様　式　C－19、F－19－1、Z－19　（共通）

科研費 KAKENHI

科学研究費助成事業　　研究成果報告書

令和　４年　５月３１日現在

機関番号：１４５０１
研究種目：研究活動スタート支援
研究期間：2019～2021
課題番号：１９K２４３４４

研究課題名（和文）ゼロショットの一般物体認識ための意味表現開発

研究課題名（英文）Zero-shot recognition of generic objects

研究代表者
ＨＡＳＣＯＥＴ　ＴＲＩＳＴＡＮ（Hascoet, Tristan）

神戸大学・経営学研究科・助教

研究者番号：６０８４８４４８

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：数年前まで、コンピューターは視覚を出来なかった。 たとえばある写真に人間が写っているかどうかを判断できなかった。十年程前、人、犬、車等、写真に写っている多くのものを高精度で認識できるようになった。自動運転車やロボット等の多くの技術の開発は視覚機能に限られていた：たとえば、自動運転車は、道路上の歩行者を認識できない場合、運転できません。
現在「男性」や「女性」等の限られた物の種類しか認識できないが人間は「自転車に乗っている若い女性」など、より詳細で多様性のあるものを認識できます。 この研究では、コンピューターがより複雑で事前定義されていないものを認識する能力を与えることに向けて取り組んできました。

研究成果の学術的意義や社会的意義
本研究では、写真から不特定の物の秋類を認識するために必要な情報を研究しました。 視覚情報の特定の処理は、意味表現よりも重要であり、そのようなプログラムを生成する能力は、実行できる計算の量によって制限されることがわかりました。

研究成果の概要（英文）：Until a few years ago, computers could not recognize things in pictures.
For example, computers were not capable to tell whether any human is in a given picture or not.
Around ten years ago, computer programs became capable to recognize a number of things in pictures with high precision, including humans, dogs, cars, etc. The development of many technologies such as self-driving vehicles and robots were previously limited by the inability of computers to recognize such objects: for example, a self-driving car can not drive if it can not recognize a pedestrian on the road.
However, computers can currently only recognize a finite number of things such as "a man" or "a woman", while humans can recognize things with more details and nuance such as "a young asian woman on a bike". This research project has worked towards giving computers the ability to recognize more complex and less predefined things, in order to allow computers to take better decisions.

研究分野：深層学習を用いたコンピュータービジョンのゼロショットラーニング

キーワード：Zero-Shot Learning　Self-Supervised Learning　Visual Representation　Feature Extraction　Semantic representations　Resource Efficiency　CNN　Computer vision

１．研究開始当初の背景

Object recognition is a foundational task for computer vision and artificial intelligence. In the past decade, Convolutional Neural Networks (CNN) have allowed for unprecedented progress in object recognition. CNN-based object recognition has become the backbone of modern computer vision: complex vision systems ranging from object detection and image segmentation systems to higher level models such as image captioning and Visual Question Answering systems, all rely on the backbone architecture of CNN classifiers. Hence, algorithmic progress on the core problem of object recognition has the potential to impact all downstream systems and applications.

One major limitation of current vision systems is that they can only recognize a finite set of visual concepts predefined by the available training data. In comparison, humans can continually define and recognize new object categories as they see them for the first time. For example, a child can recognize a zebra for the first time he sees one as his mother explains to him that “zebras look like horses with black and white stripes”. The idea of ZSL is inspired by this human ability to define and recognize new object categories from abstract descriptions. In ZSL, descriptions are referred to as semantic representations. Figure 1 illustrates the two-step process of zero-shot recognition on a toy example: In the training step, ZSL models learn a mapping between objects and their semantic representation from a set of known training classes. In the inference step, the model is shown images of new unknown classes. Using the mapping learned from the training classes, ZSL models map these inputs to their semantic representations to perform recognition. Given this formulation, ZSL research encompasses three key challenges:

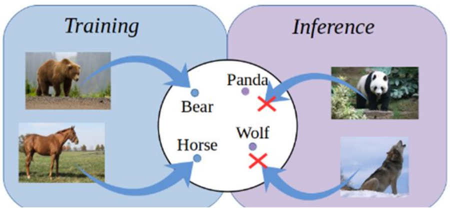

Figure 1. Illustration of ZSL process

1. Visual feature extraction: What feature representations of high-dimensional images yield best recognition accuracy and do these feature representations differ from visual features computed by traditional object recognition systems?
2. Semantic features extraction: What descriptions can enable zero-shot recognition of object categories? What feature representation of these descriptions yield best recognition ability?
3. Multi-modal mapping: How can we efficiently bridge the gap between high level semantic representations and low level image features?

２．研究の目的

Despite their great potential impact, and after a decade of active research, the accuracy of zero-shot recognition models remain too low to be considered for practical applications. The goal of this research is to improve the accuracy of zero-shot recognition models to enable their industrial deployment.

As illustrated in Figure 2, the architecture of ZSL models can be seen as the combination of three modules addressing each of the above key challenges. Concurrent works on ZSL use CNN as the visual module, word embedding models as their semantic module and focus on the core ZSL module, addressing key challenge

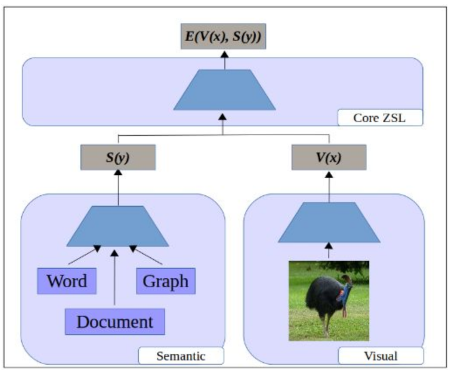

Figure 2. Illustration of ZSL architecture

３．研究の方法

Our research project is organized in three phases, as summarized in Table 1:

In phase 1, we will focus on the large-scale acquisition of semantic descriptions. Large-scale image recognition datasets feature tens of thousands of visual classes, which makes manual description annotations impractical. Hence the challenge of phase 1 consists in automating the semantic description acquisition process. In addition to word embeddings, we consider textual documents and knowledge graph as description modalities. Figure 3 illustrates these three modalities for the visual class “Cassowary”. We propose to leverage semantic web technologies to automate this large-scale acquisition process. Semantic web infrastructure provides an interlinking between the resources of large knowledge bases including Wikipedia, WordNet, and FreeBase. Following the links between these knowledge bases, as illustrated in Figure 4, we can map the visual classes of the ImageNet dataset to either Wikipedia articles or knowledge graph descriptions.

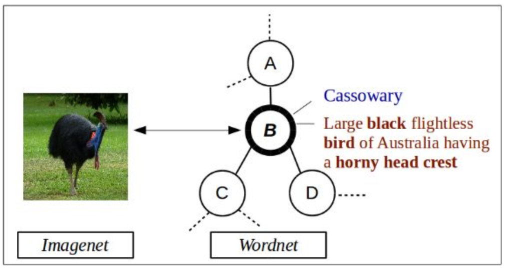

Figure 3. Illustration of class descriptions as words (blue), documents (red) and graph

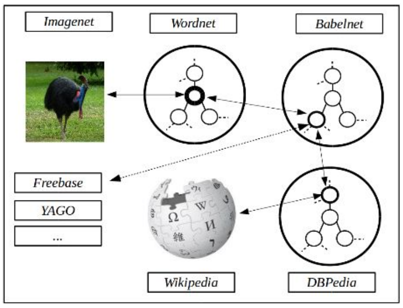

Figure 4. Illustration of data acquisition process through semantic web

Phase 2 focuses on the extraction of visual feature representation and the impact of different training paradigms (supervised, self-supervised, etc.) on Zero-Shot classification benchmarks. This phase will require optimization of our computational capacities so as to scale to large scale datasets

Finally, in the third phase, we will focus on the integration of the classifier into practical models, either semantic segmentation or object detection pipelines.

	Objective	Technologies	Start	End
Phase 1	Semantic feature learning	LOD, Semantic Web	June 2019	January 2020
Phase 2	Visual feature learning	Self-supervised learning CNN	February 2020	December 2020
Phase 3	Vision system integration	Segmentation & Detection	January 2021	December 2021

Table 1: Summary of research plan

4．研究成果

Initial efforts on visual feature extraction have to drastic memory requirement reduction: I proposed a family of architectures made of submodules whose computations either admit an analytical inverse or whose analytical inverse can be recovered with minimal memory cost. Using their analytical inverse, hidden activations necessary for the computations of the network's weight gradients can be backpropagated together with the gradient during the backpropagation step, hence bypassing the need to maintain these activations in memory. I characterized and derived a precise quantification of the numerical errors arising in the inverse reconstructions within long chains of invertible modules. I used this analysis to drastically reduce the GPU memory cost of training Convolutional Neural Networks.

This preliminary development then lead us to investigate the performance on self-supervised visual representations on the task of Generic Object ZSL (GOZ). We compared these features to traditional supervised representations and found they tend to perform better on standard zero-shot learning task whereas they do not match the supervised representations on the generalized zero-shot learning setting. Closing the gap between closely clustered supervised representations that perform well on training classes and more scattered unsupervised representations on the training classes while retaining higher accuracy on the unseen test classes has been identified as a promising research question.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Tristan Hascoet , Quentin Febvre , Weihao Zhuang , Yasuo Ariki,Tetsuya Takiguchi	４．巻 -
２．論文標題 Reversible designs for extreme memory cost reduction of CNN training	５．発行年 2022年
３．雑誌名 EURASIP Journal on Image and Video Processing	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Tristan Hascoet; Yihao Zhang; Andreas Persch; Ryoichi Takashima; Tetsuya Takiguchi; Yasuo Ariki
２．発表標題 FasterRCNN Monitoring of Road Damages: Competition and Deployment
３．学会等名 IEEE Big Data Cup Challenge 20201（国際学会）
４．発表年 2020年

１．発表者名 Tristan Hascoet, Quentin Febvre, Weihao Zhuang, Tetsuya Takiguchi, Yasuo Ariki
２．発表標題 Layer-Wise Invertibility for Extreme Memory Cost Reduction of CNN Training
３．学会等名 Neural Architects Workshop, International Conference on Computer Vision 2019（国際学会）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
フランス	Sicara			