
In machine learning and computer vision , M-theory is a learning framework inspired by feed-forward processing in the ventral stream of visual cortex and originally developed for recognition and classification of objects in visual scenes. M-theory was later applied to other areas, such as speech recognition . On certain image recognition tasks, algorithms based on a specific instantiation of M-theory, HMAX, achieved human-level performance. [1]

The core principle of M-theory is extracting representations invariant under various transformations of images (translation, scale, 2D and 3D rotation and others). In contrast with other approaches using invariant representations, in M-theory they are not hardcoded into the algorithms, but learned. M-theory also shares some principles with compressed sensing . The theory proposes multilayered hierarchical learning architecture, similar to that of visual cortex.

Intuition

Invariant representations

A great challenge in visual recognition tasks is that the same object can be seen in a variety of conditions. It can be seen from different distances, different viewpoints, under different lighting, partially occluded, etc. In addition, for particular classes objects, such as faces, highly complex specific transformations may be relevant, such as changing facial expressions. For learning to recognize images, it is greatly beneficial to factor out these variations. It results in much simpler classification problem and, consequently, in great reduction of sample complexity of the model.

A simple computational experiment illustrates this idea. Two instances of a classifier were trained to distinguish images of planes from those of cars. For training and testing of the first instance, images with arbitrary viewpoints were used. Another instance received only images seen from a particular viewpoint, which was equivalent to training and testing the system on invariant representation of the images. One can see that the second classifier performed quite well even after receiving a single example from each category, while performance of the first classifier was close to random guess even after seeing 20 examples.

Invariant representations has been incorporated into several learning architectures, such as neocognitrons . Most of these architectures, however, provided invariance through custom-designed features or properties of architecture itself. While it helps to take into account some sorts of transformations, such as translations, it is very nontrivial to accommodate for other sorts of transformations, such as 3D rotations and changing facial expressions. M-theory provides a framework of how such transformations can be learned. In addition to higher flexibility, this theory also suggests how human brain may have similar capabilities.

Templates

Another core idea of M-theory is close in spirit to ideas from the field of compressed sensing . An implication from Johnson–Lindenstrauss lemma says that a particular number of images can be embedded into a low-dimensional feature space with the same distances between images by using random projections. This result suggests that dot product between the observed image and some other image stored in memory, called template, can be used as a feature helping to distinguish the image from other images. The template need not to be anyhow related to the image, it could be chosen randomly.

Combining templates and invariant representations

The two ideas outlined in previous sections can be brought together to construct a framework for learning invariant representations. The key observation is how dot product between image I {\displaystyle I} and a template t {\displaystyle t} behaves when image is transformed (by such transformations as translations, rotations, scales, etc.). If transformation g {\displaystyle g} is a member of a unitary group of transformations, then the following holds:

■ g I , t ■ = ■ I , g – 1 t ■ (1) {\displaystyle \langle gI,t\rangle =\langle I,g^{-1}t\rangle \qquad (1)}

In other words, the dot product of transformed image and a template is equal to the dot product of original image and inversely transformed template. For instance, for image rotated by 90 degrees, the inversely transformed template would be rotated by –90 degrees.

Consider the set of dot products of an image I {\displaystyle I} to all possible transformations of template: { ■ I , g ′ t ■ ■ g ′ ∈ G } {\displaystyle \lbrace \langle I,g^{\prime }t\rangle \mid g^{\prime }\in G\rbrace } . If one applies a transformation g {\displaystyle g} to I {\displaystyle I} , the set would become { ■ g I , g ′ t ■ ■ g ′ ∈ G } {\displaystyle \lbrace \langle gI,g^{\prime }t\rangle \mid g^{\prime }\in G\rbrace } . But because of the property (1), this is equal to { ■ I , g – 1 g ′ t ■ ■ g ′ ∈ G } {\displaystyle \lbrace \langle I,g^{-1}g^{\prime }t\rangle \mid g^{\prime }\in G\rbrace } . The set { g – 1 g ′ ■ g ′ ∈ G } {\displaystyle \lbrace g^{-1}g^{\prime }\mid g^{\prime }\in G\rbrace } is equal to just the set of all elements in G {\displaystyle G} . To see this, note that every g – 1 g ′ {\displaystyle g^{-1}g^{\prime }} is in G {\displaystyle G} due to the closure property of groups , and for every g ′ ′ {\displaystyle g^{\prime \prime }} in G there exist its prototype g ′ {\displaystyle g^{\prime }} such as g ′ ′ = g – 1 g ′ {\displaystyle g^{\prime \prime }=g^{-1}g^{\prime }} (namely, g ′ = g g ′ ′ {\displaystyle g^{\prime }=gg^{\prime \prime }}). Thus, { ■ I , g – 1 g ′ t ■ ■ g ′ ∈ G } = { ■ I , g ′ ′ t ■ ■ g ′ ′ ∈ G } {\displaystyle \lbrace \langle I,g^{-1}g^{\prime }t\rangle \mid g^{\prime }\in G\rbrace =\lbrace \langle I,g^{\prime \prime }t\rangle \mid g^{\prime \prime }\in G\rbrace } . One can see that the set of dot products remains the same despite that a transformation was applied to the image! This set by itself may serve as a (very cumbersome) invariant representation of an image. More practical representations can be derived from it.

In the introductory section, it was claimed that M-theory allows to learn invariant representations. This is because templates and their transformed versions can be learned from visual experience – by exposing the system to sequences of transformations of objects. It is plausible that similar visual experiences occur in early period of human life, for instance when infants twiddle toys in their hands. Because templates may be totally unrelated to images that the system later will try to classify, memories of these visual experiences may serve as a basis for recognizing many different kinds of objects in later life. However, as it is shown later, for some kinds of transformations, specific templates are needed.

Theoretical aspects

From orbits to distribution measures

To implement the ideas described in previous sections, one need to know how to derive a computationally efficient invariant representation of an image. Such unique representation for each image can be characterized as it appears by a set of one-dimensional probability distributions (empirical distributions of the dot-products between image and a set of templates stored during unsupervised learning). These probability distributions in their turn can be described by either histograms or a set of statistical moments of it, as it will be shown below.

Orbit O I {\displaystyle O_{I}} is a set of images g I {\displaystyle gI} generated from a single image I {\displaystyle I} under the action of the group G , ∀ g ∈ G {\displaystyle G,\forall g\in G} .

In other words, images of an object and of its transformations correspond to an orbit O I {\displaystyle O_{I}} . If two orbits have a point in common they are identical everywhere, [2] i.e. an orbit is an invariant and unique representation of an image. So, two images are called equivalent when they belong to the same orbit: I ∼ I ′ {\displaystyle I\sim I^{\prime }} if ∃ g ∈ G {\displaystyle \exists g\in G} such that I ′ = g I {\displaystyle I^{\prime }=gI} . Conversely, two orbits are different if none of the images in one orbit coincide with any image in the other. [3]

A natural question arises: how can one compare two orbits? There are several possible approaches. One of them employs the fact that intuitively two empirical orbits are the same

irrespective of the ordering of their points. Thus, one can consider a probability distribution P I {\displaystyle P_{I}} induced by the group's action on images I {\displaystyle I} (g I {\displaystyle gI} can be seen as a realization of a random variable).

This probability distribution P I {\displaystyle P_{I}} can be almost uniquely characterized by K {\displaystyle K} one-dimensional probability distributions P ■ I , t k ■ {\displaystyle P_{\langle I,t^{k}\rangle }} induced by the (one-dimensional) results of projections ■ I , t k ■ {\displaystyle \langle I,t^{k}\rangle } , where t k , k = 1 , … , K {\displaystyle t^{k},k=1,\ldots ,K} are a set of templates (randomly chosen images) (based on the Cramer–Wold theorem [4] and concentration of measures).

Consider n {\displaystyle n} images X n ∈ X {\displaystyle X_{n}\in X} . Let K ≥ 2 c ε 2 log ■ n δ {\displaystyle K\geq {\frac {2}{c\varepsilon ^{2}}}\log {\frac {n}{\delta }}} , where c {\displaystyle c} is a universal constant. Then

with probability 1 – δ 2 {\displaystyle 1-\delta ^{2}} , for all I , I ′ {\displaystyle I,I^{\prime }} ∈ {\displaystyle \in } X n {\displaystyle X_{n}} .

This result (informally) says that an approximately invariant and unique representation of an image I {\displaystyle I} can be obtained from the estimates of K {\displaystyle K} 1-D probability distributions P ■ I , t k ■ {\displaystyle P_{\langle I,t^{k}\rangle }} for k = 1 , … , K {\displaystyle k=1,\ldots ,K} . The number K {\displaystyle K} of projections needed to discriminate n {\displaystyle n} orbits, induced by n {\displaystyle n} images, up to precision ε {\displaystyle \varepsilon } (and with confidence 1 – δ 2 {\displaystyle 1-\delta ^{2}}) is K ≥ 2 c ε 2 log ■ n δ {\displaystyle K\geq {\frac {2}{c\varepsilon ^{2}}}\log {\frac {n}{\delta }}} , where c {\displaystyle c} is a universal constant.

To classify an image, the following "recipe" can be used:

Memorize a set of images/objects called templates;

Memorize observed transformations for each template;

Compute dot products of its transformations with image;

Compute histogram of the resulting values, called signature of the image;

Compare the obtained histogram with signatures stored in memory.

Estimates of such one-dimensional probability density functions (PDFs) P ■ I , t k ■ {\displaystyle P_{\langle I,t^{k}\rangle }} can be written in terms of histograms as μ n k (I) = 1 / | G | ∑ i = 1 | G | η n (■ I , g i t k ■) {\displaystyle \mu _{n}^{k}(I)=1/\left|G\right|\sum _{i=1}^{\left|G\right|}\eta _{n}(\langle I,g_{i}t^{k}\rangle)} , where η n , n = 1 , … , N {\displaystyle \eta _{n},n=1,\ldots ,N} is a set of nonlinear functions. These 1-D probability distributions can be characterized with N-bin histograms or set of statistical moments. For example, HMAX represents an architecture in which pooling is done with a max operation.

Non-compact groups of transformations

In the "recipe" for image classification, groups of transformations are approximated with finite number of transformations. Such approximation is possible only when the group is compact .

Such groups as all translations and all scalings of the image are not compact, as they allow arbitrarily big transformations. However, they are locally compact . For locally compact groups, invariance is achievable within certain range of transformations. [2]

Assume that G 0 {\displaystyle G_{0}} is a subset of transformations from G {\displaystyle G} for which the transformed patterns exist in memory. For an image I {\displaystyle I} and template t k {\displaystyle t_{k}} , assume that ■ I , g – 1 t k ■ {\displaystyle \langle I,g^{-1}t_{k}\rangle } is equal to zero everywhere except some subset of G 0 {\displaystyle G_{0}} . This subset is called support of ■ I , g – 1 t k ■ {\displaystyle \langle I,g^{-1}t_{k}\rangle } and denoted as supp ■ (■ I , g – 1 t k ■) {\displaystyle \operatorname {supp} (\langle I,g^{-1}t_{k}\rangle)} . It can be proven that if for a transformation g ′ {\displaystyle g^{\prime }} , support set will also lie within g ′ G 0 {\displaystyle g^{\prime }G_{0}} , then signature of I {\displaystyle I} is invariant with respect to g ′ {\displaystyle g^{\prime }} . [2] This theorem determines the range of transformations for which invariance is

guaranteed to hold.

One can see that the smaller is supp ■ (■ I , g − 1 t k ■) {\displaystyle \operatorname {supp} (\langle I,g^{-1}t_{k}\rangle)} , the larger is the range of transformations for which invariance is guaranteed to hold. It means that for a group that is only locally compact, not all templates would work equally well anymore. Preferable templates are those with a reasonably small supp ■ (■ g I , t k ■) {\displaystyle \operatorname {supp} (\langle gI,t_{k}\rangle)} for a generic image. This property is called localization: templates are sensitive only to images within a small range of transformations. Although minimizing supp ■ (■ g I , t k ■) {\displaystyle \operatorname {supp} (\langle gI,t_{k}\rangle)} is not absolutely necessary for the system to work, it improves approximation of invariance. Requiring localization simultaneously for translation and scale yields a very specific kind of templates: Gabor functions . [2]

The desirability of custom templates for non-compact group is in conflict with the principle of learning invariant representations. However, for certain kinds of regularly encountered image transformations, templates might be the result of evolutionary adaptations. Neurobiological data suggests that there is Gabor-like tuning in the first layer of visual cortex. [5] The optimality of Gabor templates for translations and scales is a possible explanation of this phenomenon.

Non-group transformations

Many interesting transformations of images do not form groups. For instance, transformations of images associated with 3D rotation of corresponding 3D object do not form a group, because it is impossible to define an inverse transformation (two objects may looks the same from one angle but different from another angle). However, approximate invariance is still achievable even for non-group transformations, if localization condition for templates holds and transformation can be locally linearized.

As it was said in the previous section, for specific case of translations and scaling, localization condition can be satisfied by use of generic Gabor templates. However, for general case (non-group) transformation, localization condition can be satisfied only for specific class of objects. [2] More specifically, in order to satisfy the condition, templates must be similar to the objects one would like to recognize. For instance, if one would like to build a system to recognize 3D rotated faces, one need to use other 3D rotated faces as templates. This may explain the existence of such specialized modules in the brain as one responsible for face recognition . [2] Even with custom templates, a noise-like encoding of images and templates is necessary for localization. It can be naturally achieved if the non-group transformation is processed on any layer other than the first in hierarchical recognition architecture.

Hierarchical architectures

The previous section suggests one motivation for hierarchical image recognition architectures. However, they have other benefits as well.

Firstly, hierarchical architectures best accomplish the goal of ‘parsing’ a complex visual scene with many objects consisting of many parts, whose relative position may greatly vary. In this case, different elements of the system must react to different objects and parts. In hierarchical architectures, representations of parts at different levels of embedding hierarchy can be stored at different layers of hierarchy.

Secondly, hierarchical architectures which have invariant representations for parts of objects may facilitate learning of complex compositional concepts. This facilitation may happen through reusing of learned representations of parts that were constructed before in process of learning of other concepts. As a result, sample complexity of learning compositional concepts may be greatly reduced.

Finally, hierarchical architectures have better tolerance to clutter. Clutter problem arises when the target object is in front of a non-uniform background, which functions as a distractor for the visual task. Hierarchical architecture provides signatures for parts of target objects, which do not include parts of background and are not affected by background variations. [6]

In hierarchical architectures, one layer is not necessarily invariant to all transformations that are handled by the hierarchy as a whole. Some transformations may pass through that layer to upper layers, as in the case of non-group transformations described in the previous section. For other transformations, an element of the layer may produce invariant representations only within small range of transformations. For instance, elements of the lower layers in hierarchy have small visual field and thus can handle only a small range of translation. For such transformations, the layer should provide covariant rather than invariant, signatures. The property of covariance can be written as distr ■ (■ μ l (g I) , μ l (t) ■) = distr ■ (■ μ l (I) , μ l (g − 1 t) ■) {\displaystyle \operatorname {distr} (\langle \mu _{l}(gI),\mu _{l}(t)\rangle)=\operatorname {distr} (\langle \mu _{l}(I),\mu _{l}(g^{-1}t)\rangle)} , where l {\displaystyle l} is a layer, μ l (I) {\displaystyle \mu _{l}(I)} is the signature of image on that layer, and distr {\displaystyle \operatorname {distr} } stands for "distribution of values of the expression for all g ∈ G {\displaystyle g\in G} ".

Relation to biology

M-theory is based on a quantitative theory of the ventral stream of visual cortex. [7] [8] Understanding how visual cortex works in object recognition is still a challenging task for neuroscience. Humans and primates are able to memorize and recognize objects after seeing just couple of examples unlike any state-of-the art machine vision systems that usually require a lot of data in order to recognize objects. Prior to the use of visual neuroscience in computer vision has been limited to early vision for deriving stereo algorithms (e.g., [9]) and to justify the use of DoG (derivative-of-Gaussian) filters and more recently of Gabor filters. [10] [11] No real attention has been given to biologically plausible features of higher complexity. While mainstream computer vision has always been inspired and challenged by human vision, it seems to have never advanced past the very first stages of processing in the simple cells in V1 and V2. Although some of the systems inspired – to various degrees – by neuroscience, have been tested on at least some natural images, neurobiological models of object recognition in cortex have not yet been extended to deal with real-world image databases. [12]

M-theory learning framework employs a novel hypothesis about the main computational function of the ventral stream: the representation of new objects/images in terms of a signature, which is invariant to transformations learned during visual experience. This allows recognition from very few labeled examples – in the limit, just one.

Neuroscience suggests that natural functionals for a neuron to compute is a high-dimensional dot product between an "image patch" and another image patch (called template)

which is stored in terms of synaptic weights (synapses per neuron). The standard computational model of a neuron is based on a dot product and a threshold. Another important feature of the visual cortex is that it consists of simple and complex cells. This idea was originally proposed by Hubel and Wiesel. [9] M-theory employs this idea. Simple cells compute dot products of an image and transformations of templates ■ I , g i t k ■ {\displaystyle \langle I,g_{i}t^{k}\rangle } for i = 1 , … , | G | {\displaystyle i=1,\ldots ,|G|} (| G | {\displaystyle |G|} is a number of simple cells). Complex cells are responsible for pooling and computing empirical histograms or statistical moments of it. The following formula for constructing histogram can be computed by neurons:

where σ {\displaystyle \sigma } is a smooth version of step function, Δ {\displaystyle \Delta } is the width of a histogram bin, and n {\displaystyle n} is the number of the bin.

Applications

Applications to computer vision

In [clarification needed] [13] [14] authors applied M-theory to unconstrained face recognition in natural photographs. Unlike the DAR (detection, alignment, and recognition) method, which handles clutter by detecting objects and cropping closely around them so that very little background remains, this approach accomplishes detection and alignment implicitly by storing transformations of training images (templates) rather than explicitly detecting and aligning or cropping faces at test time. This system is built according to the principles of a recent theory of invariance in hierarchical networks and can evade the clutter problem generally problematic for feedforward systems.

The resulting end-to-end system achieves a drastic improvement in the state of the art on this end-to-end task, reaching the same level of performance as the best systems operating on aligned, closely cropped images (no outside training data). It also performs well on two newer datasets, similar to LFW, but more difficult: significantly jittered (misaligned) version of LFW and SUFR-W (for example, the model's accuracy in the LFW "unaligned & no outside data used" category is 87.55±1.41% compared to state-of-the-art APEM (adaptive probabilistic elastic matching): 81.70±1.78%).

The theory was also applied to a range of recognition tasks: from invariant single object recognition in clutter to multiclass categorization problems on publicly available data sets (CalTech5, CalTech101, MIT-CBCL) and complex (street) scene understanding tasks that requires the recognition of both shape-based as well as texture-based objects (on StreetScenes data set). [12] The approach performs really well: It has the capability of learning from only a few training examples and was shown to outperform several more complex state-of-the-art systems constellation models, the hierarchical SVM-based face-detection system. A key element in the approach is a new set of scale and position-tolerant feature detectors, which are biologically plausible and agree quantitatively with the tuning properties of cells along the ventral stream of visual cortex. These features are adaptive to the training set, though we also show that a universal feature set, learned from a set of natural images unrelated to any categorization task, likewise achieves good performance.

Applications to speech recognition

This theory can also be extended for the speech recognition domain.

As an example, in [15] an extension of a theory for unsupervised learning of invariant visual representations to the auditory domain and empirically evaluated its validity for voiced speech sound classification was proposed. Authors empirically demonstrated that a single-layer, phone-level representation, extracted from base speech features, improves segment classification accuracy and decreases the number of training examples in comparison with standard spectral and cepstral features for an acoustic classification task on TIMIT dataset. [16]

References