Concepts, Proto-Concepts, and Shades of Reasoning in Neural Networks

A. Augello, S. Gaglio, G. Oliveri, G. Pilato

University of Palermo

ICAR - CNR

gianluigi.oliveri@unipa.it salvatore.gaglio@unipa.it

June 22, 2018

Abstract

One of the most important functions of concepts is that of producing classifications; and since there are at least two different types of such things, we better give a preliminary short description of them both.

The first kind of classification is based on the existence of a property common to all the things that fall under a concept. The second, instead, relies on similarities between the objects belonging to a certain class A and certain elements of a subclass A_S of A, the so-called 'stereotypes.' In what follows, we are going to call 'proto-concepts' all those concepts whose power of classification depends on stereotypes, leaving the term 'concepts' for all the others.

The main aim of this article is showing that, if a proto-concept is given simply in terms of the ability to make the appropriate distinctions, then there are stimulus-response cognitive systems — whose way of manipulating information is based on Neural Networks (NN) — able to make the appropriate distinctions typical of proto-concepts in the absence of high-level cognitive features such as consciousness, understanding, representation, and intentionality. This, of course, implies that either proto-concepts cannot be given simply in terms of the ability to make the appropriate distinctions, or that we need to modify our traditional conception of mind, because the induction-like procedure followed by a NN in producing its classifications, far from being the ultimate product of a 'linguistic mind,' is, rather, inscribed in the nuts and bolts of the system's biology/electronics to which the NN belongs.

1 Introduction

A standard way of producing a classification is that obtained by means of sharp concepts. A concept C is *sharp* if and only if C refers to a property P such that, if $\mathcal{D}$ is the domain of quantification, there exists a class A such that $A = \{x \mid P(x)\}$, $A \cap \overline{A} = \emptyset$, and $A \cup \overline{A} = \mathcal{D}$. For example, if $\mathcal{D} = \mathbb{N}$, the concept *x is prime* is sharp. And we say that 'the number 2 falls under the concept *x is prime*' or, more simply, that '2 is prime.'

As is well known, when we are dealing with sharp concepts the law of excluded middle holds, whereas this is not the case with concepts that are not sharp like *x is a heap*. Within ordinary language we make an extensive, and productive, use of fuzzy concepts like *x is a heap*, *y is bald*, *z is tall*, etc. without even being aware of their problematic logical status.

However, besides the kind of classification we obtain by means of sharp/fuzzy concepts, there is a different type of classification which is not based on the existence of one and the same property common to all the members of a class A. But, it, rather, relies on similarities between the objects belonging to a certain class A and the elements of a subclass A_S of A, elements that we are going to call 'stereotypes.' By way of example, take A_s to contain two elements: a shark and a mullet. Starting from A_s we could generate A by taking some similarities between our stereotypes and other things.

Note that here by 'similarity between a and b' we mean a property common to a and b, where a and b belong to $\mathcal{D}$, e.g. being an animal, having fins, having scales, etc. Of course, more are the properties that a and b have in common the more similar a and b are to one another. The limiting case being that expressed by the identity $a = b$ where the objects denoted by a and b have the same properties, that is, when a and b denote the same object.

It is not difficult to see how, from such similarities with our stereotypes, we can generate a concept of fish given in terms of: animal living in water having either fins or scales or both. And, of course, from such a way of characterising the concept of fish it would follow that whales, dolphins, etc. are fish. But, we know that contemporary zoology has successfully challenged the above mentioned use of the term 'fish' by introducing a distinction between mammals and fish, a distinction according to which whales and dolphins are not fish, but mammals.

The quasi-accidental, purely phenomenical nature of the classifications obtained by means of brute correlations, such as those arising from mere similarities between objects and a set of stereotypes, has led us to call the cognitive representatives of these classifications 'proto-concepts.' We are going to use the term 'concepts' only for the cognitive representatives of all the other kinds of classification.

The main aim of this article is showing that, if a proto-concept is given simply in terms of the ability to make the appropriate distinctions, then there are stimulus-response cognitive systems[1] — whose way of manipulating information is based on Neural Networks (NN) — able to make the appropriate distinctions typical of proto-concepts in the absence of high-level cognitive features such as consciousness, understanding, representation, and intentionality. This, of course, implies that either proto-concepts cannot be given simply in terms of the ability to make the appropriate distinctions, or that we need to modify our traditional conception of mind, because the induction-like procedure followed by a NN in producing its classifications, far from being the ultimate product of a 'linguistic mind,' is, rather, inscribed in the nuts and bolts of the system's biology/electronics to which the NN belongs.

The present paper is a follow up to 'Wittgenstein, Turing, and Neural Networks' by G. Oliveri and S. Gaglio where, among other things, the authors endeavour to bring out the genuine cognitive character of Neural Networks (NN), cognitive character exhibited, primarily, by their ability to learn and being trained to perform a certain task.

2 The three main functions of concepts

Concepts have always played a central rôle in philosophy and, especially, in logic. One of the most important philosophical disputes in which medieval philosophers engaged — the so-called 'dispute about universals' — was directly related to the attempt to provide a plausible explanation of the classifying power of concepts. In fact, the *realists* argued, the reason why the concept *red* is so useful in classifying certain objects, separating them out from all the others, is that to such a concept there corresponds a property, a universal, that is present in all and only those things of which we correctly say that they are red.

On the other hand, the *nominalists* thought that, in contrast with what asserted by the realists, the universals do not exist. For if you take two red things a and b you, immediately, realise that the shade of red of a is different from that of b and that, therefore, 'red' is just a word, a name, to which no universal property corresponds. Therefore, according to the medieval nominalist, 'red' is a name whose usefulness in classifying boils down to the possibility of putting together all and only those things that are *similar* to one another with respect to (a certain) colour.

[1] See on this [13], Chapters 2 and 3.

Concerning the importance of concepts in logic, we can mention here, by way of example, the peculiar relation that, in pre-Fregean logic, a concept/predicate was supposed to have to the subject in a judgment, a relation exploited by Kant in the *Critique of Pure Reason* to draw the important distinction between analytic and synthetic judgments.

However, the modern theory of concepts starts with Frege for whom a concept is not an object, but a function[2] that takes proper names (or expressions performing the rôle of proper names) as arguments, and truth-values as values. From this it, immediately, follows that, according to Frege, *x is bald*, *y is a heap*, *z is tall*, etc. are not concepts, because the expressions '*a* is bald,' '*b* is a heap,' '*c* is tall,' etc. may not have a truth-value for certain proper names a, b, c.

Although in Frege we do not find the analytical philosopher's commitment to the idea that only a theory of language can provide a safe basis for the construction of a theory of thought,[3] nevertheless, for Frege, concepts have an essential function in thought that consists in presiding over the formulation and justification of judgments. It is only with the advent of *Gestalt* psychology,[4] Husserl's phenomenology,[5] and the philosophy of psychology of the later Wittgenstein — embodied, in particular, in the study of the phenomenon known as *seeing-as*[6] — that some non-idealist philosophers and psychologists discovered the very important rôle performed by concepts in perception.

Consider the Necker cube given in Figure 1. Now, apart from the well known possibility of shifting from perceiving face ABCD *as* 'coming forward' to seeing, instead, face 1234 *as* 'coming forward' — depending on which of the two faces you are focussing your attention — the really interesting thing here is that one of the necessary conditions for you to see the object in Figure 1 *as* a cube is having the concept of cube.
In fact, although a young child with no knowledge of mathematics would be able to perceive the face-shifting phenomenon, and draw a fairly resembling picture of the object in Figure 1; if he were asked to say what he sees the object *as*, he would probably reply 'a box,' 'a lump of sugar,' 'a brick,' 'a wire frame,' etc. but, certainly, not 'a cube,' because he does not know what a cube is.

If what we have said so far is correct, concepts have at least three different

[2]See [4] and [5].

[3]See on this [3], especially Chapter 10.

[4]See on this [20] especially in relation to the impact that *Gestalt* psychology has on what he calls 'productive thinking.'

[5]See on this [7], Part Three, Chapter Three.

[6]See on this [22], Part II, Chapter XI.

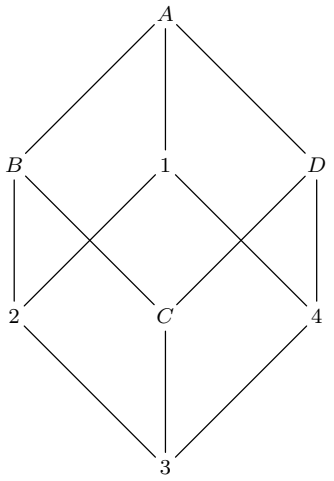

Figure 1: Necker cube

important functions: classifying, being integral part of judgments, affecting some of our perceptions. Of these three functions of concepts two of them — being integral part of judgments, affecting some of our perceptions — presuppose the existence of a cognitive system able to produce judgments/thoughts and representations of objects. The classifying function, instead, seems to us to be somewhat independent of the existence of such a cognitive system. For, although, when a competent speaker of English says 'Socrates is a man' the very meaningfulness of the assertion, and of the thought expressed by it, presuppose the speaker's ability to classify some of the objects of his domain of quantification as men; and that when someone sees something as a cube, the very possibility of his perception depends on his ability to classify certain objects of his domain of quantification as cubes; the ability, for example, to classify something as a fish (see §1) may not presuppose either judging or seeing-as (representing).

Perhaps, some light on these matters will be obtained from considering how concepts are given to us, because in so doing we might come across concepts that are given to us as means of classification and not as instruments for judging or representing.

3 Concepts and proto-concepts

In investigating how concepts are given to us, one of the obvious things to look at is language. Two of the main concept-producing devices present

in language are the so-called 'prototypes,'[7] and 'stereotypes.'[8] A prototype is an individual belonging to the domain of quantification that has certain features 'at their best.' Take, by way of example, $\mathcal{D}$ to be the set of the elements of the colour spectrum projected on to a wall by a prism when this is hit by a pencil of light rays (see Figure 2). The colour spectrum $\mathcal{D}$, with the Euclidean distance defined on it, is a metric space $(\mathcal{D}, d)$.

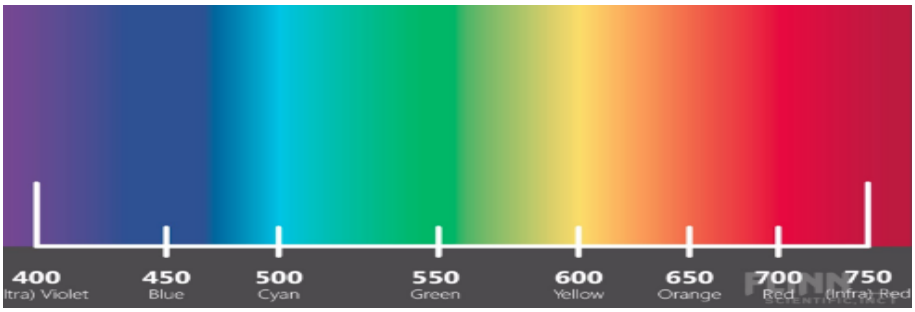

Figure 2: Colour spectrum

Now, for each colour band i present in the colour spectrum, where $i \in \mathbb{R}$, choose the middle element of the band as the prototype of that colour, and call it '$\mathbf{p}_i$.' If $k, \mathbf{p} \in \mathcal{D}$, where $\mathbf{p}$ represents a prototype, as is shown by the colour spectrum, the shorter is the distance between k and $\mathbf{p}$ the more similar the colour represented by k is to the colour represented by $\mathbf{p}$. Clearly, if $\mathbf{p}_i$ is the prototype of the red colour band then the set $R = \{x \,|\, d(x, \mathbf{p}_i) < d(x, \mathbf{p}_j)$, for any j such that $j \neq i\}$ can be considered as a *classification* of elements of $\mathcal{D}$ which can, eventually, be turned into the extension of a concept and, precisely, of the concept *x is red*.

A stereotype, on the other hand, is an individual $\mathbf{s}$ belonging to the domain of quantification that appears to have certain features, but such features are not necessarily given at their best in $\mathbf{s}$. Take $(\mathcal{D}, d)$ to be, as above. If $k, \mathbf{s} \in \mathcal{D}$, where $\mathbf{s}$ represents a stereotype then, as above, the smaller is the distance between k and $\mathbf{s}$ the more similar the colour represented by k is to the colour represented by $\mathbf{s}$. As colour stereotypes $\mathbf{s}$ consider the colours of the paints present on the palette of a Renaissance painter.

Assuming that on the palette of a Renaissance painter there could be a finite number of colour stereotypes $\mathbf{s}_1, \ldots, \mathbf{s}_n$, and that $\mathbf{s}_i$ is the only stereotype

[7] On a related concept of prototype see [6], especially Chapter 3, §3.9. See also [10] on prototypes and theory-like representations; and [18] on incorporating prototype theory in Convolutional Neural Networks.

[8] On a linguistic concept of stereotype see [16].

of red, we have that $C_i = \{x \mid d(x, \mathbf{s}_i) < d(x, \mathbf{s}_j)$, for any j such that $j \neq i\}$ is a *classification* of elements of $\mathcal{D}$ which can be, eventually, turned into the extension of the concept *x is red*.

Although, as we have seen in the examples above, both prototypes and stereotypes generate potential extensions of concepts, there are some analogies and differences existing between these two types of objects that deserve some attention.

One of the main differences between prototypes and stereotypes is that there is only one prototype of something, but you could have several stereotypes of the same kind of thing. Indeed, the prototype of red is that electromagnetic wave having a wavelength of 700 nanometers (see Figure 2), whereas the so-called 'Titian red' and 'Pompeian red' are two different possible stereotypes of red. Of course, if there is more than one stereotype of, say, red the classification induced by all the stereotypes of red available is going to be the union of the classifications induced by each single stereotype, and the larger is the number of different stereotypes of red the better are the chances of producing a correct classification of red objects.

Secondly, the very idea of 'features at their best,' present in the definition of prototype, requires some form of theorising and judging to distinguish, for example, between a fin at its best and a fin that is not at its best. On the other hand, when it comes to producing stereotypes, the situation looks rather different.

To see this consider the well known ethological phenomenon of imprinting. Imprinting is that type of learning that takes place only within a certain number of hours from birth. As Konrad Lorenz has shown,[9] if, within a certain number of hours from its birth, a gosling is exposed to a human being, rather than to a goose, it ends up considering the human being as its parent following him, etc. (see Figures 3 and 4). In other words, imprinting is that form of learning whereby geese, and other animals, form a stereotype not only of their mother/parent, but also of the species to which they belong. This has two very important consequences for us. First, the phenomenon of imprinting, clearly, shows that the formation of some stereotypes is brute, that is, it does not take place through theorising, or any sort of reasoning or representation influenced by previously acquired concepts.

Secondly, there are stereotypes, some of which play a crucial rôle in producing very important, basic classifications, that are independent of a linguistic mind.

But, be as it may with regard to the connection between imprinting and stereotypes formation, we believe that, if sufficiently many samples are

[9] [11].

Figure 3: A goose and her goslings

Figure 4: Lorenz and his goslings

provided, then the classification obtained of the elements of the domain of quantification $\mathcal{D}$ can be turned into the extension of the corresponding proto-concept by means of CNNs. Although the following section gestures in this direction, substantiating this claim will be one of the objects of our future investigations.

4 CNNs and Inductively Generated Proto-Concepts?

Traditional Feed-Forward Neural Network architectures receive a single vector as an input and process it through a series of hidden layers. Each hidden layer is constituted by a set of artificial neurons, where each unit is fully connected to all the units belonging to the previous layer. The neural units belonging to the same layer make their computation in parallel with the other units of the same layer. Furthermore, the neurons of the same layer do not share any connections.

Traditional feed-forward architectures do not perform well on image recognition and image segmentation tasks. In the last years a category of Neural Network architectures, known in the literature as Convolutional Neural Networks (CNNs) and inspired by the mammalian visual system [23][Fukushima,1980] [25], have proven to be very effective in performing tasks like image recognition and segmentation. The very first convolutional neural network architecture was LeNet, developed by Yann LeCun and it was effectively used for character recognition tasks. This kind of architectures gave rise to the general paradigm of Deep Learning.
The main operations performed in Convolutional Neural Network are Convolution, Pooling or Sub Sampling, and Classification.
Traditionally, ConvNets assume that the kind of inputs are images.
A typical representation of a convolutional network is shown in Figure 5.

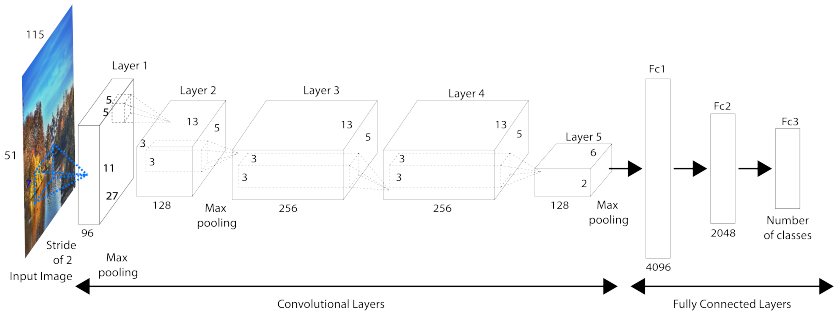

Figure 5: A typical representation of a CNN

The network processes the original image layer by layer from the original pixel values to the final classification output. The input layer specifies the dimensions of the input images. CNNs derive their name from the "convolution" operator. The main goal of Convolution in this kind of architectures is to extract features from the input image. The Convolution operation allows the network to learn image features using small portions of input data.
A set of parameterized kernels constitutes the convolution layer. Every kernel is spatially small, and it is applied to the whole image through a spanning process.
The input image is convolved with these multiple learned kernels which exploit shared weights. This operation generates a bidimensional activation map that gives the responses of that filter at every spatial position. The size of the kernels gives rise to the locally connected structure and produces a set of feature maps.

Then, the pooling layer reduces the size of the image, attempting to maintain the information.
The combination of the convolution and the pooling layers realizes the feature extraction part. Subsequently, the features are weighted and combined in the fully-connected layer, which constitutes the classification layer of the network [26].
Convolutional Neural Networks (CNNs) are powerful models capable of achieving outstanding results in particular for image classification and segmentation tasks. Nowadays this kind of neural architectures has been extremely successful in identifying faces, objects, traffic signs and are widely used in vision for robots and self-driving cars. Moreover, ConvNets have been effectively used also in several Natural Language Processing tasks as well.
Furthermore, they are capable to effectively extract features from images: pre-trained models are used as generic feature extractors[29]. This goal is reached by removing the last layer which gives the output classification scores. The activations from the last fully connected layer define the features extracted from the input image [31].
These kinds of features extracted from pre-trained CNN have been successfully used in computer vision tasks such as scene recognition or object attribute detection, yielding better results concerning traditional handcrafted features [29].
Moreover, Athiwaratkun et al. [30] have also demonstrated that Random Forest and SVM can be used with features extracted from CNN to obtain better a prediction accuracy compared to the original CNN.
Recent results indicate that very deep networks achieve even better results on various benchmarks [27], [28]. One drawback of this trend, however, is the time required to train such kind of neural architecture.

Summing up we can say that CNNs can be used to perform a generalization/classification based on the stereotypes belonging to the training set and, as discussed in this section, with considerable improvements with regard to other traditional types of neural networks (even if the limitations typical of neural networks, relating to the size of the training set and the curse of dimensionality, still remain). This is a good baseline to investigate, in future works, if the classification obtained by a CNN can be turned into the extension of corresponding proto-concepts.

5 Conclusions

This is a philosophy and ethology inspired paper relating to cognition. Starting from a discussion of various kinds of classification, we are led to distin-

guishing between classifications operated on the basis of concepts, and classifications driven by what we have called 'proto-concepts.' The difference between the two different kinds of classifications being that whereas concepts appeal to the existence of a property common to all the objects falling under them, proto-concepts, instead, derive their classification power from a set of what we call 'stereotypes' and the relevant similarities existing between these stereotypes and the objects falling under the proto-concepts.

Having discussed the difference between prototypes and stereotypes and their rôle in producing classifications — classifications that are presented as potential extensions of proto-concepts — we discuss the ethological phenomenon of imprinting discovered and studied by Konrad Lorenz. As is well known, imprinting is that cognitive phenomenon whereby goslings, if exposed to a certain object K within a certain time from birth — a goose, a human being, etc. — elect K as a stereotype (in our sense) of parent/representative of their species and behave accordingly following K, etc. This, of course, implies that goslings subject to imprinting classify K, and themselves, as belonging to the same class $\mathcal{K}$ that becomes the potential extension of a proto-concept.

We then engage in a discussion of a particular type of Neural Network, the so-called 'Convolutional Neural Network' (CNN). What we intend to show in our discussion of CNNs is that cognitive agents that operate on the basis of CNNs are able to produce classifications typical of proto-concepts in the absence of high-level cognitive features such as consciousness, understanding, representation, and intentionality. On the basis of this result we ask ourselves whether this means that proto-concepts cannot be given simply in terms of the ability to make the appropriate distinctions or that we should, instead, modify our traditional conception of mind.

References

[1] Bonomi, A. (ed.): 1973, *La struttura logica del linguaggio*, Valentino Bompiani, Milano.

[2] Dummett, M. A. E.: 1991, *The Logical Basis of Metaphysics*, Duckworth, London.

[3] Dummett, M.A.E.: 1993, *Origins of Analytical Philosophy*, Harvard University Press, Cambridge, Massachusetts.

[4] Frege, G.: 1973, 'Funzione e Concetto', in [1], pp. 411–423.

[5] Frege, G.: 1973, 'Concetto e Oggetto', in [1], pp. 373–386.

[6] Gärdenfors, P.: 2004, *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, Massachusetts.

[7] Husserl, E.: 1998, *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy*, transl. by F. Kersten, first book, Kluwer Academic Publishers, Dordrecht.

[8] Kant, I.: 1787 (1990), *Critique of Pure Reason*, transl. by Norman Kemp Smith, Macmillan, London.

[9] Kohonen, T.: 2001, *Self-Organizing Maps*, Springer, Berlin.

[10] Lieto, A.: 2018, 'Heterogeneus Proxytypes Extended: Integrating Theory-like Representations and Mechanisms with Prototypes and Exemplars,' *BICA* 2018, Springer, Advances in Intelligent Systems and Computing.

[11] Lorenz, K.: 2012, *L'anello di Re Salomone*, Adelphi eBook, Milano.

[12] McCulloch, W., and Pitts, W.: 1943, 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *Bulletin of Mathem.Biophysics*, 5:115-133.

[13] Nilsson, N. J.: 2002, *Intelligenza artificiale*, edited by S. Gaglio, Apogeo, Milano.

[14] Oliveri, G..: 1984, 'Le *Ricerche* di Wittgenstein nella lettura di S. Kripke', *Paradigmi*, Anno II, n. 6.

[15] Oliveri, G. & Gaglio, S.: *In press*, 'Wittgenstein, Turing, and Neural Networks', *Giornale di Metafisica*, vol. 1, 2018.

[16] Putnam, H.: 1975, 'The meaning of 'meaning", *Minnesota Studies in the Philosophy of Science*, vol. 7, pp. 131–193.

[17] Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: 1986, 'Learning Internal Representations by Error Propagation', in Rumelhart, D. E., and McClelland, J. L. (Eds.) (1983), *Parallel Distributed Processing*, MIT Press, Boston, Vol. 1, pp. 318-362.

[18] Saleh, B., Elgammal, A., Feldman, J.: 2016, 'Incorporating Prototype Theory in Convolutional Neural Networks', Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence.

[19] Varela, F. J., Thompson, E., Rosch, E.: 1993, *The Embodied Mind*, The MIT Press.

[20] Wetheimer, M.: 1965, *Il pensiero produttivo*, transl. by M. Giacometti and R. Bolletti, Giunti e Barbera, Firenze.

[21] Wittgenstein, L.: 1981 (1921), *Tractatus Logico-Philosophicus*, transl. by D.F. Pears & B.F. McGuinness, with the introduction by B. Russell, Routledge & Kegan Paul, London & Henley.

[22] Wittgenstein, L.: 1983, *Philosophical Investigations*, Basil Blackwell, Oxford.

[23] D. H. Hubel and T. N. Wiesel, Receptive fields of single neurones in the cats striate cortex, J. Physiol., vol. 148, no. 1, pp. 574591, 1959.

[Fukushima,1980] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern., vol. 36, no. 4, pp. 193202, Apr. 1980.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE, vol. 86, no. 11, pp. 22782324, Nov. 1998.

[26] Lars Hertel, Erhardt Barth, Thomas Kater, Thomas Martinetz, "Deep Convolutional Neural Networks as Generic Feature Extractors" 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, 2015, pp. 1-4.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, presented at the Workshop ImageNet Large Scale Visual Recognition Challenge (ILSVRC), 2014.

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

[29] Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off the- shelf: an astounding baseline for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806813.

[30] B. Athiwaratkun, K. Kang, "Feature representation in convolutional neural networks", CoRR, vol. abs/1507.02313, 2015.

[31] Garcia-Gasulla, Dario, Ferran Pars, Armand Vilalta, Jonatan Moreno, Eduard Ayguad, Jess Labarta, Ulises Corts, and Toyotaro Suzumura. "On the Behavior of Convolutional Nets for Feature Extraction." Journal of Artificial Intelligence Research 61 (2018): 563-592.