view of the category relies on the concept of shared properties instead of affordances. For example, both sparrows and penguins have wings and thus they are birds in a biological sense, even though penguins cannot fly.

There is a good reason why a modern view on categorization (based on prototype theory and related to affordances) has gained prominence—it has been found that categorization in people is complex beyond what could be explained by shared properties [18]. This does not mean that the classical view itself is wrong, but its applicability is limited. This research focuses on substituting a human helper with an AI-powered machine and therefore, a modern view on categorization has to be taken into account.

There are, of course, many use cases for AI-powered classification, where the classical view of the categorization (also called objectivist view [18]) is more appropriate, and the categories have to be rigidly defined by an expert. However, the scope of this research is limited to the instances where human reasoning regarding the categorization is imitated, which brings unique challenges, addressed in this work.

In this work, a new user study-based approach to transform existing datasets into the datasets that are tailored to the specific user population is introduced. This approach has two benefits: it can reevaluate both algorithm and human performance, and increase user satisfaction with the output of the classification algorithm. As shown in Figure 2, images from the selected dataset are shown to the target user population. Members of a population perform grouping, category naming, and recognition tasks on a subset of images. The results are used to find a Cognitive Relevance Transform (CRT) that modifies the number, grouping, and naming of categories. By user population re-targeting with the CRT dataset-provided labels, CNNs’ outputs, and human-provided labels are then modified to be able to compare humans to machines.

The developed methodology is generic to the point that can be used for existing datasets and algorithms to properly compare humans to algorithms. The CRT can be employed by anyone seeking to develop a user-friendly AI algorithm or to find one among existing ones. It can also be used for future datasets to determine more relevant categories. Note that the proposed methodology specifically addresses the use of classification methods in situations where the end user of a classification system is a human. In many applications this is not the case (e.g., autonomous driving or industrial applications), but with AI-powered appliances having ever greater roles in human lives and decisions, this aspect of classification will only increase in importance.


The paper is organized as follows: After the related work (Section 2), a **Cognitive Relevance Transform** that transforms the structure of the dataset is defined in Section 3. How to utilize user population re-targeting is demonstrated in Section 4. In the experiments (Section 5), the materials and methodology for conducting similar experiments on ILSVRC2012 and VireoFood-172 datasets are described. The results and discussion for the ILSVRC2012 dataset are presented in Section 6. Section 7 presents the results and discussion for the VireoFood-172 dataset. This paper concludes with Section 8.

### 2. Related Work

After CNNs outperformed other classifiers on the ILSVRC2010 challenge [19], deep learning has become a de facto methodology [20]. Still, CNNs are not perfect and many challenges still need to be solved [21]. One of the challenges is human’s trust in their answers. As researchers demonstrated, there exist colored patterns that are unrecognizable to humans, but CNNs recognize familiar objects with \( \geq 99.6\% \) certainty [22]. Others have shown that the so-called adversarial examples—images with perturbations, not visible to human eye—can also fool the CNNs [23]. Previous studies [22,24,25] pointed out that the problem arises due to adversarial examples occupying a much larger area in input space than training examples.

The literature did not provide any research that considers **user population re-targeting** as an approach to get closer to human classification. Most of the work only considers the comparison of humans to a machine classification performance. The majority is focused on low-level recognition performance studies, based on visual distortions of images or the viewpoint variations of objects. Dodge and Karam [26,27] studied the performance with Gaussian blur and additive Gaussian noise. Both works neglected the fact that the high distortions could be irrelevant to humans with higher cognitive recognition capabilities, similarly to Wichmann et al. [28] (contrast reduction). The results have indicated that human observers and CNNs exhibited a similar performance, but the humans were more robust to contrast changes. The viewpoint variation comparisons between humans and CNNs were done by Kheradpisheh et al. [29,30]. The humans and the CNNs were correlated in the viewpoint variation comparisons, but the error distributions of computational models were different from the humans.
In contrast, Stabinger et al. [31] studied higher-level abstraction capabilities of CNNs and human subjects. The CNNs were trained and tested on the SVRT framework presented by Fleuret et al. [2]. The framework consisted of a series of 23 classification problems with randomly generated shapes. The shapes were unknown to humans, and so the prior knowledge of humans to solving problems was minimized. The classification problems were relationships between the shapes on images. Stabinger et al. [31] reported that CNNs were generally not capable of solving problems containing shape comparison.

Abstraction experiments could be used for higher cognitive capabilities, but researchers Linsley et al. [8], Pramod and Arun [4] showed that humans and algorithms had distinct strategies to solve the problems. Linsley et al. [8] identified visual features used by humans and CNNs during object recognition. The visual features were represented as importance maps for individual images. The authors observed that the maps were strongly stereotyped across the subjects and the CNNs favored different visual features. Pramod and Arun [4] used a different approach with perceived dissimilarity measurements. They collected a large dataset of perceived dissimilarity measurements and used it to train and test computational models. They perceived that the dissimilarity estimation for human subjects was defined as the reciprocal of the average search time of the target among multiple identical distracting objects. The authors reported the existence of systematic differences between the object representations in humans and machines.

The discovery of systematic differences between humans and machines contributed to the development of bio-inspired computational models that would overcome the performance gap. It is believed that a human IT cortex is mainly responsible for the higher cognitive performance in recognition tasks. By Yamins et al. [32] it would thus be necessary for computational models to strongly correlate to the IT. Yamins et al. [32] showed that hierarchical neural networks (HMO models) were highly predictive of the IT cortex. By testing the models on recognizing photorealistic 3D models with different position, scale, and pose, the models matched the human performance. Similarly, Cadieu et al. [33] compared the IT cortex responses to CNNs on recognizing objects with viewpoint variations. They also found that CNNs achieved equal performance to the IT cortex and that they were even better than the bio-inspired HMO models. Still, both works tested only low-level cognitive capabilities. A more high-level study was done by Rajalingham et al. [34]. They observed the object recognition behavior of humans, monkeys, and CNNs on the object-level and image-level discriminability. Rajalingham et al. [34] showed that the CNNs accurately predicted primate patterns of how often the object was incorrectly categorized, but the discriminability of each image from all other objects was significantly different.

All related research shows that regardless of the type of an experiment, there exists a gap between computational models and humans. However, there is no simple approach or metric to evaluate the models, according to their human-like capabilities. Therefore, in this work, it is demonstrated how to use a simple tool (Cognitive Relevance Transform) to truly compare models to humans to get closer to human classification. This tool can be used to normalize already established metrics in object classification to human cognitive relevance capabilities. Transformed metrics are comparable to humans and can clearly show which model has more human-like properties.

3. The Cognitive Relevance Transform

The Cognitive Relevance Transform (CRT) is defined as a sequence of operations on image categories, with each operation belonging to one of the three operation classes: (1) Merging of the multiple dataset categories into a new category, denoted by $E$; (2) Separating the dataset category into multiple categories, denoted by $S$; and (3) Renaming the category, denoted by $R$. 
The CRT is obtained in three steps (see Algorithm 1). Intention is to assess the cognitive relevance of the whole dataset $D \subset \mathcal{H} \times \mathcal{C}$ where $\mathcal{H}$ is a set of images and $\mathcal{C}$ is a set of categories. A modern object detection/recognition dataset contains a huge number of both categories $\mathcal{C}$ and images $\mathcal{H}$. The cost of testing humans on such a dataset would be prohibitive. So, a representative subset $D' \subset \mathcal{H}' \times \mathcal{C}'$, using fewer categories and images per category, needs to be extracted. In the second step, a battery of tests (experiment tasks) on the human population $P$ using subset $D'$ is performed to get image groups as feature vectors $X$ and labels $A$. This work presents the methodology on how to do this properly, to reduce the human workload, but still get useful results. In the third step, the CRT is derived from analyzing the grouping and labeling tasks in user studies.

**Algorithm 1** The Cognitive Relevance Transform.

The Cognitive Relevance Transform (CRT) is a sequence of operations $(E, S, R)$ in image categories. The CRT is obtained in three steps. First (operation 1), the dataset is subsampled by Algorithm 2. Second (operation 2), user studies are performed (see Section 3.2 and Algorithm 3). Third (operation 3), merging operations $E$ and separation operations $S$ are determined by Algorithm 4. Renaming operations $R$ are determined by Algorithm 5. $=$ denotes definition and $\leftarrow$ denotes calculation.

**Require**: $D \subset \mathcal{H} \times \mathcal{C}, P$

**Ensure**: CRT

1: $D' \leftarrow$ DATASET SUBSAMPLING($D$)
2: $(X, A) \leftarrow$ USER STUDIES($P, D'$)
3: CRT = $(E, S, R) \leftarrow$ GET CRT($X, A, D'$)

4: function GET CRT($X, A, D'$)
5: $(E, S) \leftarrow$ DETERMINING MERGING AND SEPARATION OPERATIONS($X, D'$)
6: $R \leftarrow$ DETERMINING RENAMING OPERATIONS($A$)
7: return $(E, S, R)$
8: end function

**3.1. Reduction of the Dataset Size**

Reduction of the dataset size can be done by Algorithm 2, which requires the total number of categories $|\mathcal{C}|$, a maximal number of categories per task $\nu_c$, a number of images per category $\nu_i$, categories’ effect size $h_c$, images’ effect size $h_i$, significance level $\alpha$, and statistical power $1 - \beta$. The algorithm ensures the minimal number of categories $|\mathcal{C}'|$, the minimal number of images per category $|\mathcal{H}'|$, and the number of batches $|B|$ that are needed to evaluate all images from a subset $D'$. Batch $B$ is a set of categories and images $B \subset D'$ that can fit into one task.
Algorithm 2 Dataset Subsampling.
To reduce the dataset size, the minimum number of categories $n_c$ is calculated by the statistical power of two-sample test for proportions with unequal sample sizes. For details, see [35]. The minimum number of images per category $n_i$ is calculated by the statistical power of $\chi^2$ statistical test of independence [35]. Both, $n_c$ and $n_i$ are then corrected to a discrete number by operation 6. Finally, the number of batches $|B|$ is calculated. $\leftarrow$ denotes calculation.

Require: $|C|, \nu_c, \nu_i, h_c, h_i, \alpha, 1 - \beta$
Ensure: $|C'|, |H'|, |B|

1: $n_c \leftarrow \text{PWR.2P2N.TEST}(|C|, h_c, \alpha, 1 - \beta)$
2: $n_i \leftarrow \text{PWR.CHISQ.TEST}((n_c - 1)^2, h_i, \alpha, 1 - \beta)/n_c$
3: $|C'| \leftarrow \text{TO INTEGER}(n_c, \nu_c)$
4: $|H'| \leftarrow \text{TO INTEGER}(n_i, \nu_i)$
5: $|B| \leftarrow (|C'||H'|)/(|C|\nu_c)$
6: function TO INTEGER($n, \nu$)
7: return $\lceil n \nu \rceil / \nu$
8: end function

The minimum number of categories $n_c$ is determined by the statistical power of two-sample test for proportions with unequal sample sizes (operation 1 in Algorithm 2). The null hypothesis for the operation is no difference in the size proportion of the original dataset and its subset. $n_c$ is then corrected to final count $|C'|$ by function TO INTEGER in Algorithm 2 operation 6. With this operation, $|C'|$ is established as: (1) a discrete number; and (2) divisible by $\nu_c$. The first condition ensures whole categories. The second condition ensures the categories can be equally divided among tasks in user studies.

The minimum number of images $n_i$ is determined by the statistical power of $\chi^2$ statistical test of independence (operation 2 in Algorithm 2), where the null hypothesis is two random variables are independent. The reason behind using such a test is that one of the results after the human experiment will be a human confusion matrix for the selected categories. The confusion matrix consists of actual and predicted class variables. The basic idea is that a predicted class would mainly indicate the same actual class (they are dependent). Then, $n_i$ is corrected to final count $|H'|$ by function 6 in Algorithm 2 to ensure the same conditions described for $|C'|$.

The number of batches $|B|$ is calculated by operation 5 in Algorithm 2. Because the categories and images can be equally divided among tasks (ensured by function 6 in Algorithm 2), all batches have equal size.

3.2. User Studies

To get the most out of user studies, the proper size of human population $P$ must be first determined by Algorithm 3. The algorithm requires batch size $|B|$, minimum probability of successful classification by humans $p_m$, the estimated probability of successful classification by humans $p_e$, and statistical power $1 - \beta$. It ensures the total number of human subjects $|P|$ and an estimation of a number of observations per image $n_o$. 
Algorithm 3 Population size.

Effect size $h_o$ for the number of observations $n_o$ is calculated by operation 1 (see [35]). Next, $n_o$ is determined by the statistical power of the binomial test [35]. Finally, the number of batches $|B|$ and the number of human subjects $|P|$ are calculated. $\leftarrow$ denotes calculation.

Require: $|B|, p_m, p_e, 1 - \beta$

Ensure: $|P|, n_o$

1: $h_o \leftarrow 2 \arcsin(\sqrt{p_m}) - 2 \arcsin(\sqrt{p_e})$
2: $n_o \leftarrow \text{PWR.P.TEST}(h_o, 1 - \beta)$
3: $|P| \leftarrow |B| n_o$

Estimation of a number of observations per image $n_o$ is done by the statistical power of the binomial test (operation 2 in Algorithm 3) with the effect size $h_o$. $n_o$ significantly determines a change in the image’s category as each image is checked if there exists a difference between a dataset-provided and a human-labeled category.

The total number of human subjects $|P|$ for the whole experiment is determined by operation 3 in Algorithm 3, as each batch $B$ can fit in one task and all images in $B$ must be evaluated $n_o$ times.

The experiment on the human population $P$ is divided into two tasks, with clearly defined goals. A sequence of tasks guarantees that the information is revealed to the participants incrementally, so no cross-contamination occurs between the tasks. In the first task, the participants are asked to group similar images as they see fit (Figure 3). The goal is to evaluate which categories from the dataset $D'$ are perceived as one category and which are separate categories by the human population $P$. The number of categories $\nu_c$ and the number of images in the category $\nu_i$ are unknown to the participants.

**Figure 3.** The participants are asked to group similar images as they see fit in Task 1. <jacamar> photos by Bernard Dupont licensed under CC BY-SA 2.0 and by David Surtees licensed under CC BY 2.0. <siamang> photos by cuatrok77 licensed under CC BY-SA 2.0 and by Abi Skipp licensed under CC BY 2.0. <gyromitra> photos by Tatiana Bulyonkova licensed under CC BY-SA 2.0 and by Andrey Gaverdovsky licensed under CC BY 2.0.

In the second task, the participants are shown the next sequence of images. They are asked to label the images as they see fit (Figure 4). The goal of this task is to determine the renaming operation.
3.3. Deriving CRT operations

The grouping task gives a set of feature vectors $X = \{x_j\}_{j \leq |D'|}$ where every feature vector $x_j = \{v_h\}_{h \leq |P|}$ represents an image from a subset $D'$ and its attributes $v_h$ represent group identification number (ID) from each human grouping. Group IDs are unique in such a way that two groups created by different human subjects have different IDs despite being created from the same batch of images $B$. Set $X$ is then used in Algorithm 4.

Algorithm 4 Merging and Separation operations.

To determine merging operations $E$ and separation operations $S$, a set of feature vectors $X$ is clustered into a set of $C$ clusters. Then, for each cluster $C_i$ and category $C'_j$ get $n'_j$ as a relative frequency of $C'_j$ clustered into $C_i$. Use $n'_j$ to separate the category $C'_j$. In context of a cluster $C_i$, separation of $C'_j$ is denoted by $s'_j$. A set of separation operations for all categories $C'$ in the context of a cluster $C_i$ is denoted by $S_i$. Merging operation $E_i$ in the context of cluster $C_i$ is calculated by merging results from separation operations $S_i$. = denotes definition and $\gets$ denotes calculation.

**Require:** $X = \{x_j\}_{j \leq |D'|}, x_j = \{v_h\}_{h \leq |P|}, D' \subset C' \times H'$

**Ensure:** $E, S$

1. $C \gets \text{CLUSTERING}(X)$
2. $(E, S) \gets \text{GET CRT OPERATIONS}(C, D')$

3. **function** GET CRT OPERATIONS$(C, D')$
   4. for $i \leq |C|$
      5. for $j \leq |C'|$
         6. $n'_j \gets \text{FREQUENCY}(C'_j, C_i) / |H'|$
         7. $s'_j \leftarrow \text{SEPARATE}(C'_j, n'_j)$
      8. end for
   9. $S_i = \{s'_j\}_{j \in |C'|}$
   10. $E_i = \text{MERGE}(S_i)$
   11. end for
12. return $(E = \{E_i\}_{i \leq |C|}, S = \{S_i\}_{i \leq |C|})$
13. **end function**
The clustering operation 1 in Algorithm 4 returns $C = \{C_i\}_{i \leq k}$ clusters where each cluster $C_i$ contains images that the majority of human subjects agree to belong together. Clusters $C$ are then used to determine the operations to transform the original categories to the newly created clusters by operation 3 in Algorithm 4.

For each cluster $C_i$, a set of separation operations $S_i = \{s_{ij}\}_{j \leq |C_i'|}$ and a merging operation $E_i$ are proposed. Each separation operation $s_{ij}$ defines a separation of category $C_i'$ in the context of a cluster $C_i$. Separation operation $s_{ij}$ is calculated using the relative frequency $n_{ij}$ (see operation 6 in Algorithm 4), which is the percentage of images from category $C_i'$ that can be assigned to cluster $C_i$. Merging operation $E_i$ in the context of cluster $C_i$ is calculated by merging results from separation operations $S_i$. In other words, merging operation $E_i$ combines images that can be assigned to cluster $C_i$ regardless to which category they originally belong.

Clusters $C$ can be described by the most frequent label derived from Algorithm 5. A set of labels $A$ from the second task is cleaned (denoted by $A^*$) and structured into a set of feature vectors $Y$, where every feature vector $y_j$ represents an image from a subset $D'$ and its attributes $a_k^*$ represent a cleaned label from each human labeling. $Y$ is then clustered in the same way as $X$ because they are dependent. Each cluster $C_i$ is then assigned the most frequent label from the corresponding label cluster $Z_i$.

**Algorithm 5** Determining Renaming operations.

A set of labels $A$ is cleaned by the number of text operations. Cleaned labels $A^*$ are structured into a set of feature vectors $Y$ and clustered into a set of $Z$ clusters. Renaming operation $R_i$ for cluster $Z_i$ is determined by the most frequent label.

Require: $A = \{a_{jh}\}_{j \leq |D'|, h \leq |P|}$

Ensure: $R$

1. $A^* \leftarrow$ CLEAN($A$)
2. $Y = \{y_j\}_{j \leq |D'|}, y_j = \{a_{kh}\}_{h \leq |P|}$
3. $Z \leftarrow$ CLUSTERING($Y$)
4. $R = \{R_i\}_{i \leq |Z|} \leftarrow$ MOST FREQUENT LABEL($Z$)

5: function CLEAN($A$)
6: $A^* \leftarrow$ TO LOWERCASE($A$)
7: $A^* \leftarrow$ DECODE TO CLOSEST ASCII CODE($A^*$)
8: $A^* \leftarrow$ REMOVE NON-ALPHABETIC CHARACTERS($A^*$)
9: $A^* \leftarrow$ ABBREVIATIONS TO WHOLE WORDS($A^*$)
10: $A^* \leftarrow$ STRIP WHITE SPACE($A^*$)
11: $A^* \leftarrow$ WORD SEGMENTATION($A^*$)
12: $A^* \leftarrow$ TOKENIZE WORDS($A^*$)
13: $A^* \leftarrow$ KEEP NOUNS, ADJECTIVES AND PREPOSITIONS($A^*$)
14: $A^* \leftarrow$ SPELL CHECK($A^*$)
15: $A^* \leftarrow$ LEMMATIZE($A^*$)
16: return $A^*$
17: end function

4. User Population Re-Targeting

The task is to compare and adapt machine algorithms to the human population $P$. First, machine classification is done on a subset $D'$. For images in $H'$, the algorithm normally outputs
a score or a confidence value for categories in $C'$. The algorithm’s top guess is then considered as the correct answer (Top-1 classification). Second, $n_h$ unique human subjects categorize each image in $H'$. They can only choose one category from $C'$ for each image.

Given both the output of the human and machine classification on a subset $D'$, classifications can be evaluated by comparing confusion matrices as they are the most information-rich representations of classifier performance. However, there exist fundamental differences: the machine algorithm provides one result (albeit it can, in some cases, provide the confidence scores for multiple hypotheses), but the human population will provide one answer per human subject. The comparison between a population and a machine is therefore not entirely straightforward, as subjects differ in their knowledge (and possibly attention).

To address this problem, the individual answers by human subjects are essentially transformed into the population answers by Algorithm 6. The initial assumption is that original labels provided by the dataset creators $L'$ are correct (but of course, not optimal), as the dataset was built using human annotators. Answers by human subjects are considered a two-class problem. Each image can be classified as a positive class (most frequent label) or a negative class (all the other labels). The minimum probability $p_e$ is then required for human subject consensus to use the positive class label $\hat{l}_j$ as the consensus label $l^*_j$ instead of the original label $l'_j$. As shown in experiments, this results in very small changes to the labeling of the subset, which confirms that initial labeling did not have many gross errors.

**Algorithm 6 Human subject consensus.**

Individual answers by human subjects $L$ are transformed to the population answers $L^*$. First, calculate number of successes and trials. Success is a human label $l_j^h$ that is equivalent to original label $l'_j$. A binomial test (for details see [35]) is used to determine consensus label $l^*_j$ as a population answer. If more than $p_e$ human subject consensus exists, consensus label $l^*_j$ becomes the most frequent label $\hat{l}_j$, otherwise it becomes the original label $l'_j$. = denotes definition, $\equiv$ denotes equivalence, and ← denotes calculation.

**Require:** $L = \{l_j^h\}_{h \leq |P|}$, $L' = \{l'_j\}_{j \leq |D'|}$, $P_e$, $\alpha$

**Ensure:** $L^* = \{l^*_j\}_{j \leq |D'|}$

1: for $j \in D'$ do
2: successes ← $\sum_{h \leq |P|} l_j^h \equiv l_j'$
3: trials ← $|P|
4: $\hat{l}_j$ ← THE MOST FREQUENT LABEL($\{l_j^h\}_{h \leq |P|}$)
5: p-value ← BINOMIAL TEST(successes, trials, $p_e$)
6: if p-value < $\alpha$ then
7: $l^*_j$ ← $\hat{l}_j$
8: else
9: $l^*_j$ ← $l'_j$
10: end if
11: end for

Given the population and machine classification results on a dataset $D$, it can formally be defined how the CRT influences confusion matrices (see diagram in Figure 5). An untrained classification map is denoted by $\delta : H \rightarrow C$ where $H$ is a set of images and $C$ is a set of classification classes. The testing sets are denoted by $D \subseteq H \times C$, and a trained classifier on such testing set by $\delta(D)$. As human classification is not trained from any test set involved in these experiments, a trained human classifier is denoted by $\delta_H(\cdot)$. In particular, the machine classifier is denoted by $\delta_M$, and human classification
is denoted by $\delta_H$. Since a confusion matrix is computed from a trained classifier $\delta(D)$ on a testing subset $D'$, it is denoted by $\text{CFM}(\delta(D), D')$. The comparison of results of mAP (mean average precision) among human and machine classification is denoted by $\ominus$.

$$\text{CFM}(\delta_H(\cdot), D') \xrightarrow{\text{Algorithm 6}} \text{CFM}_P \xrightarrow{\text{CRT}} \text{CFM}_P^* \xrightarrow{\text{mAP}} \text{mAP}_H$$

$$\text{CFM}(\delta_M(D), D') \xrightarrow{\text{CRT}} \text{CFM}_M^* \xrightarrow{\text{mAP}} \text{mAP}_M$$

**Figure 5.** The data processing diagram of user population re-targeting. Dataset is denoted by $D$ its subset is denoted by $D'$, confusion matrix computed from human results is denoted by $\text{CFM}(\delta_H(\cdot), D')$ and confusion matrix computed from machine classification is denoted by $\text{CFM}(\delta_M(D), D')$. $\text{CFM}_P$ represents population confusion matrix, $\text{CFM}_P^*$ denotes population confusion matrix transformed by CRT operations, $\text{CFM}_M^*$ is a transformed machine confusion matrix, mAP denotes mean average precision metric, and the comparison of metrics is denoted by $\ominus$.

5. Experiments

Two similar experiments on different datasets were conducted. First, the ILSVRC2012 dataset (http://image-net.org) was used since it was widely used in the past seven years, and as far as it is evident, never put to scrutiny the way it was done in this work. The population $P$ in this experiment was very diverse, subjects from six English-speaking countries with no additional control. Additionally, this experiment was used to verify whether the obtained CRT is only a statistical fluke, and whether people actually like new categories.

A human studies methodology was also used in the VireoFood-172 dataset (http://vireo.cs.cityu.edu.hk/VireoFood172). The dataset is not so widely used and it is also much smaller in size. Because it contains only Asian food, it is somewhat ideal for experimenting on a more specific population that would result in stronger effects. For this experiment, two different populations ENG and ASIA were used. The ENG population consisted of subjects from two English-speaking countries, and the ASIA population contained subjects from two Asian countries.

5.1. Materials

Reduction in the dataset size (Algorithm 2) and determining population size (Algorithm 3) were performed with RStudio version 1.1.442 [36], R version 3.3.2 [37] and pwr package version 1.2-0 [38]. Additionally, the results were cross-checked with G* Power 3 version 3.1.9.2 [39].

For human studies (see Section 3.2), subjects were recruited using the Clickworker (http://www.clickworker.com) platform. The experimental environment was arranged on a custom server, where each subject received a unique link to a sequence of tasks, which they solved in the web browser. For the experimental environment, Django version 2.1.7 with Python version 3.6.7 was used and it was running on Nginx server version 1.16.0. Data was saved in the Postgresql database version 11.3.

For the clustering method in Algorithm 4, the Python package kmodes version 0.10.1 [40] was used as the implementation of the k-Modes algorithm [41,42] with the improved initialization method [43]. A number of clusters $k$ was defined by the elbow method. The algorithm was run with different numbers of $k$ to get the graph of the cost function $P(W, Q)$ where $W$ is a partition matrix and $Q$ is a vector of modes (for details see [41,42]). The optimal $k$ was selected on the location of an elbow (see Sections 6.1 and 7.3 for details).
When determining renaming operations by Algorithm 5, the SymSpell algorithm [44] was selected for word segmentation. For the implementation of the algorithm, a Python package symspellpy version 6.3.8 [45] was used. Newly created words were then tokenized by NLTK version 3.4.3 [46]. The lemmatization process was also done by NLTK using WordNet [47]. Non-existent lemmas were replaced by ‘unknown’. For other cleaning functions, standard Python version 3.6.7 procedures were used.

MXNet version 1.5.0 and GluonCV framework version 0.5.0 [48,49] were used for deep learning models. Evaluation and comparison of population and machine classification was done using Python packages Pandas version 0.24.2 [50], Scipy version 1.3.1 [51], Scikit-learn version 0.21.2 [52], and Numpy version 1.17.0 [53,54].

5.2. Experiments on ILSVRC2012 Dataset

The ILSVRC2012 object detection dataset [19] was used as the dataset D. It contains |C| = 1000 categories. Each category contains from 732 to 1300 images. Unfortunately, the ground truth for the ILSVRC2012 test images was not available. For all tests, ILSVRC2012 validation images were used. While this did not influence human performance, it is plausible that it overestimated machine performance to a certain extent.

5.2.1. Reduction of Dataset Size

The preliminary testing suggested that on average, the maximum time for a subject to still have enough concentration to end the task is 20 min. Based on the experience with preliminary test subjects (did not participate in the main experiment), it was established that it would be appropriate to have \( \nu_c = 5 \) categories and \( \nu_i = 8 \) images per category. This totals in 40 images per batch \( B \). See Section 3.1 for definitions.

To detect large changes between the smaller experimental dataset \( D' \) and the original dataset \( D \) at least \( |C'| = 10 \) categories, \( |H'| = 16 \) images per category, and \( |B'| = 4 \) batches were needed. The numbers were obtained by Algorithm 2 from Section 3.1 with parameters \( |C| = 1000, \nu_c = 5, \nu_i = 8, h_c = 0.9, h_i = 0.5, \alpha = 0.05, \) and \( 1 - \beta = 0.8 \).

5.2.2. User Studies

For proper user studies, at least \( |P| = 92 \) subjects and \( n_o = 23 \) observations per image were required. The numbers were obtained by Algorithm 3 from Section 3.2 with parameters \( |B'| = 4 \) batches, \( p_m = 0.5, p_c = 0.75 \) and \( 1 - \beta = 0.8 \).

Please note, \( n_o = 23 \) observations per image is a number that is needed for statistical significance. It is much larger than the number of observations usually used when annotating computer vision datasets!

The only constraint placed on the population of subjects \( P \) in user studies was that they live in the countries where the primary language is English (the UK, the USA, Canada, Australia, Ireland, New Zealand), as Task 2 requires text entry of labels, which assumes familiarity with the English language. Each subject was paid 3 EUR for 20 min of work to motivate them to approach the tasks seriously. In practical applications, the human subjects would be sampled according to some predefined criteria (e.g., sampled from a target market or a target demography for an AI-powered machine).

5.2.3. Choice Of Categories

Eight ILSVRC2012 categories that have the lowest frequency (popularity) in English texts were chosen to simulate a challenging task for human observers. As a control, two categories that are most frequent in a written text were added. When choosing eight low-frequency categories, those that have a high-level concept in their names were skipped. It is obvious to humans that <frilled lizard> is a kind of a lizard, and in the absence of multiple lizard species, such results would not reflect the ability of humans to correctly classify categories they never heard of. The categories are shown in Table 1.
Table 1. ILSVRC2012 categories, selected to form smaller dataset $D'$. `<web site>` and `<library>` are assumed to be easy for people to recognize, the rest are expected to be hard. Descriptions provided by WordNet [47].

| Category | WordNet Description |
|----------|---------------------|
| `<site>` | A computer connected to the internet that maintains a series of web pages on the World Wide Web. |
| `<library>` | A building that houses a collection of books and other materials. |
| `<dunlin>` | Small common sandpiper that breeds in northern or Arctic regions and winters in southern United States or Mediterranean regions |
| `<bolete>` | Any fungus of the family Boletaceae |
| `<jacamar>` | Tropical American insectivorous bird having a long sharp bill and iridescent green or bronze plumage |
| `<gyromitra>` | Any fungus of the genus Gyromitra |
| `<dhole>` | Fierce wild dog of the forests of central and southeast Asia that hunts in packs |
| `<kakatoe galerita>` | White cockatoo with a yellow erectile crest |
| `<earthstar>` | Any fungus of the family Geastraceae; in form suggesting a puffball whose outer peridium splits into the shape of a star. |
| `<siamang>` | Large black gibbon of Sumatra having the 2nd and 3rd toes partially united by a web. |

5.2.4. Deep Learning Models

For comparison of humans to the machine, widely known AlexNet [55], VGG19 [56], and ResNet152v2 [57] were used. All of them were pretrained on the ILSVRC2012 dataset. AlexNet is an older architecture and it was used on purpose, to give it a better chance in the improvement of performance between original dataset categories, and the new, CRT-transformed categorization.

5.3. Experiments on VireoFood-172 Dataset

The VireoFood-172 dataset [58] was also used as the dataset $D$. It contains $|C| = 172$ categories of popular Asian dishes. Each category contains from 191 to 1061 images. The categories cover eight groups of foods Bean products, Egg, Fish, Meat, Seafood, Soup, Staple, and Vegetables. The distribution of food categories under eight groups is represented in Figure 6.

![Figure 6. Distribution of food categories under eight groups from VireoFood-172 dataset [58].](image)

5.3.1. Reduction of the Dataset Size

Following the methodology from Section 3.1, at least $|C'| = 16$ categories, $|H'| = 18$ images per category, and $|B| = 6$ batches were needed to detect large changes between the smaller experimental dataset $D'$ and the original dataset $D$. The numbers were obtained by the Algorithm 2 with parameters $|C| = 172$, $\nu_c = 8$, $\nu_i = 6$, $h_c = 0.8$, $h_i = 0.5$, $\alpha = 0.05$, and $1 - \beta = 0.8$. 

5.3.2. User Studies

For user studies (see Section 3.2), \( n_o = 24 \) observations per image and \( |P| = 144 \) subjects were required. The numbers were obtained by Algorithm 3 with parameters \( |B| = 6, p_m = 0.5, p_e = 0.75, \) and \( 1 - \beta = 0.8. \)

In the case of the VireoFood-172 dataset, two experiments with different populations of subjects were performed. The first population of 144 subjects \( P_1 \) lives in the UK or Ireland. With such a constraint, a large pool of culturally similar subjects were obtained.

The second population of 144 subjects \( P_2 \) only live in Malaysia or Singapore. The constraint also brought a large pool of culturally similar subjects. Note that using the Clickworker platform, there were limited options for Asian countries to select. China was not among them.

The age constraint (18–50) was set to both populations. With the age constraint, experiments were focused on the population, which was expected to be more computer literate.

Each subject was initially paid 2.25 EUR for 15 min of work, to motivate them to approach the tasks seriously. To be able to recruit enough subjects from Ireland and Singapore, the payment was gradually increased to 8 EUR.

5.3.3. Choice of Categories

Before choosing the categories, a demography experiment was conducted where the subjects from the chosen population were asked about their eating habits. Both populations were mainly omnivores (they eat everything). By also considering the distribution of category groups (Figure 6), the most representative categories would be the ones from the meat group. From this group, 16 random categories were chosen, which are shown in Table 2.

| No. | Category                                      |
|-----|----------------------------------------------|
| 1.  | <pickles, shredded pork & vermicelli>        |
| 2.  | <fried lamb with cumin>                     |
| 3.  | <four-joy meatballs>                        |
| 4.  | <sauteed bullfrog with pickled peppers>     |
| 5.  | <saute spicy chicken>                       |
| 6.  | <spare ribs with garlic>                    |
| 7.  | <chicken feet with pickled peppers>         |
| 8.  | <sauteed shredded pork with skin of tofu>   |
| 9.  | <braised beef with brown sauce>             |
| 10. | <sauteed shredded pork with garlic sprout>  |
| 11. | <roast chicken wings>                       |
| 12. | <braised intestines in brown sauce>         |
| 13. | <braised pork>                              |
| 14. | <sauteed snails>                            |
| 15. | <beefsteak>                                 |
| 16. | <pork with garlic sauce>                    |

5.3.4. Deep Learning Models

For comparison of humans to the machine, the same pre-trained models from the ILSVRC2012 experiment were used. The models were fine-tuned on the VireoFood-172 dataset. The models were fine-tuned with the parameters momentum 0.9, weight decay 0.0001, base learning rate 0.001 with decay each 10th epoch by the factor 0.75, and 60 training epochs were used.

The data for training was augmented in the following way. First, the original image was cropped with a random size 0.08 to 1.0 and a random aspect ratio 0.75 to 1.33, and then resized to 224. Next, the image was randomly flipped left or right with a probability of 0.5. Image brightness, contrast, saturation, and hue were randomly jittered by a factor 0.4. Then, AlexNet-style PCA-based noise was added to the image with an intensity of 0.1. Finally, the image was normalized with ImageNet mean \((0.485, 0.456, 0.406)\) and standard deviation \((0.299, 0.224, 0.225)\).
6. ILSVRC2012 Results and Discussion

CFM analysis has shown human subjects created many false positive and false negative results for each category (see Figure A1 in the Appendix A for details). Also, many images, regardless of the category, were classified as "<unknown>". This suggests the classification task was hard as subjects did not know which category to choose. "<site>" category had the best result as it was correctly recognized 96% of the time. Also, "<kakatoe galerita>" and "<library>" were easily recognizable. That the "<site>" and "<library>" were one of the most recognizable categories, is consistent with the hypothesis that the popular categories will be easily recognized by humans (see Section 5.2.3). In contrast, the ability to separate other classes, excluding "<kakatoe galerita>", was very low in humans. Metrics Top-1 ACC (top-1 accuracy), Precision, Recall, and F1-Score were calculated 51.4%, 65.2%, 46.8%, and 53.0% respectively. The numbers are very low, but that does not necessarily mean people are bad at classifying. The results rather indicate that given labels cannot be correctly associated with given images by more than 75% of the population.

There is a significant difference between the average human confusion matrix CFM_H (Figure A1a), obtained by averaging categorization results of \( n \) unique human subjects, and the population confusion matrix CFM_P (Figure A1b), which was determined by the methodology from Section 4. CFM_H results show there were many outliers, and with human subject consensus (Algorithm 6) they were successfully eliminated. For example, the algorithm removed the potential answers from humans that were not taking the experiment seriously (randomly selecting the answers) or answers that were selected by mistake. These kinds of errors cannot be successfully supervised with online experiments.

The machine confusion matrices CFM_M were obtained simply by running respective algorithms on the validation set of the selected ILSVRC2012 categories. AlexNet, the oldest CNN architecture had the worst classification results (Figure A2a). The best CNN results were obtained by ResNet architecture (Figure A2b). CNN metric results can be viewed in Section 6.2. Note that the term ‘original’ is strictly used for CFM matrices before applying CRT and ‘changed’ for transformed CFM matrices that have gone through a CRT transformation.

The comparison of CFM_P and CFM_M implies CNNs are not superior in the classification task. Humans outperformed all CNN models as later classified more images into the "<unknown>" category.

It should be noted that human and machine error modes are somewhat different: humans will label the image as "<unknown>" if they do not know the answer, but the machine, pre-trained on ILSVRC2012 will choose the label outside of selected categories from \( D' \). Both of these categories were denoted as "<unknown>" even though they represent conceptually different things. This is necessary to keep the human experiment within the manageable limits.

6.1. Deriving CRT operations

A total of 17 optimal clusters were determined by the elbow method and cost function \( P(W, Q) \) (Figure 7). There was a clear change of slope at 17 clusters and a noticeable drop after three clusters. It was verified that the three clusters did not provide any meaningful clustering, and the value of 17 was used.

The results indicate that even though humans created a larger amount of clusters, the most frequent labels were sometimes the same for multiple clusters. This was an important revelation: people knew that there were two different kinds of birds in images, but they were OK if they got the same label. These labels were merged to get more general categories.

Transformation of the dataset \( D' \) into a new one is represented by Table 3. All bird types were merged into "<bird>" category and all fungi types were merged into "<mushroom>" category. This was consistent with the results from Task 3 where the population CFM was obtained. The matrix indicated bird and mushroom category types as one of the hardest categories to distinguish.
Figure 7. Optimal cluster number analysis for ILSVRC2012 experiment. An analysis was done by the k-Modes algorithm [41–43]. A clear change of slope is visible at 17 clusters (orange markings).

Table 3. Transformation of the dataset $D'$ labels into the cognitive relevant labels.

| No. | Transformed Category       | Original Categories                  |
|-----|----------------------------|-------------------------------------|
| 1.  | <bird>                     | <dunlin>, <jacamar>, <kakatoe galerita> |
| 2.  | <book>                     | <library>                           |
| 3.  | <fox>                      | <dhole>                             |
| 4.  | <library>                  | <library>                           |
| 5.  | <unknown>                  | <library>, <bolete>                 |
| 6.  | <monkey>                   | <siamang>                           |
| 7.  | <mushroom>                 | <bolete>, <gyromitra>, <earthstar>   |
| 8.  | <website>                  | <site>                              |

<siamang>, <dhole>, and <site> were renamed to <monkey>, <fox>, and <website> respectively. The results indicate that the users did not associate the dataset-provided labels with the corresponding images. Thus, more general labels should be considered. Even more, users thought that images of <dhole> category depict a fox. In reality, <dhole> is a wild dog and despite being in the same taxonomic family, Canidae does not fit into the subfamily of a fox. Nevertheless, <dhole> visually looks like a fox, and that is the most obvious explanation for the observed change in labeling.

<library> was split by users into the categories <book> and <library>. Upon examination of the data, it was found that the original category was quite heterogeneous—some images were more easily associated with the <book> rather than the <library>. Labels are conceptually close, as libraries contain books, but there is sufficient ambiguity that it gave rise to the two distinct categories.

Hard to recognize images were put into the new <unknown> category. This category represented the images that were hard to recognize or hard to name.

6.2. Human Population and Machine Classification after Applying the CRT

The Top-1 ACC metric of population results did not change and neither did they for the ResNet model. A significant change was perceived in all other metrics. The precision of human results increased by 6.0 pp (percentage points), but the precision of ResNet dropped by 1.3 pp (Table 4). The increase of human results clearly shows the importance of choosing correct category labels and correct images that are associated with those labels. Similarly, the drop in the precision metric for CNN indicates that their performance can be overestimated if a dataset is not constructed with a human population in mind. Note, transformed CFM matrices $\text{CFM}_P$ and $\text{CFM}_M$ can be viewed in Figure A3a and Figure A3b respectively.

In the recall metric human results increased by 9.4 pp and ResNet’s increased 7.5 pp. Similarly, there was an 8.2 pp increase from original to changed in human’s F1-Score, and 0.7 pp drop in ResNet’s F1-Score.
Similar results of the CRT transformation were obtained for other computational models too. The results can be seen in Table 4 for VGG19 and AlexNet as well. They are consistent with the results on ResNet as Precision dropped, Recall increased, F1-Score dropped as well, and Top-1 ACC did not change.

Table 4. Metrics of human subjects and the selected deep neural networks AlexNet [55], VGG19 [56], and ResNet152v2 [57] on the pre-CRT (original) and post-CRT (changed) ILSVRC2012 subset $D'$.

| Model               | Top-1 ACC (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---------------------|---------------|---------------|------------|--------------|
|                     | Pre | Post | Pre | Post | Pre | Post | Pre | Post |
| Human population    | 99.38 | 99.38 | 90.91 | 96.88 | 90.34 | 99.74 | 90.62 | 98.08 |
| AlexNet             | 88.13 | 88.13 | 90.91 | 88.13 | 80.11 | 84.38 | 84.66 | 79.23 |
| ResNet152v2         | 96.88 | 96.88 | 90.91 | 89.58 | 88.07 | 95.57 | 89.42 | 88.74 |
| VGG19               | 94.38 | 94.38 | 90.91 | 88.75 | 85.80 | 91.92 | 88.08 | 85.26 |

6.3. Qualitative Results

Four examples of images where CNN clearly missed the category are shown in Figure 8. The first image represents <kakatoe galerita> and was correctly recognized by humans. AlexNet recognized it as <ox>. This is not the kind of mistake that humans would make, and consequently, the CNN can be hardly defined as intelligent in this case. Such mistakes could potentially evoke frustration in users of the real appliance, relying on the same CNN method.

6.4. Verification

The first question is, whether the changed metrics are only a statistical fluke. To examine this, 10.00 random CRT transforms were generated, which all contained nine final categories, and observed the range of possible metrics for human and machine. The results, which are shown in Figure 9, were evaluated by the Wilcoxon signed-rank test as distributions. The null hypothesis was defined as Performance of random CRT transforms is equal to measured performance for humans and a machine. The null hypothesis was rejected as $p$-values for human and CNN results were $> 0.001$. The obtained CRT is, thus, very likely not a statistical fluke.

Figure 8. Classification on ILSVRC2012 dataset [19]. When using original categories CNN classifications are absurdly different from true categories. Images are only examples to visually represent a dataset category. They are not part of the dataset. <kakatoe galerita> photo by Lip Kee Yap licensed under CC BY-SA 2.0, <library> photo by Christopher John SSF and <bolete> photo by Jason Hollinger licensed under CC BY 2.0.
Figure 9. Changed human (a,e) and machine (b,d,f) metrics (in blue) on ILSVRC2012 subset are not accidental, since almost all randomly generated CRTs performed worse in the case of human and machine (orange bars represent results of random test in histogram form, heights are bin frequencies). Note that almost all orange bars are to the left of the blue line.

The final question is, whether people actually like new categories. The new human verification test was set up with subjects who did not participate in any of the previous tests but have been drawn from the same population. They were presented with images, and the choice of pre-CRT label and post-CRT label for the image. One hundred people were asked simply “Which label do you like the most?” and all images from $D'$ have been reevaluated this way. The results, shown in Figure 10, are conclusive—people prefer CRT-transformed categories by a large margin.
Figure 10. Results of the final verification of the ILSVRC2012 experiment. Even if the disjointed sample of the population is tested, they still prefer CRT-transformed categorization by a large margin. The blue bars denote the percentage of votes for the original category, and orange bars denote votes for the transformed category. Note that most of the previous categories did not get even a single vote.

7. VireoFood-172 Results and Discussion

Different from the previous experiment, this experiment involved two distinct populations of human subjects. The first population of 144 subjects was recruited from the UK and Ireland and denoted ENG. Their mean age was 31.45 ± 8.54. Of the total, 48% of participants were female and 52% were male; 67% of participants had normal vision, 24% had mild vision loss, and 9% moderate vision loss. The second population of 144 subjects was recruited from Malaysia and Singapore and denoted ASIA. Their mean age was 29.34 ± 7.69. Of the total, 45% of participants were female and 55% were male; 62% had normal vision, 27% had mild vision loss, 11% had moderate loss; and 0.6% had severe loss of vision.

Their self-identification by their diet is shown in Figure 11a. ANOVA statistical power analysis for diet groups (effect size $\eta^2 = 0.40$, $\alpha = 0.05$, statistical power $1 - \beta = 0.80$, and 4 groups) has shown that at least 76 subjects were needed to correctly identify the diet of the population. As 144 subjects for each demography were used, the statistical power for the ANOVA test rose to 0.99.

More than 80% of all subjects were identified as omnivorous and ate meat at least once in the last week. To test the significance of the diet preference a $\chi^2$ test of independence with null hypothesis diet preference is not associated with demography
was used. The hypothesis was rejected for both categories as the $p$-value was $>0.001$. The result shows that the population was appropriate to use in the experiment, as a significant amount of people eat meat.

Figure 11b represents answers about what did people eat in the last week at least once. The biggest differences between ASIA and ENG demography have been observed for soup, seafood, and fish food types, which is not considered a problem, as these categories were not considered in the experiment. Regarding meat, more than 80% were eating it in the last week. This also indicates suitability for using the selected populations in the experiment.

7.1. Training CNNs

Differently from the previous experiment, CNNs were additionally fine-tuned on the VireoFood-172 dataset. Training and validation accuracy for selected CNNs is shown in Figure 12a. Validation accuracy rose above train accuracy for all algorithms and had settled down by 60 epochs. As validation accuracy did not start to fall, the algorithms did not overfit. An additional confirmation for that are loss charts (Figure 12b). Validation loss was nearly the same as train loss, but it did not start to rise. Based on this analysis it can be concluded that the deep neural networks were trained relatively well. As accuracy did not change much from epoch 50 onwards, it was assumed that additional training would not improve the performance by a significant margin.

![Train and val accuracy](image)

(a) Accuracy

![Train and val loss](image)

(b) Loss

Figure 12. Accuracy (a) and loss (b) graphs of training and validating CNNs on VireoFood-172 subset. All graphs show relatively well-trained CNNs. The algorithms did not overfit, and additional training would not significantly improve the performance.

7.2. Pre-CRT Results

Original CFMs for ASIA population (Figure A4a in Appendix B) and ENG population (Figure A5a), obtained by subject categorization, have shown quite dispersed classifications with many mistakes. People disagreed to a certain extent as to which label should be appended to which image. Top-1 ACC, Precision, Recall, and F1-Score for ASIA population were 51.0%, 57.4%, 48.0%, and 51.7% respectively. This indicates a degree of disagreement among the human population. The results were slightly worse than the ones from the ILSVRC2012 experiment. This shows that the classification task was harder. Images are visually very similar and the type of meat is difficult to recognize from the image. Top-1 ACC, Precision, Recall, and F1-Score for ENG population were 40.4%, 47.0%, 38.0%, and 41.4% respectively. The original results for the ENG population were even worse. This was expected as food is only Asian and is considered harder to recognize by non-Asian subjects.

Using the human subject consensus (Algorithm 6 from Section 4), both CFMs (Figures A4b and A5b for ASIA and ENG population respectively) became diagonal matrices after adjusting them for population consensus. The analysis concluded there is no category, where more than 75% of the population would agree on a label change. This is true for both populations.
CFMs for AlexNet and ResNet deep neural networks (Figure A6) were very similar to each other, but metrics (see Section 7.4 for details) show ResNet outperformed AlexNet. Compared to ASIA and ENG human results, neural network performances on the original subset were worse. In addition to mistakes, many images were classified into categories that were out of the scope of the experiment and thus fell into the <unknown> category. The worst results for ResNet were obtained for <pork with garlic sauce>, where 22% of images were classified as <unknown>.

7.3. Deriving CRT Operations

Sixteen optimal clusters were determined by the elbow method and by cost function $P(W, Q)$ for both ASIA and ENG demography (Figure 13). The change in slope for selected points was not very noticeable as it was for six clusters, but the cost was significantly lower. The selected values also corresponded to the original number of categories.

![Optimal cluster number analysis](image)

**Figure 13.** Optimal cluster number analysis for the VireoFood-172 experiment. The analysis was done by the k-Modes algorithm for ENG population (a) and ASIA population (b). Sixteen clusters were selected (orange markings).

As before, despite creating 16 clusters, human labels were sometimes the same, so the clusters were merged and got 10 different categories for ENG population, and seven different categories for the ASIA population. The finding was surprising as more categories were suspected coming out of the ASIA population. Because the food represented was Asian, it was assumed ASIA subjects would have an easier time separating different dishes.

The differences of categories for each demography can be shown in Tables 5 and 6. Both populations defined <meat>, <vegetable>, <meatball>, and <noodle> as new categories. They contained similar images. ASIA population additionally defined <intestine>, <pork>, and <seafood> categories that were not present in ENG population. In contrast, <chicken foot>, <fry stir>, <soup>, <pasta>, <snail>, and <stew> were defined by ENG population and were not present in ASIA population.

7.4. Post-CRT Results

After transforming CFM$_{P1}$ (ENG demography) and ResNet CFM$_{M}$ (see Section 3 for details), the CFM$_{P1}$ (Figure A7a) remained the perfect diagonal matrix. ResNet CFM$_{M}$ (Figure A7b) also stayed similar at a glance, but the metrics in Table 7 shows there was a change in performance. Top-1 ACC increased by 0.7 pp, Precision dropped by 4.5 pp, Recall also dropped by 3.2 pp, and F1-Score dropped by 3.7 pp. Similar results were obtained from other CNN models. It was also observed that CFM$_{P1}$ does not have any <unknown> category, but ResNet’s does. This shows that humans identified all images by new categories and ResNet categorized some of the images to the categories that are out of the scope of the dataset $D'$. 
Table 5. Transformation of subset $D'$ labels into transformed labels for ENG demography.

| Original Categories | Transformed Category |
|---------------------|----------------------|
| <meat>              | <meat> <meatball>    |
| <braised intestines in brown sauce> | <chicken foot> |
| <fried lamb with cumin> | <noodle> |
| <roast chicken wings> | <stew> |
| <beefsteak>         | <snail> <vegetable> <fry stir> |
| <braised pork>      | /ok /ok /ok          |
| <braised beef with brown sauce> | /ok /ok |
| <sauteed bullfrog with pickled peppers> | /ok /ok /ok /ok |
| <pork with garlic sauce> | /ok /ok /ok /ok /ok /ok |
| <four-joy meatballs> | /ok /ok /ok /ok /ok /ok /ok /ok |
| <pork, shredded pork & vermicelli> | /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /ok /o...
a glance. ResNet’s Top-1 ACC (Table 8) was the same, but Precision, Recall, and F1-Score were lower than with the ENG population by 1.4 pp, 1.3 pp, and 0.8 pp, respectively.

7.5. Qualitative Results

Examples of images where CNN clearly missed the category are shown in Figures 14 and 15. The first image in Figure 14 represents classification of a <pork with garlic sauce> in ENG demography. Humans correctly classified it, but AlexNet CNN classified it as <shredded cabbage>, which is clearly incorrect, as only the meat supercategory was used. By transforming human and CNN predictions into new categories, humans still perfectly categorized the image. CNN now changed its result value to <unknown>. From a user’s perspective, it is considered a better result.

The appearance of ‘exotic’ foods (<chicken foot>, <snail>) after performing the CRT transformation that was derived using the ENG population was a clearly unexpected result, which was not observed in the ASIA group. According to results on ImageNet, a general trend towards less specific labels was expected, especially in the ENG population, which should be less familiar with Asian food. However, there may be a simple explanation: the ENG population is sensitive to the food that contains unfamiliar animals or unfamiliar parts of the animals since they want to avoid it—they want to know whether the food contains unfamiliar ingredients that may be unpalatable to them. This unpredictable result further emphasizes the importance of user tests, when evaluating classification algorithms.

### VireoFood-172 Image Classification (ENG Demography)

| Original categories | Transformed categories |
|---------------------|------------------------|
| **Pork with garlic sauce** | **Shredded cabbage** |
| **Fried lamb with cumin** | **Sauteed sea cucumber with scallion** |
| **Sauteed shredded pork with skin of tofu** | **Sauteed squid tentacles** |

| **Dataset** | **Human** | **CNN** |
|-------------|-----------|---------|
| Pork with garlic sauce | Pork with garlic sauce | Shredded cabbage |
| Fried lamb with cumin | Fried lamb with cumin | Sauteed sea cucumber with scallion |
| Sauteed shredded pork with skin of tofu | Sauteed shredded pork with skin of tofu | Sauteed squid tentacles |

| **Dataset** | **Human** | **CNN** |
|-------------|-----------|---------|
| Meat | Meat | Unknown |
| Meat | Meat | Unknown |
| Noodle | Noodle | Unknown |

**Figure 14.** Classification of selected images from VireoFood-172 dataset [58] by ENG population. When using original categories CNN classification are different from true categories. Images are only examples to visually represent a dataset category. They are not part of the dataset. <pork with garlic sauce> photo by Irrational cat and <fried lamb with cumin> photo by Alchen_x licensed under CC BY-SA 2.0, <sauteed shredded pork with skin of tofu> photo by T. Tseng licensed under CC BY 2.0.

The first image in Figure 15 represents a classification of a <four-joy meatballs> in the ENG demography. Humans still correctly classified it, but AlexNet CNN classified it as <braised pork>. By transforming human and CNN predictions into new categories, humans still perfectly categorized the image. CNN now changed its result value to <pork>. The reason, why CNNs transformed the label into <pork> is that the original prediction is part of the dataset D’ labels.
VireoFood-172 Image Classification (ASIA Demography)

| Original categories | Transformed categories |
|---------------------|------------------------|
| **Dataset**         | **Human**              | **CNN**          |
| Four-joy meatballs   | Four-joy meatballs     | Braised pork     |
| Pork with garlic sauce | Pork with garlic sauce | Fried and stewed hairtail in soy sauce |
| Sauteed shredded pork with garlic sprout | Sauteed shredded pork with garlic sprout | Steamed fish |

| **Dataset**         | **Human**              | **CNN**          |
|---------------------|------------------------|
| Meat                | Meat                   | Pork             |
| Meat                | Meat                   | Unknown          |
| Vegetable           | Vegetable              | Unknown          |

**Figure 15.** Classification of selected images from VireoFood-172 dataset [58] by ASIA population. When using original categories CNN classifications are different from true categories. Images are only examples to visually represent a dataset category. They are not part of the dataset. <four-joy meatballs> photo by Yoppy licensed under CC BY 2.0, <pork with garlic sauce> photo by Irrational Cat and <sauteed shredded pork with garlic sprout> photo by Kent Wang licensed under CC BY-SA 2.0.

8. Conclusions

The majority of the computer vision develops methods on the datasets with hard annotations. Hard annotations are the ones that are normally not changed. However, as this work showed, this may lead to sub-optimal results when such algorithms are used in real-world applications. A frequently heard prognosis that CNNs will soon outperform humans is to be interpreted strictly with the understanding that the current resemblance between evaluation methods and human reasoning and performance is not sufficient.

Some argue that the problem can be addressed by simply adding more data, but this route has its limits. It is unsustainable to use millions of images on models that do not generalize well and will eventually overfit. It should be minded that humans need only a couple of images to learn new concepts. Another possibility is to increase the model capacity, but this is unsustainable as well. There are already CNN models with millions of parameters and more than 100 layers. In many cases, the research resorts to active learning, but the validity of this approach is questionable for practical applications [59]. Active learning is prone to runaway feedback loops or even abuse. In practice, AI-powered devices should be ready for consumer use immediately after purchase, or the customers may feel they are not getting good value for money.

A deeper look into the problem of poor dataset annotations would better address the problem. One of the reasons behind poor annotations is the classical view of the categorization, which assumes that categories are based on shared properties. When creating benchmark datasets, the categories are normally chosen according to narrow research goals. Researchers obtain a huge amount of images (possibly from the internet) and label them as they see fit. However, newer research shows that categorization is, in fact, a matter of both human experience and imagination [18]. The preferred categorization between people with different experiences and cultures will be different. Therefore, it is essential to introduce the concept of a target user in popular benchmarks and performance tests. Category selection is, therefore, not a trivial task.

How to decide which possibility of categorization fits the target user population? Which one categorization will encourage algorithms to learn the answers that will not frustrate the target users? To answer these questions, this work has presented a novel method for the transformation of categorization tasks. The method can be used to transform a categorization problem, defined by
the dataset and human as a target-user, into the problem (dataset) that is more relevant for the target population. A sufficient number of samples have to be used to make the results statistically significant—this is a topic that is often overlooked in the comparison between human subjects and automated recognition algorithms.

8.1. Summary of Findings

• The performance of human observers on ImageNet problem is probably better than previously thought.

In [19] they have reported 94.9% Top-1 ACC, but this research has measured 99.4% Top-1 ACC. Note that there is a significant difference in the methodology. In [19] the metric was estimated from one “trained” human, who annotated 1500 images. A human was trained on 500 sample images. This work, on the other hand, used a much more sophisticated methodology. True classification tasks were used, where a human can choose among a set of categories and not only an annotation task which is similar to the labeling task. Using a classification task provides equal treatment of people and machines and is much more appropriate for comparison of humans to machines. Training humans, as done by [19], does not reflect the human experience and reasoning—it assumes the actual users of AI will yield to the methodology of AI community, and not vice-versa. Twenty-three participants per image were used and not only one, to properly sample experimental populations. The result of this work is, therefore, statistically significant and adapted to a selected population. The methodology of [19] did not even consider the question of different populations, and thus, their result could vary significantly if they would use different human annotators. Also note that their annotator labeled 1500 images, but each of the participants in this paper classified only 40 images, which is a considerable difference regarding human fatigue.

• The performance of CNNs is target user population-dependent.

The performance of CNNs changed in almost all cases after the user population re-targeting (that is, applying the CRT). The changes are especially striking in the VireoFood-172 experiment when CRT was derived from the ASIA population. It should be minded that improvements of just a few percentage points over the state-of-the-art provide good chances of acceptance to most competitive computer vision conferences, but here it is demonstrated that 7.5 pp change is possible just by adapting the categories to the target population. Improvements from the older AlexNet CNN architecture to much more advanced ResNet CNN are very slim when tested on a population re-targeted dataset.

• Not all dataset biases are created equal.

Dataset bias has a negative connotation in the field of AI—it suggests poorly curated data. However, this work essentially introduces the bias into the dataset, with the aim to bias the dataset towards human subjects, who will eventually be likely users of the technology in question. Even more, as the users may come from different sociological and cultural backgrounds, the dataset should be biased specifically with the target population in mind.

• User testing in the evaluation of CNNs is needed.

The qualitative analysis has shown some surprises that were not expected. The appearance of ‘exotic’ food categories in Chinese food after adapting the dataset to European users is a strong signal that only user tests can provide insight into what people expect from an AI-powered appliance. The findings of the research can be interpreted as: European users would like to be alerted to the food that they do not usually consume, and this is their serious concern when recognizing food.

• In many cases, mistakes, made by CNNs are grave.

The qualitative analysis also revealed that mistakes made by CNNs are grave in the sense that are below what humans expect from state-of-the-art technology (e.g., mixing fish and birds in ImageNet
test, mixing meat and sea cucumbers in the food recognition test). Figures 8, 14 and 15 show the natural consequence of CNNs relying only on low-level visual features, and still being unable to infer high-level concepts.

The research, documented in this article, presents a complete methodology that can be employed by anyone seeking to develop or deploy a user-friendly AI device or algorithm where the target-user of the system is a human.

The main disadvantage of the proposed method is using online user-studies as they cannot be fully controlled. Online user-studies produce many outliers as humans are not taking the experiment seriously enough. Furthermore, the cost of the user-studies can be prohibitive if statistical tests output a large number of subjects needed for the evaluation. As the datasets are getting larger by the day the method’s statistical power could become irrelevant. Next, the method uses the manual method to determine the optimal number of clusters. In this regard, it cannot be fully automated. Finally, there are multiple parameters that must be considered when designing tests for comparing humans to a machine.

Still, there is room for improvement. To overcome problems that are inevitably brought by using online experiments, one could use offline experiments for better control or more sophisticated web platforms that would automatically detect if humans are taking the experiment seriously.

Next, a better method to determine the number of clusters could be used, for example, adding the means to automatically seek a change of slope in a graph. The different cost functions for categorical data could be used in the clustering process, possibly reducing the number of clusters that got the same label in the labeling task. Using fuzzy set-based k-Modes algorithms [60] could also be considered.

The definition of a labeling task as an open-world problem poses certain challenges. People can label the image as anything, and it is down to a spell checker that will try to correct miss-spelled words. The online correction and suggestion tool would probably dramatically reduce the number of spelling errors and would give users a chance to benefit from the online spelling dictionary. Furthermore, different noise handling techniques for crowd-sourcing could be used [61].

Furthermore, there is still a question what will happen if CNN is trained on labels already adapted to the target user population? Similar to semantic segmentation, it could be only fine-tuned (only the top-most classification layers re-trained). The process is more time-consuming than the approach presented in this work, but it could bring improvements to the methodology.

What this work opens up is a new methodology to examine the differences between a human and a machine in object recognition tasks. With the preposition, the output of already pre-trained object classification algorithms is re-targeted to the user-friendly categorization. The methodology could also be efficiently used the other way around. A CNN pre-trained on an arbitrary dataset could be checked to which population the dataset fits the most and then used for that population. This is similar to transfer learning, except a CNN is transferred to the best target user population.

**Author Contributions:** Conceptualization, G.K., A.L. and J.P.; methodology, G.K. and J.P.; software, G.K.; validation, G.K., A.K. and J.P.; formal analysis, G.K.; investigation, G.K.; resources, G.K. and J.P.; data curation, G.K.; writing—original draft preparation, G.K.; writing—review and editing, A.K., A.L. and J.P.; visualization, G.K.; supervision, J.P.; project administration, J.P.; funding acquisition, J.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Gorenje, d. o. o. and by the Slovenian Research Agency (ARRS) research project J2-9433 (B) and research programs P2-0095 and P2-0246.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

**Dataset:** Research data related to this submission will be made available on request.
Abbreviations

The following abbreviations are used in this manuscript:

- **AI** artificial intelligence
- **CNN** Convolutional Neural Network
- **HMO** hierarchical neural network model
- **ID** identification number
- **ILSVRC2012** ImageNet Large Scale Visual Recognition Challenge 2012
- **mAP** mean average precision
- **ACC** accuracy
- **CFM** confusion matrix
- **CRT** cognitive relevance transform
- **pp** percentage points
- **ENG** English population
- **ASIA** Asian population

Appendix A. ILSVRC2012 Results

Confusion Matrices (CFMs) obtained by the methodology for the ILSVRC2012 dataset are presented here. The results can be easily reproduced for comparative studies. The matrix in Figure A1a is the average confusion matrix for human subjects, directly generated from the human subject classification and Figure A1b is the population confusion matrix, as follows from the transformation by Algorithm 6.

The confusion matrices for CNNs $\text{CFM}(\delta_{M}(D), D')$ were obtained simply by running respective algorithms on the validation set of the selected ILSVRC2012 categories. The worst and best results are shown in Figure A2.

Figure A1. Obtained confusion matrices (CFM) for human subjects from the ILSVRC2012 experiment. The matrix in (a) represents averaged results for all human subjects. (b) shows a population CFM.
Figure A2. Obtained confusion matrices for CNNs from ILSVRC2012 experiment. The matrix in (a) was derived from AlexNet classification results. (b) is ResNet’s CFM.

Results after applying CRT to the dataset $D'$ are shown in Figure A3a,b for humans and the ResNet, respectively.

Figure A3. The human population CFM (a) and ResNet CFM (b) after applying CRT to the dataset $D'$, which was derived from the ILSVRC2012 dataset.

Appendix B. VireoFood-172 Results

Confusion Matrices (CFMs) obtained by the methodology for the VireoFood-172 dataset are presented here. The results can also be easily reproduced for comparative studies. The obtained CFMs for ASIA and ENG subjects and populations are shown in Figures A4 and A5 respectively.
Figure A4. Average human (a) and population (b) CFMs for the ASIA population of the VireoFood-172 subset.
Figure A5. Average human (a) and population (b) CFMs for the ENG population of the VireoFood-172 subset.

CFMs for AlexNet and ResNet deep neural networks are shown in Figure A6.
Figure A6. AlexNet (a) and ResNet (b) confusion matrices in VireoFood-172 experiment.

Figure A7 shows ENG population and ResNet CFM which were changed by CRT obtained from ENG population.
ResNet CFM was also changed by CRT obtained from the ASIA population. It is shown in Figure A8b alongside ASIA population CFM.

Figure A7. The population (a) and ResNet (b) CFM from after applying CRT. CFMs are obtained from the ENG population and VireoFood-172 subset.

Figure A8. The population (a) and ResNet (b) CFM after applying CRT. CFMs are obtained from the ASIA population and VireoFood-172 subset.

References

1. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. Neurocomputing 2016, 187, 27–48.
2. Fleuret, F.; Li, T.; Dubout, C.; Wampler, E.K.; Yantis, S.; Geman, D. Comparing machines and humans on a visual categorization test. Proc. Nat. Acad. Sci. USA 2011, 108, 17621–17625.
3. Lake, B.M.; Salakhutdinov, R.; Tenenbaum, J.B. Human-level concept learning through probabilistic program induction. Science 2015, 350, 1332–1338.
4. Pramod, R.T.; Arun, S.P. Do Computational Models Differ Systematically from Human Object Perception? In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1601–1609.
5. Geirhos, R.; Janssen, D.H.J.; Schütt, H.H.; Rauber, J.; Bethge, M.; Wichmann, F.A. Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv 2017, arXiv:1706.06969.
6. Lake, B.M.; Ullman, T.D.; Tenenbaum, J.B.; Gershman, S.J. Building Machines That Learn and Think Like People. Behav. Brain Sci. 2017, 40, e253.
7. Das, A.; Agrawal, H.; Zitnick, L.; Parikh, D.; Batra, D. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? *Comput. Vis. Image Underst.* 2017, 163, 90–100.

8. Linsley, D.; Eberhardt, S.; Sharma, T.; Gupta, P.; Serre, T. What are the visual features underlying human versus machine vision? In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2706–2714.

9. Norman, D. *The Design of Everyday Things: Revised and Expanded Edition*; Basic Books: New York, NY, USA, 2013.

10. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In Proceedings of the Seventh International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.

11. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.

12. Wilson, D.; Sperber, D. Relevance Theory. In *The Handbook of Pragmatics*; Blackwell: Oxford, UK, 2004; Chapter 27, pp. 607–632.

13. Torralba, A.; Efros, A.A. Unbiased Look at Dataset Bias. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1521–1528.

14. Tommasi, T.; Patricia, N.; Caputo, B.; Tuytelaars, T. A Deeper Look at Dataset Bias. In *Domain Adaptation in Computer Vision Applications*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 37–55.

15. Zendel, O.; Honauer, K.; Murschitz, M.; Humenberger, M.; Dominguez, G.F. Analyzing Computer Vision Data—The Good, the Bad and the Ugly. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6670–6680.

16. Ponce, J.; Berg, T.L.; Everingham, M.; Forsyth, D.A.; Hebert, M.; Lazebnik, S.; Marszalek, M.; Schmid, C.; Russell, B.C.; Torralba, A.; et al. Dataset Issues in Object Recognition. In *Toward Category-Level Object Recognition*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 29–48.

17. Pinto, N.; Cox, D.D.; DiCarlo, J.J. Why is Real-World Visual Object Recognition Hard? *PLoS Comput. Biol.* 2008, 4, e27.

18. Lakoff, G. *Women, Fire, and Dangerous Things*; The University of Chicago Press: Chicago, IL, USA, 1987; p. 632.

19. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 2015, 115, 211–252.

20. Stock, P.; Cisse, M. ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. In *Computer Vision—ECCV 2018*; Springer: Munich, Germany, 2018; pp. 504–519.

21. Marcus, G. Deep Learning: A Critical Appraisal. *arXiv* 2018, arXiv:cs.AI/1801.00631.

22. Nguyen, A.; Yosinski, J.; Clune, J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 427–436.

23. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.

24. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.

25. Tabacof, P.; Valle, E. Exploring the Space of Adversarial Images. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 426–433.

26. Dodge, S.; Karam, L. A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions. In Proceedings of the 2017 26th International Conference on Computer Communication and Networks (ICCCCN), Vancouver, BC, Canada, 31 July–3 August 2017; pp. 1–7.

27. Dodge, S.; Karam, L. Can the Early Human Visual System Compete with Deep Neural Networks? In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2798–2804.
28. Wichmann, F.A.; Janssen, D.H.J.; Geirhos, R.; Aguilar, G.; Schütt, H.H.; Maertens, M.; Bethge, M. Methods and measurements to compare men against machines. *Electron. Imaging* 2017, 14, 36–45.

29. Kheradpisheh, S.R.; Ghodrati, M.; Ganjtabesh, M.; Masquelier, T. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Sci. Rep.* 2016, 6, 32672.

30. Kheradpisheh, S.R.; Ghodrati, M.; Ganjtabesh, M.; Masquelier, T. Humans and Deep Networks Largely Agree on Which Kinds of Variation Make Object Recognition Harder. *Front. Comput. Neurosci.* 2016, 10, 92.

31. Stabinger, S.; Rodriguez-Sánchez, A.; Piater, J. 25 years of CNNs: Can we compare to human abstraction capabilities? In *International Conference on Artificial Neural Networks*; Springer International Publishing: Cham, Switzerland, 2016; pp. 380–387.

32. Yamins, D.L.K.; Hong, H.; Cadieu, C.F.; Solomon, E.A.; Seibert, D.; DiCarlo, J.J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Nat. Acad. Sci. USA* 2014, 111, 8619–8624.

33. Cadieu, C.F.; Hong, H.; Yamins, D.L.K.; Pinto, N.; Ardila, D.; Solomon, E.A.; Majaj, N.J.; DiCarlo, J.J. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Comput. Biol.* 2014, 10, e1003963.

34. Rajalingham, R.; Issa, E.B.; Bashivan, P.; Kar, K.; Schmidt, K.; DiCarlo, J.J. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* 2018, 38, 7255–7269.

35. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988; p. 567.

36. Allaire, J. *RStudio: Integrated Development Environment for R*; RStudio, Inc.: Boston, MA, USA, 2016.

37. Team, R.C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2016.

38. Champely, S.; Ekstrom, C.; Dalgaard, P.; Gill, J.; Weibezahl, S.; De Rosario, H. *pur: Basic Functions for Power Analysis*; R Foundation for Statistical Computing: Vienna, Austria, 2016.

39. Faul, F.; Erdfelder, E.; Lang, A.G.; Buchner, A. *G*’* S* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 2007, 39, 175–191.

40. De Vos, N. *Kmodes*. 2016. Available online: https://github.com/nicodv/kmodes (accessed on 21 May 2020).

41. Huang, Z. Clustering large data sets with mixed numeric and categorical values. In Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference (PAKDD), Singapore, 23–24 February 1997; pp. 21–34.

42. Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov.* 1998, 2, 283–304.

43. Garbe, W. SymSpell. Available online: https://github.com/wolfgangbym/symspell (accessed on 5 February 2019).

44. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O’Reilly Media, Inc.: Boston, MA, USA, 2009.

45. Fellbaum, C. *WordNet: An Electronic Lexical Database*; MIT Press: Cambridge, MA, USA, 1998.

46. Chen, T.; Li, M.; Li, Y.; Lin, M.; Wang, N.; Wang, M.; Xiao, T.; Xu, B.; Zhang, C.; Zhang, Z. *MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems*. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Workshop on Machine Learning Systems (NIPS), Barcelona, 5–10 December 2016.

47. McKinney, W. *Data Structures for Statistical Computing in Python*. In *Proceedings of the 9th Python in Science Conference*; van der Walt, S., Millman, J., Eds.; SciPy: Austin, TX, USA, 2010; pp. 56–61, doi:10.25080/Majora-92bf1922-00a.

48. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* 2020, 17, 261–272. doi:10.1038/s41592-019-0686-2.
52. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825–2830.

53. Oliphant, T.E. *A Guide to NumPy*; Trelgol Publishing: Austin, TX, USA, 2006; Volume 1.

54. Van der Walt, S.; Colbert, S.C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* 2011, 13, 22–30.

55. Krizhevsky, A. *One Weird Trick for Parallelizing Convolutional Neural Networks*; Technical Report; Google Inc.: Mountain View, CA, USA, 2014.

56. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.

57. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *Computer Vision—ECCV 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.

58. Chen, J.; Ngo, C.W. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*; ACM: New York, NY, USA, 2016; pp. 32–41.

59. Settles, B. From theories to queries: Active learning in practice. *Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010*; Microtome Publishing: Brookline, MA, USA, 2011; pp. 1–18.

60. Yanto, I.T.R.; Ismail, M.A.; Herawan, T. A modified Fuzzy k-Partition based on indiscernibility relation for categorical data clustering. *Eng. Appl. Artif. Intell.* 2016, 53, 41–52.

61. Li, C.; Jiang, L.; Xu, W. Noise correction to improve data and model quality for crowdsourcing. *Eng. Appl. Artif. Intell.* 2019, 82, 184–191.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).