INCREMENTAL IMAGE LABELING VIA ITERATIVE REFINEMENT

Fausto Giunchiglia*, Xiaolei Diao* , Mayukh Bagchi*

* DISI, University of Trento, Italy
{fausto.giunchiglia, xiaolei.diao, mayukh.bagchi}@unitn.it
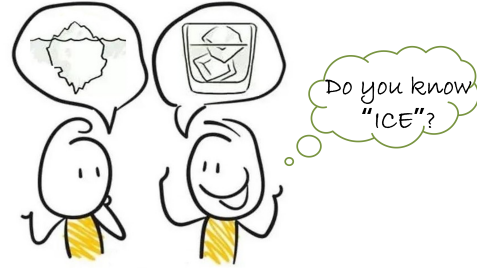
ABSTRACT

Data quality is critical for multimedia tasks, while various types of systematic flaws are found in image benchmark datasets, as discussed in recent work. In particular, the existence of the semantic gap problem leads to a *many-to-many mapping* between the information extracted from an image and its linguistic description. This unavoidable bias further leads to poor performance on current computer vision tasks. To address this issue, we introduce a Knowledge Representation (KR)-based methodology to provide guidelines driving the labeling process, thereby indirectly introducing intended semantics in ML models. Specifically, an iterative refinement-based annotation method is proposed to optimize data labeling by organizing objects in a classification hierarchy according to their visual properties, ensuring that they are aligned with their linguistic descriptions. Preliminary results verify the effectiveness of the proposed method.

Index Terms— Data Quality, Semantic Gap Problem, Image Datasets, Visual Properties

1. INTRODUCTION

Data is critical in machine learning (ML) systems, e.g., multimedia understanding, since it is one of the core infrastructures[1]. As a data-driven science, ML is affected by data quality, which in turn impacts downstream work. Sambasivan et al. [2] explore the downstream effects of data problems, and find data cascades usually exist in tasks that underestimate the quality of data. Thus, early intervention, especially in the data collection and labeling process which is the one step that impacts the downstream most, is of vital importance to improve the performance of various tasks.

Recent work reports various types of systematic flaws in the development of object recognition benchmark datasets. For example, Dimitris et.al. provide an extensive analysis of the labeling mistakes in ImageNet [3], and point out that noisy data collection pipelines lead to systematic misalignment between generated benchmarks and real-world tasks. Shreya et.al. [4] critically analyze the *geodiversity* of ImageNet as well as OpenImages [5] and show that they exhibit *Amerocentric* and *Eurocentric* representation bias. Lucas et. al. [6] describe the bias inherent in the annotation pipeline via which ImageNet was constructed. Terrance et. al. [7] report

Fig. 1: An interesting case of the semantic gap problem.

significant discrepancies in the classification accuracy of six *'in production'* object recognition systems, and ties the discrepancies to diversity in *socio-economic status*, *culture* and *language* from where the images were sourced.

Our intuition is that these flaws are grounded in the way language and perception interact. The key observation is that there is a misalignment between what computer vision (CV) systems perceive from media and the words that humans use to describe the same sources. Specifically, current datasets utilize words or phrases to label images. For example, all category labels in ImageNet are words/phrases taken from WordNet, and we call them *lexical labels*. During the labeling process, the use of such *lexical labels* will make a significant impact on the quality of the constructed dataset: the ground truth of the dataset will be directly affected by user experience, and users with different backgrounds may give inconsistent labeling results since they have different understandings to the same *lexical labels* and images [8]. This problem has been identified as *Semantic Gap Problem* (SGP) [9], where it was crystallized in [10] as the fact that there is a *many-to-many mapping* between the information extracted from the visual data and their possible contextual linguistic interpretations. The SGP is actually a consequence of the fact that the linguistic descriptions of an image are subjective and context-dependent. An interesting example is different objects icebergs and ice cubes are both regarded as "ice" in Figure 1. The SGP exists not only between different people but also appears in the same person in different scenarios.

To address the above issues, we introduce a knowledge representation (KR) theory [11] and approach into the process of data collection and labeling. We propose an image labeling process based on iterative refinement, aiming to generate high-quality ground truth datasets. The intuition stems

from the fact that the current labeling process is largely unspecified, leaving much freedom to annotator subjective judgments, who can then select among the many SGP mappings. In this paper, we focus on alleviating two types of ambiguity caused by SGP, i.e., *object ambiguity* and *visual ambiguity*. The idea is to apply KR-base methodology to introduce the human experience to provide guidance that drives the labeling process, thereby indirectly introducing *intended semantics* in the ML model, codified in a natural language format. During this process, machines and humans will refine the ground truth through iterative interaction and collaboration to obtain higher-quality datasets. We organize the labeling process based on a two-step labeling strategy, where each step is responsible for a specific aspect of the SGP many-to-many mapping, as (i) localizing objects in an image to eliminate a possible source of *object ambiguity*.(ii) identifying *visual properties* used to characterise objects rather than *lexical labels* to eliminate a possible source of *visual ambiguity*.

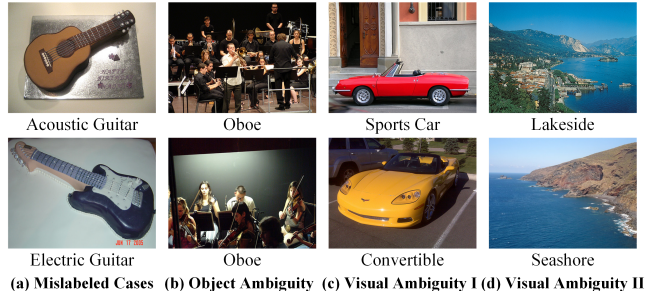The contributions of our work are summarized as follows:
- We introduce a KR-based methodology to guide image labeling and constitute a paradigm shift by attempting to integrate CV with KR.
- During dataset labeling, we refine the ground truth of the dataset by introducing humans into an iterative process, and improve dataset quality by asking humans to provide feedback and supervision.
- Preliminary experiment results demonstrate that the datasets constructed based on our proposed methodology keep a higher inter-annotator agreement.

The remainder of the paper is organized as follows. Sec. 2 the mislabeling types that arise in current image datasets. We illustrated the proposed labeling strategy in Sec. 3 and detail the labeling process in Sec. 4. Preliminary results are given in Sec. 5. Finally, Sec. 6 concludes the paper.

2. TYPES OF MISLABELING

We take ImageNet as an example to explore the ground truth in object recognition datasets and analyze the mislabeling in it. We divide the mislabeling cases into two types, i.e., object ambiguity and visual ambiguity. We give some examples of the above two types in Fig. 2, and analyze the details below. Note that in this process, we ignore the "simple" mistakes caused by the carelessness of the annotators since they are easily identifiable, such as an image labeled by ImageNet as *'acoustic guitar'* is a 'fake' guitar shaped on a birthday cake, as the two images shown in Fig. 2(a).

Object Ambiguity. This type arises from the presence of multiple objects in an image. It occurs when there is an *systematic incongruence* between the ImageNet label and the label of the most likely main object, as deemed by humans. An example is given in Fig. 2(b), which is labeled *'oboe'* in ImageNet but labeled *'orchestral'* by humans. In fact, empirical evidence from cognitive psychology [12] suggests that humans select the main object by perceptual attributes (*stimuli*) when looking to multi-object images, since it is usually

(a) Mislabeled Cases (b) Object Ambiguity (c) Visual Ambiguity I (d) Visual Ambiguity II

Fig. 2: Mislabeled samples of different types, where the labels are from ImageNet.

the most visually salient. However, we observe that ImageNet exhibits biases for many multi-object images. The labels of these images correspond to a *distinctive object* rather than the main object in the image, making it a challenge for models to extract features from the corresponding categories[13].

Visual Ambiguity. When annotators label images, ambiguity arises in the understanding of visual data. This is determined by the background and experience of different annotators. Images are visually polysemic when their visual *"semantics are described only partially"* [9]. Interpretations of images are not unique, so an image may be described with different labels by different annotators. As shown in Fig. 2(c), an example is labeled by different annotators as *'sports car'* and *'convertible'*, which are two same-level classes in ImageNet. Furthermore, the two similar images in Fig. 2(d), labeled *'seashore'* and *'lakeside* in ImageNet, are also representative examples that confuse the annotators. Such confusing class pairs are caused by *design factors* of ImageNet, namely labeled only based on *lexical labels*. As a result, choosing disjoint labels grounded in linguistic properties is insufficient for humans to visually disambiguate confusing class pairs in the face of potential overlapping image distributions.

Due to the semantic gap problem described above, even aside from occasional annotator errors, the resulting dataset may not accurately capture the ground truth, which seriously affects the quality of the dataset.

3. TWO-STEP LABELING STRATEGY

In this section, we introduce a two-step labeling strategy and analyze the reasons that this strategy is able to effectively bridge the mislabeling caused by the SGP. Before that, we summarize the characteristics of *good images* in a dataset by analyzing images in ImageNet with high inter-annotators consistency. Firstly, these images are almost always those containing a single object (the *'main object'*, as called in [13]). Secondly, these images are less noisy, in the sense that they have the minimum influence of confounding variables such as occlusion and clutter distorting them. Thirdly, all of these images are captured from an optimal viewpoint leading to clear visibility of their defining visual characteristic.

The proposed two-step labeling strategy is as follows:

- *S1: Object localization.* Identify object(s) in an image, thus eliminating the possible *object ambiguity*.
- *S2: Visual classification.* Identify the *visual properties* used to characterize objects, thus eliminating a possible source of *visual ambiguity*.

Note that such two steps are not splits of currently popular task *object detection*, but are designed based on different purposes. Localization is inherently visual, while classification is inherently *semantics-intensive* and is performed after object localization. Based on this distinction, and following the empirically validated theory of *teleosmantics* [11, 14], we distinguish two types of concepts, *substance concept* (SC) and *Category Concept* (CC) [15], to represent objects. As a result of object localization, SC can be seen as a visual representation of a single object, which is in turn amenable for visual classification. As the representations of entity concept-oriented language descriptions, CC is described by natural language descriptions (i.e., lexical labels) in the current datasets. In this process, given a continuous feed of images, the goal is to organize objects into a classification hierarchy based on their pre-defined visual features.

S1: Object Localization. Object localization refers to activities in which all objects in an image are localized (but not identified), for instance via bounding polygons [16]. This step aims to eliminate the possibility of *object ambiguity* by identifying and extracting relevant objects in the images, as far as possible to ensure that they satisfy the basic characteristics of a *good images*. During this process, SC is represented by localizing objects in multiple images, which does not distinguish between individuals (e.g., *'oboe#123'*) and classes (e.g., *'oboe'*). Meanwhile, the key to the continuous localization of a substance is grounded in its internal *causal factor* [11] which is incrementally manifested and extracted as perceivable visual properties.

S2: Visual Classification. The goal of step S2 is to identify the visual properties used to characterize objects, thus eliminating possible sources of *visual ambiguity*. We predefine two sets of visual properties for objects, called *visual genus* and *visual differentia*, and build *visual subsumption hierarchies* by perceiving visual properties from the new images. Specifically, *visual genus* is a set of visual properties shared across different objects, while *Visual differentia* refers to another set of visual properties different from *visual genus*, which are exploited to distinguish objects within the same genus. For example, the *visual genus* of the class *'lobster'* is "marine creatures with carapace", and one pre-eminent way of visually classifying further can be based on the *visual differentia* "the presence and shape of claws" with its different instantiations, i.e. "without claws" (Spiny lobster) and "with large tender claws"(American lobster).

One observation is that the many-to-many mapping of the SGP still occurs in the *good image* category, which is caused exactly by the choice of different differentia for the same genus. That is why the classification based on category labels

suffers from the SGP many-to-many mapping. SGP appears when annotators, even the same annotator, implicitly apply a different visual differentia to two images of the same object when selecting category labels based on their personal experience. Based on the above considerations, we make three fundamental assumptions in S2:

- The object hierarchy is built based on the visual genus and differentia of objects rather than category labels.
- The visual properties on which the differentia is computed are consistent across all objects in that category.
- The visual properties used to compute the differentia is consistent with the modeling decisions that are taken linguistically, i.e., with the genus and differentia defined by the gloss of the corresponding WordNet class.

Note that our approach differs significantly from mainstream CV, especially from the way ground truth datasets have been generated so far. Furthermore, the visual classification proceeds by a (successive) selection of visual differentia assume in an egocentric setting. In other words, how to select the visual differentia completely depend on the background, point-of-view, experience, and purpose of the annotators. In practice, the selection of what we define as visual differentia depends on the highly egocentric differentiation of *affordances* [17], with the guidance of cognitive psychology [18], in particular tends to visual properties that correspond to object functions (e.g., the visual property "a pair of joined reeds that vibrate together" for an oboe denotes the function of playing them to produce sound).

4. THE ITERATIVE LABELING PROCESS

In this section, we detail the image labeling methodology. This is an iterative refinement process, which includes three loops, namely the top-level loop, the vertical loop, and the horizontal loop. The first step object localization is applied in the top-level loop, and the second step visual classification produced by the vertical loop and horizontal loop.

The Top-level Loop. The top-level loop is designed to continuously offer *good images* for the iterative labeling method. In this process, the *object localization* step is introduced to avoid *object ambiguity*. An object localization model [19] is introduced to automatically locate objects by machine in an image, and crop the image based on the coordinates of objects, aiming to obtain images with a single object. As a result, we can obtain multiple single-object images from one multi-object image. Note that, although the square bounding box sometimes contains parts of other objects in the cropped image, the main object of these images will be more defined than the original image, thus, we treat the cropped image as a single-object image. Then, all images are input to the following vertical loop and horizontal loop one by one for labeling.

The Vertical Loop. The goal of the vertical loop is to iteratively refine the label of the input image through layer-by-layer computation to label it precisely. There is also a hierarchy that is input at the same time as the image, which will be

gradually enriched by adding nodes and samples in the current iterative process. When inputting a new object, the similarity is applied to compare with the categories already stored in the hierarchy, and the category with the highest similarity is regarded as the "candidate". Next, humans join this loop to determine whether the object has the common *visual genus* as the "candidate". If the answer is "False", proceed to compare this object to other categories at the same layer with "candidate" in the hierarchy, and enter the horizontal loop until "True" is obtained. None of "True" obtained means the object does not share a *visual genus* with any categories, then it will be labeled as a new category and stored in the hierarchy. During this process, the machine is responsible for recommending initial labeled options and performing the method, while humans take charge of determining the *visual genus* to decide whether further labeling refinement is needed.

The Horizontal Loop. The goal of the horizontal loop is to label the most refined category for the input object by comparing them within a domain category. It is triggered by the vertical loop and compares the input object with the subcategories of the candidate (if the candidate category has no subcategories, the input object is labeled as the candidate category). The process starts from the subcategories with the highest similarity of the object and asks humans whether they have *visual differentia*. If the answer is "False", it means that they belong to the same category, and the input object is labeled as the current subcategory; if the answer is "True", the same comparison is proceed with the next subcategory. If all subcategories in the candidate has *visual differentia* from the input image, it means the object does not belong to any subcategory, and it is labeled as a new category and added to the hierarchy as a new subcategory of the candidate. In the process, we complete the labeling and enrich the hierarchy at the same time. This incremented hierarchy will also continue to be utilized in the next top-level loop.

Observation and Analysis There are two important observations. Firstly, the fact that, though we have a detailed set of canonical principles for ensuring the *visual subsumption hierarchy* to be ontologically thorough, the task becomes particularly critical due to the tradeoff between the appropriate vertical and horizontal choice in uniquely classifying an object. The choice must be guided by the specific object recognition task that must be performed by the model trained using the dataset generated. In other words, there cannot be a *fits-it-all* dataset. Instead, we envisage a future where this methodology will allow the construction of datasets with clear and precisely specified *semantic* properties, which will then be introduced in the ML models by using them for training. Secondly, as a consequence of the first observation, human supervision can often be necessary for determining the exact (succession of) differentia in sync with the egocentric hierarchy in the mind of the user. The key observation underlying both observations, also factoring in other phases, is that the faceted classification process, while (of course) not eliminating human subjectiv-

Table 1: The image labeling results by two annotators, where "1, 1_1, ..." represent nine different categories, respectively, and the "Alpha" is Krippendorff's alpha measure.

Categories	1	1_1	1_1_1	1_1_1_1	1_1_1_2	1_1_2	1_1_3	1_2	1_3	Alpha
Expert1	17	42	21	21	22	13	12	33	10	0.9832
Expert2	17	42	20	22	22	13	12	33	10	

Table 2: Classification results on two datasets.

Methods	Accuracy				
	VGG [21]	GoogleNet [22]	ResNet [23]	RAN [24]	SENets [25]
ImageNet (subset)	0.699	0.727	0.538	0.706	0.734
Refine dataset (ours)	0.762	0.825	0.741	0.790	0.804

ity, does provide the guidelines for *enforcing a one-to-one mapping* between visual and linguistic properties. Overall, although the proposed iterative refinement process of image labeling still suffers from human subjectivity, it does provide the guidelines for *enforcing a one-to-one mapping* between visual and linguistic properties.

5. PRELIMINARY RESULTS

Inter-annotator agreement. To evaluate the inter-annotator agreement, we collect 191 musical instrument images of nine categories from ImageNet and invite two annotators to label these images based on our proposed iterative refinement process. After finishing labeling, the number of images for each category is shown in Table 1. We introduce Krippendorff's alpha[20] for agreement measure. Results as high as 0.9832 demonstrate near-perfect agreement between two annotators, which verifies the reliability of our method.

Machine Learning Experiment. In this experiment, we used two different datasets to train five classic ML methods. ImageNet (subset) refers to a subset containing nine musical instrument categories collected from ImageNet with the original labels. The refined dataset (ours) is labeled by our proposed iterative refinement strategy with the same images. During training, we keep all parameters consistent and use the same test set. The results of object recognition are shown in Table 2. It can be found that the accuracy of the same methods is significantly improved when trained on our dataset. These results confirm that our dataset has higher data quality, and verifies the validity of the proposed methodology.

6. CONCLUSION

In this paper, we propose a general iterative refinement process based on KR methodology to generate high-quality ground truth image datasets, aiming to overcome the limitations imposed by SGP on current labeling processes. In the future, we will focus on the construction of large benchmark datasets by extending the data labeling methodology.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Alon Halevy, Peter Norvig, and Fernando Pereira, "The unreasonable effectiveness of data," *IEEE intelligent systems*, vol. 24, no. 2, pp. 8–12, 2009.

[2] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *CHI*, 2021, pp. 1–15.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255.

[4] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley, "No classification without representation: Assessing geodiversity issues in open data sets for the developing world," *arXiv preprint arXiv:1711.08536*, 2017.

[5] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy, "Openimages: A public dataset for large-scale multi-label and multi-class image classification.," *Dataset available from https://github.com/openimages*, 2016.

[6] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord, "Are we done with imagenet?," *arXiv preprint:2006.07159*, 2020.

[7] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten, "Does object recognition work for everyone?," in *CVPR*, 2019, pp. 52–59.

[8] Fausto Giunchiglia, Mayukh Bagchi, and Xiaolei Diao, "Visual ground truth construction as faceted classification," *arXiv preprint arXiv:2202.08512*, 2022.

[9] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, "Content-based image retrieval at the end of the early years," *IEEE TPAMI*, vol. 22, no. 12, pp. 1349–1380, 2000.

[10] Fausto Giunchiglia, Luca Erculiani, and Andrea Passerini, "Towards visual semantics," *Springer Nature Computer Science (SNCS)*, vol. 2, no. 6, 2021.

[11] Fausto Giunchiglia and Mattia Fumagalli, "Concepts as (recognition) abilities," in *FOIS*, 2016, pp. 153–166.

[12] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem, "Basic objects in natural categories," *Cognitive psychology*, vol. 8, no. 3, pp. 382–439, 1976.

[13] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry, "From imagenet to image classification: Contextualizing progress on benchmarks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9625–9635.

[14] Ruth Garrett Millikan, "Neuroscience and teleosemantics," *Synthese*, pp. 1–9, 2020.

[15] Ruth Garrett Millikan, *Language: A biological model*, Oxford University Press on Demand, 2005.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[17] James J Gibson, "The theory of affordances," *Hilldale, USA*, vol. 1, no. 2, pp. 67–82, 1977.

[18] Carolyn F Palmer, Rebecca K Jones, Beth L Hennessy, Marsha G Unze, and Anne D Pick, "How is a trumpet known? the "basic object level" concept and perception of musical instruments," *The American journal of psychology*, pp. 17–37, 1989.

[19] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[20] John Hughes, "Krippendorffsalpha: An r package for measuring agreement using krippendorff's alpha coefficient," *The R Journal*, vol. 1, no. 1, 2021, Also: arXiv preprint arXiv:2103.12170.

[21] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *CVPR*, 2017, pp. 3156–3164.

[25] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.