Learning Without Data Collections *

Bhargav Ramudu Manam
M.Sc. Data Science and Artificial Intelligence

Université Côte d'Azur, 06410 Sophia Antipolis, France

1 Introduction

In this report, we ponder upon the concept of "Learning Without Data Collections" using the ideas from the lectures and the research materials, **Deep Learning to See- Towards New Foundations of Computer Vision** by (2) and (3). The authors made their discussions predominantly for Vision modality.

Currently in Computer Vision (CV), there are several deep learning (DL) methods based on convolutional neural networks (CNNs) that perform exceedingly well for several CV related tasks. Models based on CNNs are currently have the state-of-the-art (SOA) performance in several CV tasks such as Image classification, Object detection, etc. The authors acknowledge that the main reason for this is the great representational power of CNNs (scale and translational invariances) and also the elegance and efficiency (over classical numerical schemes) of Back propagation(BP). However, they argue that most of these significant results are based on the truly artificial supervised learning communication protocol, which is achieved due to usage of vast computational resources that is far from being natural. The SOA results are made possible due to highly complex Neural Network (NN) models of millions or billions of parameters trained on huge training datasets (with labelled images). The authors believe that solving Vision tasks using supervised learning techniques take lot of computational resources and more difficult to solve than the one offered by the Nature itself.

The main point of contention is that all these amazing DL or machine learning (ML) models can be deceptive and unreliable, in the sense that incorporating noise to the data often make these models easily prone to Adversarial attacks. Does that mean the features learned by these models are not robust enough? The authors opined that this can be due to several reasons. Namely:

1. it has to do with the diversifying (bottom-up) approach of pattern recognition community whose main focus is in details rather the underlying principles that drive the Vision.
2. all the analysis is done using the labelled images by simplifying the setup of Vision tasks as to identifying the patterns at spatial level by completely ignoring **time**, where in fact the true visual cognition happens in **motion**

* This report is to fulfill the requirements of **Advanced Learning** course instructed by **Prof. Marco Gori & Dr. Alessandro Betti** of Inria, Sophia Antipolis. Throughout this report, they are referred as the main authors along with others.

3. the studies by Crick (4), Poggio (7) and several others argue that the biological plausibility of BP does not hold true and hence the learning process that occurs in NN is not inspired by Nature.

In their work (2), they propose that the massive image supervision can be replaced with natural communication protocols arising from living in a visual environment with motion as the prominent driver of the feature learning process which takes spatio-temporal coherence into account, just like animals do. The authors aim to capture the object perception by following the Unifiers (Top-down) generalising approach. They pose **ten questions** that are helpful in driving the development of new approaches in Vision theory. The authors introduce two fundamental principles of motion invariance that drives the feature learning in visual perception (discussed in Section 3).

2 Learning Without Data Collections

2.1 What is learning without data?

Learning Without Data Collections means that it involves acquiring knowledge or skills without depending on conventional databases, where we need to collect high quality data and store them. This method acknowledges the significance of temporal progression and interaction in learning, particularly in sensory intense areas such as vision and speech. Instead of relying on an existing labelled database, the emphasis is on actively participating in the learning process in real-time.

2.2 Context

In the context of visual learning, it is observed that animals can gain visual skills through their interactions with the environment without accessing a stored visual database. This raise questions about the specific biological reasons behind this ability and whether machines can replicate it. Authors argue that even machines can acquire similar visual skills like animals by adopting a learning protocol that mirrors the continuous online processing through acquired video signal rather than random labelled images and involves interactions with humans (supervision) to describe the visual scenes. Figure (1) shows a context of learning without data, where once an agent is born it learns continuously in time by having interactions with its environment and some human supervision.

2.3 Ergodicity

Traditionally, in deep networks the learning is achieved by minimizing the empirical (or if possible functional risk) risk function (inverse of objective function) over the space (inputs and outputs) of all the training samples. But when it comes to learning without data, the risk is reformulated as the temporal empirical risk (Learning over time) and it is generalized and solved using Lagrangian

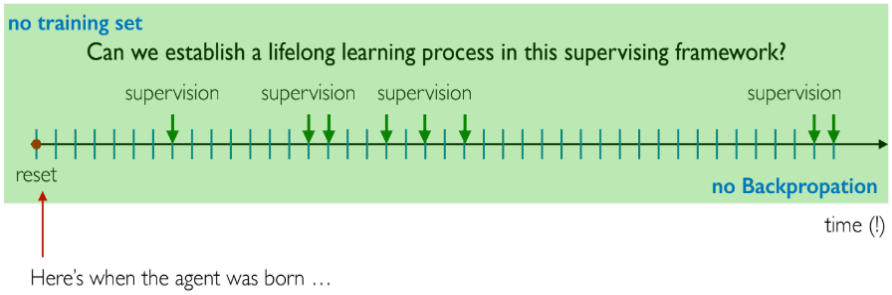

Fig. 1: Lifelong learning without data (figure taken from lecture slides). An agent is initialized at birth and made to learn in the visual environment through processing continuous video signal just like humans or animals. Occasional feedback is provided through human supervision.

formulations and optimal control theory. Moreover, the spatiotemporal features, learned through graph-based NN architectures, arise from the environmental interactions. The authors in (1), has shown that the BP diffusion that occurs over time, by the forward wave due to the input and the backward wave from the output, is biologically plausible. Further, the diffusion process is local in space and time i.e., they operate as neurons with time-delayed response and has cyclic connections similar to human brain. This biological plausibility affirms that the learning process in machines is similar to that observed in humans or animals. Moreover, they conclude that the BP that occurs in the classic NNs due to abrupt changes in the input by feeding the training set and by ignoring time is not biologically plausible and in fact is a degenerate case of the BP diffusion process that is seen when time is considered.

3 Motion Invariance

[1] The strong assumption made by the authors is that the visual perception solely depends on the motion and hence they claim that the only invariance required for machines to master the visual skills is *motion invariance*.

Object identity vs. abstract categorization (affordance): Visual learning without databases also involves the distinction between object identity and abstract categorization or its affordance. While humans can recognize the identity of objects regardless of different poses or environmental noise, the abstract

[1] In this section, the explicit and rigorous mathematical formulations or notations are willfully not presented in this report and assumed to be inline from the textbook as the emphasis is given to stating the main results that drive the motion invariance principles. This is done since the same notion is followed during the lectures by Prof. Marco Gori and his advice to focus on the conceptual understanding of the topic.

categorization of objects involves understanding their general category beyond specific instances. For example, recognizing "my own chair" in relation to the abstract notion of a chair requires a deeper understanding of object categories.

In this section, some main aspects of the visual learning along with two fundamental principles of motion invariance are discussed.

3.1 Main aspects of visual learning

Object identity: Humans and animals possess spectacular ability to identify objects consistently whether they are of different scales, present at various locations or shapes (due to rotation or deformation), partially visible (partly obstructed) and presence of environmental noise. Authors find that the one and only reason for this is none other than the motion and motion is responsible for various realizations of an object. Therefore, the visual learning process should develop features of an object that are motion invariant, these features define the identity of such an object.

Affordance: One of the important aspects of visual learning is the concept of affordance. When humans grasp the identity of an object they also gain abstraction of that object through the associated relationship presented by the object called affordance. In fact, humans in most cases characterize the object's identity by the affordance it offers rather than its identity. Not all objects are characterized by the same affordances. Some objects are attached and cannot be moved or easily broken, while others are detached and afford different actions. For example, a chair affords seating, but it can also be climbed on or used for defense. Understanding affordances goes beyond mere object recognition based on shape and appearance and requires perceiving the potential use cases and actions associated with different objects and scenes.

From material points to pixels: Optical flow In computer vision, we can always consider the case where the objects in the video are always moving even in the static case (because we implement the foveated nature of animals and humans in our machines as well, more on this later) except the ill-posed cases such as a moving white screen. This means that our machines are always in front of a optical flow, which can aid in acquiring visual information by relating the 2-D projection of material points (features) to the corresponding pixels. One approach to estimate the optical flow is by imposing the *brightness invariance*, which is defined as:

For Ω belonging to the retina (visible frame space of the machine) and some $t \in (O, T)$, where T is the total time of the video stream and for some pixel $x(t)$, belonging to the material point which evolves over time its brightness $b(x, t)$, at $(x, t) \in \Omega \times (0, T) =: \Gamma$ is constant i.e., $b(x(t), t) = c$, for some $c \in \mathbb{R}, \forall t \in (0, T)$ i.e.,

$$\frac{db(x(t), t)}{dt} = \nabla b.v + b_t = 0, \forall t \in (0, T) \tag{1}$$

Where, $v = \dot{x}$ is the velocity and has infinite solutions because of the scalar equation with two unknowns. This is overcome by Horn and Schunck in their work (5) by a regularization principle subject to some constraints on v. The authors further identified that the above brightness invariance does not take into account the cases where the lightening conditions change significantly and proposed the following reformulated version of the same invariance based on the R, G, B components of the pixel:

$$v.\nabla \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \frac{\partial}{\partial t} \begin{pmatrix} R \\ G \\ B \end{pmatrix} = 0 \tag{2}$$

3.2 The two principles of Motion Invariance

First Principle (I): Material Point Invariance (MPI) From the discussion on optical flow earlier, the authors conclude that the tracking of a pixels (trajectories) $x(t)$ is more directly related to objects and their features over the single material points. This is explained by the analogy that cat chases the mouse effectively by perceiving the overall picture of the mouse through the areas of spatially or temporally variable brightness.

The MPI covers the entire process involved in the *object identity*, which pairs the object features (along with its brightness) with its own velocity.

Let the visual features of an object $\mathcal{O}$ is given by $\Phi : \Gamma \to \mathbb{R}$ (assuming Φ is well defined on any (x, t) knowing the knowledge of the brightness for a given frame at time t, $b(., t)$). Given the features of $\mathcal{O}$, then the *consistency condition* (given below) for all of these features of $\mathcal{O}$ holds true, i.e.,

$$\Phi(x_\Phi(t), t) = \Phi(x_\Phi(0), 0) = c_\Phi, \forall t \in [0, T] \tag{3}$$

Where, $x_\Phi(t)$ represents the trajectory of the associated feature Φ and $c_\Phi \in \mathbb{R}$; if $x_\Phi(t) = x$ we let $v_\Phi(x, t) = \dot{x_\Phi}(t)$. Refer to the figure (2) below for better understanding of consistency condition. Φ can be seen as a signal that is travelling with velocity v_Φ and v_Φ can be computed by a function which depends on $D\Phi(., t) = (\nabla\Phi(., t), \Phi_t(., t))'$. Equation (3) considers a single trajectory but we are actually interested in considering multiple trajectories corresponding to each point of the retina (locally defined in time). Hence, for a given pair $(x, t) \in \Gamma$, $x_\Phi(t)$ is the trajectory for which $x_\Phi(t) = x \implies \dot{x_\Phi}(t) = v_\Phi(x, t)$. Given the optical flow v_Φ, we say that Φ is a *conjugate feature* with respect to $v_\Phi : \forall(x, t) \in \Gamma$, we have:

$$MPI \sim (\Phi \bowtie v_\Phi)(x, t) := \frac{d\Phi(x_\Phi(t), t)}{dt} = \nabla\Phi(x, t) * v_\Phi(x, t) + \Phi_t(x, t) = 0 \tag{4}$$

Where $\bowtie$: $\mathbb{R}^\Gamma \times (\mathbb{R}^\Gamma)^2 \to \mathbb{R}^\Gamma$ performs the mapping[2] $(\Phi, v_\Phi) \mapsto \Phi \bowtie v_\Phi$.

From equations (3) and (4) observe that:

[2] Here the symbol $\mathbb{R}^X$ denotes the set of all maps $f : X \to \mathbb{R}$

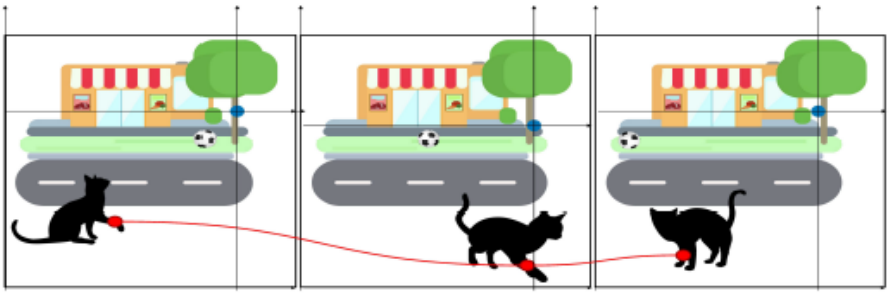

Fig. 2: MPI: **Consistent decisions** (figure taken from (2)), When considering any pixel, the correspondent visual feature does not change during its motion, which will be considered either with respect to the computer retina or to the focus of attention.

1. in case of no optical flow for some $\bar{x}, v_{\Phi}(\bar{x}, t) = 0 \forall t \in [0, T] \implies \Phi_t(\bar{x}, t) = 0$. This further results into $\Phi(\bar{x}, t) = c_{\Phi} \forall t \in [0, T]$ which is the consistency condition as one expects.
2. in case of a constant field $\Phi(x, t)$ in $C \subset \Gamma \implies \Phi \bowtie v_{\Phi} = 0$ on C, independent of v_{Φ}.
3. Also it is clear that, since $b \bowtie v = 0$ we can confidently state that b is in fact a conjugate feature of velocity v.

In essence, the pair (Φ, v_{Φ}) is an *indissoluble pair* (of features and velocities) playing the fundamental role in visual feature learning characterizing the object. Refer to the figure (3) for better understanding of this which presents a classic illusion example (Barber's pole).

When the pole rotates, it is difficult to track the individual pixels (moving horizontally) visually, but one can easily perceive the downward motion of the stripes (red or blue). Features like the stripes have their own velocities (vertically). The perception of the vertical (downward) movement is intrinsically linked to the presence and characteristics of the stripes (features). The vertical movement is directly associated with the presence and motion of the specific stripe being tracked. Likewise, when we perceive the presence and characteristics of a stripe, it naturally leads to the perception of its corresponding vertical movement. The two perceptions are intertwined and mutually dependent on each other, this characterizes their indissoluble relationship. Although achieving a similar conjunction of velocities and features in real-world scenes is challenging, the existence of distinguishing features implies their invariance.

Second Principle (II): Coupled Motion Invariance(CMI) Similar to object identification, shifting the attention at the pixel level (like foveated animals) helps in understanding the notion of affordance. The mechanisms involved in CMI are very similar to those discussed in MPI. We will now discuss CMI by considering a simple case of two entities (one human and one object) as shown

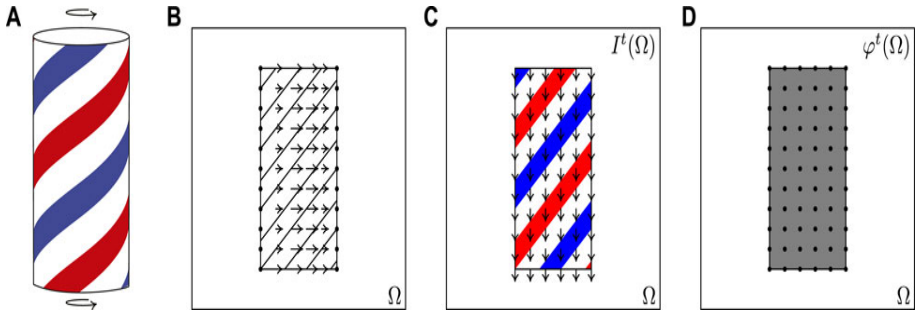

Fig. 3: MPI: **Indissoluble pair** (figure taken from (2)), Barber's pole example. (A) The $3-D$ object spinning counterclockwise.(B) The $2-D$ projection of the pole and the projected velocity on the retina Ω. (C) The brightness of the image and its optical flow pointing downwards. (D) A feature map that respond to the object and its conjugate (zero) optical flow.

in the figure (4), it is possible to have more than two entities in a real scenario. Let x be the pixel where a person with identity $p(x,t)$ and an object with identity $o(x,t)$ are identified. Denote $v_p(x,t)$ in pixel x at time t as the optical flow coming from person who is providing affordance to the object. By utilizing the conditions already established under principle I, let us consider a rapid movement. For some $(x,t) \in \Gamma$, the *coupling relation* $\between$ among the entities is defined as follows:

$$p \between o(x,t) := \gamma(p(x,t)) \wedge \gamma(o(x,t)) \tag{5}$$

Here, $\gamma : [0,1] \to \{0,1\}$ function returns a Boolean decision and is based on a suitable thresholding criteria. The significance of interaction between $p(x,t)$ and $o(x,t)$ is established by experimenting various degree of object coupling, depending upon the measure of $\mathscr{C}_{p \between o} = \{(x,t) \in \Gamma : p \between o(x,t) = 1\}$. A coupling $p \between o$ is $\epsilon-$significant, for some $\epsilon > 0$, given that[3] $\mathcal{L}^3(\mathscr{C}_{p \between o}) > \epsilon$ and is denoted by $p \between_\epsilon o$. Whenever this holds, it is convenient to introduce the notion of *degree of affordance* $\alpha_{op} \in \mathbb{R}$, of o conveyed by p. Then the Coupling Motion Invariance principle is given as follows:

$$CMI: \qquad \alpha_{op} \bowtie (v_o - v_p) = 0 \tag{6}$$

Observe that, when the interacting object p is stationary i.e., $v_p = 0$, then the CMI equation reduces to Principle I (MPI) $\implies \alpha_{op} \bowtie v_o = 0$, thus revealing the complimentary yet different nature of the two principles. By now, it is clear that Principle I (MPI) drives the object identification whereas the Principle II drives the object affordance.

Additionally, features developed due to motion invariance principles in different visual environments can be utilized to devise representations that facilitate

[3] $\mathrm{L}^3 here is the Lebesgue measure in \mathbb{R}^3$

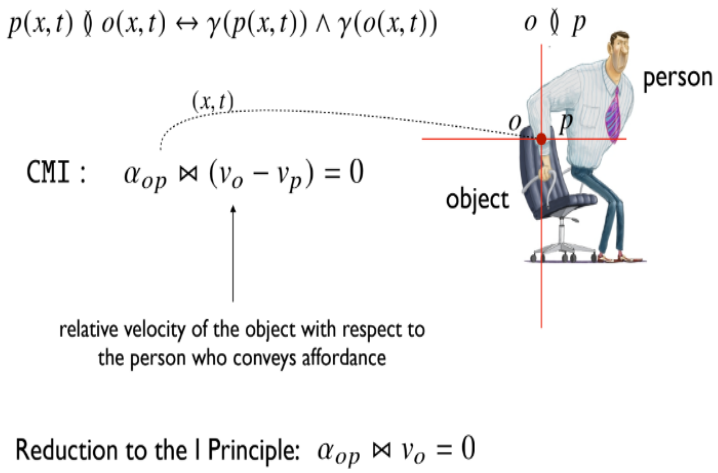

Fig. 4: Motion Invariance Principle II: Coupled Motion Invariance (CMI), figure taken from the lecture slides

subsequent learning processes suited to specific problems. This idea of leveraging previously learned features enhances the adaptability and problem-solving capabilities of machines. This is usually done as shown in figure (5), where the crucial point is *focus of attention* (FOA) at (x, t) that is responsible to perceive the optical flow. The authors showed that the all the information based laws of motion are driven by FOA. Moreover, after focusing on a certain pixel, the resolution on the retina is variable (foveated) i.e., high resolution at the center of the pixel and lower peripheral visual skills (as in animals). However, this compensated by the velocity of eye movements. Eye movements and FOA helps in estimating the probability distribution on the retina over time. Foveated eyes are proven to be very effective for scene understanding. The Foveated Neural Networks (FNNs) proposed by the authors are very effective computationally (due to variable resolution of the retina and FOA) and are argued to be biologically plausible. In turn, they also possess the hierarchical architecture paving the way for implementation of the motion invariances discussed earlier.

In the *virtuous loop of FOA*, at the birth of the agent the learning process is mainly driven by the brightness and its changes over spatio-temporal (optical flow) where attention is mainly given to the details and movements. This results in abstraction of very primitive virtual masses from these initial features. As the learning continues (through FNNs), i.e., the forward passes results in virtual masses that are more meaningful and depends on visual features that are superior than those learned only due to change in the brightness of the pixels whereas, the backward passes improves the FOA. This virtuous loop continues and the corresponding abstractions of virtual masses (acquired through even higher level

visual features) further improve the FOA. The authors interpret this interesting behaviour of the virtuous loop as the *duality principle* that regards motion and features as the two sides of the same medal. Meanwhile, during this virtuous loop directed supervision is provided at each step through scene interpretation and using interplay with language.

Moreover, the authors give emphasis to mastering the visual perception first before the introduction of language (similar to how babies master the vision). Further importance is given to learning in a truly experimental setting in the spirit of "en plein air" perspective that requires to moving into completely virtual visual environments.

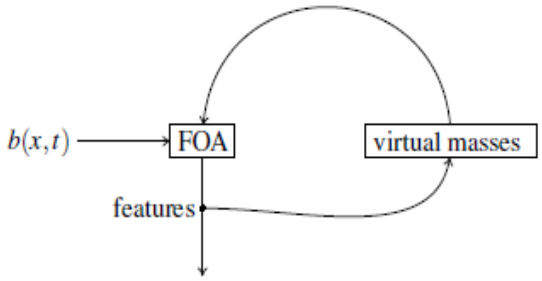

Fig. 5: The virtuous loop of FOA. After a phase in which FOA is driven by visual details that are regarded as virtual masses, as time goes by, the development of visual features results in new virtual masses that also provide focus of attention by means of the same attraction mechanisms.

3.3 Strengths

1. One obvious positives of learning over time (without data collections) is the biological plausibility of the FNNs and the Unifiers (top-down) approach of the authors truly paves the way for deep networks that have superior learning capabilities. Not to mention that this makes FNNs move a step towards achieving Artificial General Intelligence (AGI).
2. Architectural incorporation of the motion invariance principles and accommodation of FOA along with dropping the constraint for weight sharing (responsible for translational invariance) already makes FNNs favourable than CNNs, which are the current SOA architectures. Moreover, the computational cost FNNs is lower than CNNs due to variable retina resolution and FOA only at few selected locations (at a given time). Also by theoretical construction, the size of the FNNs models should be simpler than those of CNNs models solving the same problem, hence lower computational complexity.
3. Motion based visual perception seen in this report is more robust and less prone to Adversarial attacks than those of classical ones.

4. Learning without data collections benefits from data collection, quality, ethical, privacy, storage and management issues of data. Because of learning over time there is significant reduction in computation cost and resource utilization.

3.4 Weaknesses

1. **Flexibility:** As the models are designed through a generalised learning approach, these models aim to perceive all the visual abstractions at same time, though this is good but may be not suitable for simple tasks or tasks which does not need all the principles of motion invariance (or just need translation invariance).
2. **Learning in the wild:** In terms of implementation challenges, learning without databases encompasses several aspects. One challenge is designing visual agents that can operate in natural visual environments, such as for surveillance, ego-centric vision tasks or just in "en plein air" perspective. These agents should be able to recognize objects and provide pixel-level maps for object segmentation. The goal is to gain object recognition skills through human vocal interactions and supervision. This is something that is yet to be explored by the authors and there are not enough studies relating to this.
3. The motion invariance principles have ill-posed solutions, though these are less likely to encounter in nature. For example, the brightness invariance (including the color tracking) even after constrained regularization cannot guarantee singular solutions. This should not taken as very critical since the classical models are build with far worse assumptions without any regularity.
4. Though the crowd-sourcing performance evaluation scheme proposed by the authors is to be encouraged. Keeping in mind the novelty of the method, it is also important to test FNNs models on some popular vision datasets as this gives some traction in the ML community and helps to standout as an alternative to traditional CNNs based DL models.

4 Related Article

4.1 Continual Learning of Natural Language Processing Tasks: A Survey (6)

Continual Learning (CL) involves learning several tasks by a DL model over its life span. CL suffers with two key issues:

1. Catastrophic Forgetting (CF): model parameters are modified by a large extent as a result forgetting the past learnt tasks.
2. Knowledge Transfer (KT): CL should encourage both forward transfer (using previous knowledge from past tasks to learn the new task) and backward transfer (improving on previous tasks while learning similar new tasks) of knowledge.

Though the research article is about Natural Language Processing (NLP) and the survey contains lot of unrelated topics, I found the following similarities in relation to our discussion.

Similarities:

1. Generalising Approach and mimicking human behaviour: CL deals with a learning paradigm that tries to emulate human like capabilities (mimicking) in mastering multiple tasks (generalising) by knowledge acquisition through continuous learning in time without forgetting the past tasks and transferring the learnt knowledge to help with new tasks.
2. Language Invariance principles: The authors opined that many NLP tasks exhibit similar KT, which shows that sufficient level of abstraction independent of tasks have occurred during the learning of language i.e., the embedding dimension of the models has learnt enough semantics of the language.

5 Conclusion

With motion as the main proponent of visual perception, learning over time discussed in this report presents exciting possibilities in the field of Computer Vision that can truly revolutionize the learning process of visual tasks. The two motion invariance principles, MPI and CMI demonstrate the ability in object identification and object affordance, respectively. The FNNs equipped with variable resolution retina and focus of attention are capable of learning through back-propagation that is biologically plausible. FNNs with their ability to implement motion invariance principles and hierarchical architecture pose a serious challenge to CNNs dominance in the field of CV and therefore are worthy of extensive development and experimentation by the researchers.

To summarize, in this report an overview of learning without data collections taking place in sensory-rich domain of visual environments is presented. This novel approach of visual learning involves active engagement, interactions, and real-time processing of sensory information. Understanding affordances, distinguishing object identities from abstract categories, and utilizing transferable features are key elements in this approach. By addressing these challenges, machines can acquire visual skills that go beyond simple recognition and develop a deeper understanding of objects and scenes.

Bibliography

[1] Betti, A., Gori, M.: Backprop diffusion is biologically plausible. arXiv preprint arXiv:1912.04635 (2019)

[2] Betti, A., Gori, M., Melacci, S.: Deep Learning to See: Towards New Foundations of Computer Vision. Springer (2022)

[3] Betti, A., Gori, M., Melacci, S., Pelillo, M., Roli, F.: Can machines learn to see without visual databases? arXiv preprint arXiv:2110.05973 (2021)

[4] Crick, F.: The recent excitement about neural networks. Nature **337**, 129–132 (1989)

[5] Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1-3), 185–203 (1981)

[6] Ke, Z., Liu, B.: Continual learning of natural language processing tasks: A survey. arXiv preprint arXiv:2211.12701 (2022)

[7] Poggio, T.A., Anselmi, F.: Visual cortex and deep networks: learning invariant representations. MIT press (2016)