Visual semantic processing in primate frontotemporal cognitive network and machines

Morteza Mooziri[1,4], Meysam Zare[1,4], Milad Qolami[1], Mohammad Javan[1,2], Farideh Shakerian[1]*, Mohammad-Reza A. Dehaqani[3,5]*

1. Department of Brain and Cognitive Sciences, Cell Science Research Center, Royan Institute for Stem Cell Biology and Technology, ACECR, Tehran, Iran.
2. Department of Physiology, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran.
3. Cognitive Systems Laboratory, Control and Intelligent Processing Center of Excellence (CIPCE), School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran.
4. Equal contribution.
5. Lead contact.
* Correspondence: farideh.shakerian@gmail.com, dehaqani@ut.ac.ir

Abstract

Vision is a major sensory system to primates. Visual processing includes extraction of physical features and abstract information. Unlike physical features, we have limited knowledge on how the primate brain extracts semantic information of the visual input. Here, we recorded the neuronal activities of inferior temporal (ITC) and ventrolateral prefrontal cortices (PFC), while macaque monkeys viewed a series of natural and artificial visual stimuli. We found that mid-level semantic information, e.g., face vs. body distinction, is processed by ITC population and is sent to downstream regions, including PFC, in the feed-forward pathway. Contrarily, high-level abstractions, e.g., animate vs. inanimate, are solved by PFC and fed back to ITC. Furthermore, we show progressive abstractions along the feed-forward direction in state-of-the-art deep learning models of vision. These results provide insights on critical questions in the field, including the purpose of information processing by ITC and perceptual sequence of object features.

Introduction

Primates possess a remarkable ability to recognize and categorize objects rapidly and accurately, a function essential for survival, navigation through complex environments, and social interactions [1-3]. Psychophysical studies have long established that object recognition is not a singular, instantaneous process, but rather unfolds across multiple stages [4,5]. Early in this sequence, the brain extracts basic visual features—such as shapes, textures, and basic configurations—that allow for rapid, coarse distinctions between objects [3,4,6]. However, it is

the ability to categorize objects at more abstract, conceptual levels, that makes the uniqueness of this system. These high-level perceptual judgments, often delayed relative to recognition of basic visual features, point toward a hierarchical and interactive system of visual processing, where distinct neural circuits contribute to progressively refine object representations [3,4].

Behaviorally, it is shown that the brain can swiftly extract mid-level semantic features, such as distinguishing faces from bodies, while the formation of abstract categories, like animate versus inanimate, typically requires more time and likely engages additional cognitive resources [4,5]. It is not clear whether this basic level precession is critical for perception of higher-level abstractions or it is just simply indicative of a temporal ordering due to evolutionary importance or experience [4]. At the neural level, this hierarchy is thought to involve the inferior temporal (ITC) and the prefrontal (PFC) cortices, two regions critical for object representation and decision-making [2,7-11]. The ITC has long been associated with the encoding of physical features, which are necessary for differentiating visually similar objects [12-15], whereas the PFC has been implicated in abstract, context-dependent categorizations [10,16]. However, we are still lacking theories that provide a holistic understanding of how semantic information is derived from visual input.

Despite advances in understanding the roles of the ITC-PFC circuit in object recognition, key questions remain unresolved. A major one is “how these regions interact to support semantic processing?” Early perception of mid-level semantic information is believed to stem from basic level advantage of ITC [7]. In this view, mid-level semantic information is solved by ITC, which is probably fed forward to higher order cognitive areas like PFC [7]. In contrast, an alternative hypothesis proposes that ITC conveys primarily visual, non-semantic information to downstream regions, which are responsible for making abstraction by incorporating these physical features and probably benefiting prior knowledge [9,10,12-19]. Overall, although current models suggest a bidirectional flow of information between ITC and PFC [20,21], the content of information in each of these directions is a matter of debate. To our knowledge, there is no experimental evidence on how abstract representations emerge in this circuit.

Here, we attempted to address how abstract semantic information is represented in the primate brain. To that aim, we recorded the neuronal activities of ITC and PFC of macaques, while viewing a series of visual objects. We found that, mid-level semantic information is represented earlier in both ITC and PFC, compared to high-level abstractions. Interestingly, we show that this information is solved by ITC population and is transferred to PFC in the feed-forward direction. However, the high-level categorization is performed by PFC, which is then fed back to ITC. In line with these findings, we also observed progressive abstraction along the feed-forward direction in the state-of-the-art deep learning models of vision. These findings will provide insights on several existing questions of primate visual system, from debates on ITC processing to perceptual sequence of object features (see Discussion).

Results

Neurophysiology and behavior

We simultaneously recorded the neuronal activity of ITC and PFC while macaque monkeys viewed a series of visual stimuli containing isolated grayscale natural and artificial objects of several categories (Fig. 1A). Since spatial frequency profile, as a basic visual feature, can affect category processing in ITC [22], we matched it between different categories of our stimulus set (see Methods). After adjustment, there was no significant difference between categories in none of the spatial frequencies (all comparisons showed a $p > 0.05$; Supplementary Fig. 1). The stimulus set was designed to elicit both mid-level and high-level semantic distinctions, as it spanned several conceptual categories, including animate versus inanimate objects and faces versus bodies. This vastness of object categories is critical for object recognition tasks, especially in this case, where we wanted to cover a broad range of categories for each class of high-level abstractions. To achieve this goal, we leveraged an experimental setting that allows us to record the neuronal responses in thousands of trials in a technically feasible amount of time, namely rapid serial visual presentation (RSVP); in every trial, the stimulus appeared in the central 7^{o} of the animal's visual field for 80 ms, which was followed by 400 ms of blank screen (Fig. 1A). The animal received a juice reward in random periods ranging 2-4 secs for continuous fixation at the center of the screen.

Fig. 1B shows the hypothetical semantic hierarchy, similar to a previous study [7]. At the highest level of abstraction in this hierarchy, stimuli are broadly divided into two primary categories: animate and inanimate objects. Below this level, that is mid-level semantic, the stimuli can be either primate or non-primate and face or body for the category of animates and artificial or natural for inanimates. Importantly, since some of these classifications are related to human-level knowledge, and might not necessarily be perceived by macaques, we focused on animate vs. inanimate distinction for high-level abstraction and face vs. body distinction for mid-level abstraction, throughout the rest of this study. Using this clearly defined stimulus set, we aimed to disentangle the temporal dynamics of object representation in the ITC-PFC circuit, exploring how these regions process different levels of semantic and conceptual information.

Basic level advantage of ITC and PFC

Fig. 1 C-F presents sample units that show systematic selectivity for faces (Fig. 1C,E, upper panels), bodies (Fig. 1C,E, lower panels), animates (Fig. 1D,F, upper panels), and inanimates (Fig. 1D,F, lower panels) in the two regions (Fig. 1C,D for ITC and Fig. 1E,F for PFC). First, we sought to figure out the temporal dynamics of semantic representation in each of these regions. For that, we trained linear classifiers, specifically using linear discriminant analysis (LDA), to distinguish faces from bodies and animates from inanimate objects using the population neuronal activity of each region (Fig. 1G,I). We observed that mid-level semantic information was represented earlier than high-level information in ITC (mean ± SD in ms; onset times: mid-level = 76.59 ± 8.09, high-level = 83.61 ± 13.99, $p < 1e\text{-}6$, Mann-Whitney test; peak times: mid-level = 129.33 ± 14.03, high-level = 176.20 ± 17.17, $p < 1e\text{-}59$; Mann-Whitney test; Fig. 1G,H), similar to our previous study [7]. Furthermore, this early representation of mid-level information was also observed in PFC population activity (mean ± SD in ms; onset times: mid-

level = 73.24 ± 11.27, high-level = 105.24 ± 18.74, p < 1e-49, Mann-Whitney test; peak times: mid-level = 148.77 ± 24.04, high-level = 153.13 ± 15.13, p = 0.002, Mann-Whitney test; Fig. 1I,J). These results are consistent with previous findings on early representation of mid-level information in ITC [7]. Additionally, when considering PFC as the main brain area that engages in psychophysical tasks, where the subject must provide perceptual responses [16], these observations provide the rationale for faster categorization of basic level information at the behavioral or perceptual level, in line with previous reports [4,5].

Contrary to ITC, in which the temporal difference of information appearance between different levels of semanticness was more prominent in peak times (onset distributions d' = 0.61, peak distributions d' = 2.99; Fig. 1G,H), PFC, however showed a more pronounced separation in onset times (onset distributions d' = 2.07, peak distributions d' = 0.22; Fig. 1I,J). This striking difference in temporal dynamics of category representation between ITC and PFC prompted us to reconsider the hypothesis that "all this hierarchy is solved in one region." Hypothetically, if both levels of information are solved in one area and then transferred to the other, a relatively preserved temporal dynamics is expected in the two areas. Furthermore, perceptual information is not necessarily reflected in onset or peak times alone [4]. In this view, it is the speed of accumulation of perceptual evidence over time that defines the time-course of information representation, which is clearly defined in exemplar-based random walk model (Fig. 2A; see ref.[4] for a thorough discussion on this). To probe that issue, we used a metric aiming to measure the speed of information solution, not only representation, for a given problem in a brain area, namely solution time, which is defined as the difference of onset and peak times of information representation (Fig. 2B). Since it does not directly depend on any point estimate of representational times, it can capture the time required for a problem to be fully represented in a region, which can inversely show the rapidity of solution (Fig. 2B). Therefore, the combination of solution time with either one of onset or peak times can reliably point to resolution of information. With this view, we reassessed the time-course of population information in ITC and PFC for both levels semanticness (Fig. 1G,I). We found that ITC population activity solved the face vs. body separation at a shorter time than it did for animate vs. inanimate problem (mean ± SD in ms; animate vs. inanimate = 92.60 ± 23.30, face vs. body = 52.75 ± 18.38, p < 1e-46, Mann-Whitney test; Fig. 2C, upper panel), which aligns well with previous evidence [7]. Besides, this is in line with a major theory of ITC processing, which states that ITC makes the semantic perception of a face in macaques [23-25] and humans [25] (see Discussion). On the other hand, this relation was reversed in PFC population activity, with high-level abstractions being solved quicker than mid-level (mean ± SD in ms; animate vs. inanimate = 48.89 ± 24.96, face vs. body = 75.54 ± 25.91, p < 1e-21, Mann-Whitney test; Fig. 2C, lower panel). This observation can fit in the vast body of literature depicting the roles of PFC for high-level categorizations and perceptions [9,10,16,19,26] (see Discussion). This notion also explains that since mid-level information is received in in the feed-forward direction, they appear earlier in the population activity of PFC (i.e., earlier onset times), compared to high-level information that does not exist in the feed-forward information (Fig. 1I,J), while PFC does not solve this information (Fig. 2C). Therefore, earlier onset times does not necessarily mean the solution of a problem by a specific brain area.

Processing of mid-level information by ITC and high-level information by PFC

Hypothetically, faster solution times can be interpreted as the responsibility of a brain region in solving specific categorization tasks, especially when accompanied with an earlier representation in one of critical times, i.e., onset or peak time. Therefore, next, we compared the time-resolved dynamics of each of these levels of hierarchy in the ITC-PFC circuit. Supplementary Fig. 2A,E demonstrates the temporal evolution of mid- (Supplementary Fig. 2A) and high-level (Supplementary Fig. 2E) information in these regions. Specifically, face vs. body resolution happened earlier in ITC, compared to PFC; with a small negligible inverse effect in onset times (mean ± SD in ms; ITC = 76.59 ± 8.09, PFC = 73.24 ± 11.27, p = 0.001, Mann-Whitney test; Supplementary Fig. 2A,B), this difference is evident in peak (mean ± SD in ms; ITC = 129.33 ± 14.03, PFC = 148.77 ± 24.04, p < 1e-18, Mann-Whitney test; Supplementary Fig. 2A,C) and solution (mean ± SD in ms; ITC = 52.75 ± 15.38, PFC = 75.54 ± 25.91, p < 1e-20, Mann-Whitney test; Supplementary Fig. 2A,D) times. Animacy information show a relatively different dynamics in this circuit. Particularly, onset of separation happens earlier in ITC population compared to PFC (mean ± SD in ms; ITC = 83.61 ± 13.99, PFC = 105.24 ± 18.74, p < 1e-28, Mann-Whitney test; Supplementary Fig. 2E,F), which contrasts our previous result. However, this information peaked earlier (mean ± SD in ms; ITC = 176.20 ± 17.17, PFC = 153.13 ± 15.13, p < 1e-30, Mann-Whitney test; Supplementary Fig. 2E,G) and was solved quicker (mean ± SD in ms; ITC = 92.60 ± 23.30, PFC = 47.89 ± 24.96, p < 1e-43, Mann-Whitney test; Supplementary Fig. 2E,H) in PFC compared to ITC. Although these results are compatible with the theories of accumulation of perceptual evidence over time [4] (Fig. 2A), we still looked for a deeper mechanistic explanation for earlier onset of animacy separation in ITC. By visual inspection, objects in the two categories of animacy level look different in terms of physical features. Therefore, we quantified four low-level physical features of each object, namely circularity, elongation, spikiness, and contrast (see Methods). Supplementary Fig. 3A shows the 20 most circular objects, which contains a greater number of inanimates; whereas the 20 most spiky objects were predominantly animate (Supplementary Fig. 3B). Thus, we hypothesized that the earlier onset of animacy information in ITC, could potentially reflect the initiation of visual feature, and not semantic, processing. A linear classifier trained on merely these features could categorize objects as animate or inanimate (Mean ± SD of performance = 63.64 ± 4.09, one-sample Wilcoxon signed rank test from a theoretical chance level of %50 showed a p = eps, z-scored value of %50 compared to the performance distribution was -3.33; Supplementary Fig. 3C). Therefore, despite an earlier onset of animacy distinction in ITC compared to PFC, this phase of separation is more probably due to processing of physical features. Importantly, slower slope of information accumulation in ITC compared to PFC, as evidenced by longer solution time (Supplementary Fig. 2E,H), indicates that unlike PFC, ITC cannot fully solve this problem by itself and requires aid from a collaborator.

Next, we aimed to use representational similarity analysis (RSA) in parallel to the results shown in Supplementary Fig. 2 (and another purpose that will be described later). For that, we first created the ground truth representational similarity matrices (RSMs) for face vs. body (Fig. 2D) and animate vs. inanimate (Fig. 2H) categorizations. Next, instantaneous RSMs were derived from ITC and PFC population activities, and their similarities were computed to the related ground truth RSMs (Fig. 2D,H; see Methods). Similar to the results of the linear

classifier presented above (Supplementary Fig. 2), despite insignificant onset times (median ± SD in ms; ITC = 90.0 ± 51.48, PFC = 87.0 ± 45.41, p = 0.36, Mann-Whitney test; Fig. 2D,E), face-body separation peaked earlier (median ± SD in ms; ITC = 148.0 ± 22.06, PFC = 167.0 ± 7.72, p < 1e-7, Mann-Whitney test; Fig. 2D,F) and was solved faster (median ± SD in ms; ITC = 66.5 ± 55.25, PFC = 79.5 ± 45.40, p = 0.003, Mann-Whitney test; Fig. 2D,G) in the ITC. Also, animacy information initiated earlier (median ± SD in ms; ITC = 114.0 ± 41.47, PFC = 138.0 ± 69.25, p < 1e-9, Mann-Whitney test; Fig. 2H,I), but peaked later (median ± SD in ms; ITC = 205.0 $\pm$ 16.08, PFC = 177.0 $\pm$ 6.45, p < 1e-41, Mann-Whitney test; Fig. 2H,J) and was solved more slowly (median ± SD in ms; ITC = 90.0 ± 41.81, PFC = 38.0 ± 68.95, p < 1e-20, Mann-Whitney test; Fig. 2H,K) in ITC. As expected, these results resemble those observed with the earlier approach (Supplementary Fig. 2), and could be interpreted the same.

Confirmation with encoding models

To this point, we used the ITC and PFC population activities to predict the information of interest. Next, we asked whether the temporal evolution of neural response could be explained by statistical regularities of objects. For that, we used generalized linear models (GLMs) to predict the instantaneous neural response using a set of semantic properties and basic physical features of each object (see the full list of regressors and implementation details in Methods). Fig. 3A,E,I exhibits the time-course of regression coefficients, β values, for the three most important semantic regressors, with respect to the present study purpose, in both regions, namely β_{face} (Fig. 3A), β_{body} (Fig. 3E), and $\beta_{animacy}$ (Fig. 3I), which were binary categorical variables; each of these regressors were either True or False for every individual object. Units with unreliable encoding dynamics were excluded from the analyses (see Methods). Hypothetically, based on the above results (see Supplementary Fig. 2 and Fig. 2), we expected that the temporal evolution of face and body coefficients to be similar to each other and opposite of animacy. Our results showed that β_{face} initiated slightly earlier in PFC (median ± SD in ms; ITC = 85.5 ± 4.07, PFC = 74.0 ± 13.27, n = 95 and 78 units, respectively, p < 1e-38, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3A,B), but peaked earlier (median ± SD in ms; ITC = 149.0 ± 7.47, PFC = 173.0 ± 18.52, n = 95 and 78 units, respectively, p < 1e-26, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3A,C) and was solved faster (median ± SD in ms; ITC = 61.0 ± 7.71, PFC = 95.0 ± 22.69, n = 95 and 78 units, respectively, p < 1e-39, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3A,D) in ITC. Despite insignificant onset times (median ± SD in ms; ITC = 83.0 ± 8.81, PFC = 82.0 ± 9.77, n = 77 and 68 units, respectively, p = 0.33, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3E,F), β_{body} also peaked earlier (median ± SD in ms; ITC = 149.0 ± 5.79, PFC = 170.0 ± 9.49, n = 77 and 68 units, respectively, p < 1e-55, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3E,G) and was solved quicker (median ± SD in ms; ITC = 67.0 ± 11.06, PFC = 86.0 ± 13.95, n = 77 and 68 units, respectively, p < 1e-38, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3E,H) in ITC, similar to β_{face}. As expected, the temporal dynamics of $\beta_{animacy}$ differed from that of β_{face} and β_{body}.

Specifically, it showed earlier onset (median ± SD in ms; ITC = 86.0 ± 6.18, PFC = 78.0 ± 6.52, n = 82 and 69 units, respectively, p < 1e-20, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3I,J) and peak times (median ± SD in ms; ITC = 160.0 ± 5.52, PFC = 122.0 ± 27.7, n = 82 and 69 units, respectively, p = 0.01, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3I,K) as well as shorter solution time (median ± SD in ms; ITC = 76.0 ± 8.06, PFC = 48.0 ± 28.49, n = 82 and 69 units, respectively, p = 0.01, Mann-Whitney test for 200 times subsampling with replacement from each distribution; Fig. 3I,L) in PFC. Earlier in this manuscript, we argued that the sooner initiation of animacy categorization in ITC compared to PFC (Supplementary Fig. 2E,F and Fig. 2H,I) is likely to occur due to physical feature processing phase within ITC population (Supplementary Fig. 3). Here, we ruled-out the possible confounding effects due to several low-level visual features; spatial frequency profile and color were matched when preparing the stimuli, and circularity, elongation, spikiness, and contrast are accounted for analytically. Needless to say, other low-level as well as mid- and high-level visual features are yet untouched, which suggests cautious interpretation of the results. However, overall, we provide near-confirmatory evidence that exclusive evaluation of semantic features further emphasizes the roles of ITC for processing mid-level rapidly-categorized semantic information and PFC for high-level perceptually-advanced abstractions in primate brain (see Discussion).

Feed-forward transfer of mid-level and feedback transfer of high-level semantic information

The evidence up to now, is highly suggestive for directionality of information. Therefore, next we aimed to implement an approach similar to Granger causality, which measures how much a time-series is predictable from the past of another [27,28]. In short, we tried to predict a region's RSMs from earlier RSMs of the same or both regions over time. Specifically, at each timepoint t, we fitted regression models to predict the RSM_t from past RSMs of the same area or both areas; the residuals of these regression models were used similar to Granger causality [28]; a reduction in the residuals when using RSMs of both areas, suggests that the present population activity of this area (RSM_t) is predictable from the past of the other (see Methods). Fig. 4A,C illustrates the time-resolved dynamics of information transfer for face-body (Fig. 4A) and animate-inanimate (Fig. 4C) recognition in the ITC-to-PFC and PFC-to-ITC directions. We observed that mid-level information transmission tends to initiate sooner in the ITC-to-PFC pathway (median ± SD of temporal difference (PFC-to-ITC – ITC-to-PFC) of information transfer initiation in ms = 3.0 ± 34.08, p = 0.09, one-sample Wilcoxon signed rank test from a theoretical null value of 0 ms delay; Fig. 4A,B), although it did not reach the significant threshold. Interestingly, the opposite was the case for animacy. The animate-inanimate population information transmission started earlier in the PFC-to-ITC direction (median ± SD of temporal difference (PFC-to-ITC – ITC-to-PFC) of information transfer initiation in ms = 6 $\pm$ 30.61, p < 1e-4, one-sample Wilcoxon signed rank test from a theoretical null value of 0 ms delay; Fig. 4C,D). These findings are in line with the above results (Supplementary Fig. 2 and Fig. 2-3) and show that mid-level semantic information is transferred in the feed-forward direction, while high-level abstractions are feedbacks of PFC to sensory cortex.

Progressive abstraction along the feed-forward direction in deep learning models of vision

Thus far, we showed that as the visual input goes forward in the primate brain, more abstract information is derived from it. Subsequently, we asked "is this phenomenon only a feature of biological vision?" In other words, are deeper layers of artificial visual systems also more sensitive to more abstract information? To answer that, we tried to explore how semantic information is represented in state-of-the-art deep learning models of vision. We studied the models pretrained for object recognition task on various image datasets [29-37] (See Methods). Specifically, we studied CORnet-S, as the representative for CORnet family of networks [29], AlexNet [30], SqueezeNet [31], ResNet [32], DenseNet [33], Inceptionv3 [34], EfficientNet [35], VGG-16 [36], and MobileNet [37]. In each network, we extracted the activations of the layers of interest for our stimulus set (Fig. 1B), through the forward propagation, from which the layer-wise RSMs were constructed and were compared to the ground truth expectations (see Methods). We observed that, generally, as the input image reaches the deeper layers of the network, stronger abstract representations appear (all significant values had a $p < 0.05$; Fig. 5A,B). Importantly, this effect was true for both levels of abstractions, i.e., mid- and high-level semanticness. Also, VGG-16 showed the highest representational similarity to ground truth for both levels (mean ± SD of similarity; face-body = 0.62 ± 0.02, animate-inanimate = 0.21 ± 0.03; Fig. 5C,D). These results imply that, as we observed in primate brain, more advanced layers of the networks are relatively specialized for more semantic/abstract feature extractions.

Unlike many other networks of object recognition, CORnet family models are designed based on the primate visual system, with modules corresponding to primate cortical areas, including V1, V2, V4, and ITC [29]. In fact, for the above results (Fig. 5A,B), we used the activations of the output layer of these four modules in the CORnet-S. CORnet-S was chosen since it has the highest behavioral and ITC neural predictivity in the family [29]. Next, we were curious to see how much brain-like are these representations. Specifically, we asked "are these representations similar to those formed by the brain for each recognition task?" To that aim, we used the CORnet-S-extracted representation from the layer corresponding to ITC output, where the semantic information culminates, as the new ground truth, for each case of face vs. body and animate vs. inanimate distinction (see Methods). Supplementary Fig. 4A,C depicts the similarity of CORnet-S ITC layer population representation with the ITC- and PFC-derived representations for both cases over time. We found that CORnet-S ITC layer information was more similar to primate brain's ITC for both face-body (median ± SD of similarity; ITC = 0.44 ± 0.02, PFC = 0.36 ± 0.03, $p < 1e\text{-}61$, Mann-Whitney test; Supplementary Fig. 4B) and animate-inanimate (median ± SD of similarity; ITC = 0.18 ± 0.02, PFC = 0.15 ± 0.02, $p < 1e\text{-}26$, Mann-Whitney test; Supplementary Fig. 4D) separations. Subsequently, we tried to expand the same idea to the other networks; here, we also added CORnet-Z, CORnet-RT, and VGG-19. We performed the same comparison (as in Supplementary Fig. 4) while every time using the latest layer of each of these networks (as depicted in Fig. 5A,B) as the ground truth matrix. Fig. 5E,G demonstrates the temporal evolution for the similarity of ITC and PFC mid- (Fig. 5E) and high-level (Fig. 5G) semantic representations to those of deep models. We found that the primate ITC-derived representations were more similar to these networks, compared to PFC representations, for both face-body (mean ± SD of similarity; NN~ITC = 0.46 ± 0.05,

NN~PFC = 0.35 ± 0.04, n = 12 networks, p = 0.0005, one-sample Wilcoxon signed rank test from a theoretical null value of 0 for similarity difference; Fig. 5F) and animate-inanimate (mean ± SD of similarity; NN~ITC = 0.21 ± 0.04, NN~PFC = 0.16 ± 0.02, n = 12 networks, p = 0.0005, one-sample Wilcoxon signed rank test from a theoretical null value of 0 for similarity difference; Fig. 5H) cases. Also, VGG-16 showed highest similarity to ITC population activity (mean ± SD of similarity = 0.52 ± 0.02; Fig. 5I, left panel), while AlextNet was the most similar network to PFC representation (mean ± SD of similarity = 0.42 ± 0.03; Fig. 5I, right panel) for face-body discrimination. For animacy separation, AlexNet was the most similar model to ITC (mean ± SD of similarity = 0.30 ± 0.02; Fig. 5J, left panel) and MobileNet formed the most PFC-like (mean ± SD of similarity = 0.20 ± 0.03; Fig. 5J, right panel) representation. Overall, we can infer that in line with previous reports [20,21], the visual function of PFC, which is subject to growing interest [38], is most probably absent, at least partially, in currently available deep models of vision. Also, while VGG-16 has more visual cortex-like properties, which is similar to a previous study [39], AlexNet and MobileNet produce more abstract and PFC-like and representations.

Discussion

Here, we showed that the primate brain solves mid-level semantic information in higher visual cortex during feed-forward pathway, while the more abstract information remains to be extracted by more cognitive areas of neocortex, which then feedback a copy of that information to upstream regions. Our work addresses critical gaps in the literature by linking behavioral evidence of hierarchical visual processing to the neural circuits that mediate these functions. While previous studies have focused on either the ITC or PFC in isolation, our investigation of their interaction over time will provide a more comprehensive view of how the primate brain integrates sensory information and abstract knowledge to achieve robust, flexible object recognition. We also show that this progressive abstraction regime can be generalized to artificial models of vision.

The purpose of information processing by ITC is strongly controversial [12-15,23-25]. While the vast body of literature in humans [25] and monkeys [23-25] proposes that ITC makes the semantic perception of certain evolutionarily important attributes, such as being a face or a body, there is an alternative theory suggesting that ITC processes physical, and not semantic, aspects of the visual input [12-15]; the latter, suggests that these visual properties are combined by more cognitively developed areas, like PFC, to form abstract perceptions [8,16]. The observation that mid-level information is represented earlier in ITC and travels in the forward direction to PFC, is consistent with the former theory. Furthermore, there is evidence that forward propagation is not enough for solving object recognition [21]. On the other hand, PFC has established roles in more cognitively advanced behaviors, which includes abstractions [9,10,16,26]. Also, PFC sends feedback signals to ITC which helps most for processing late-stage difficult-to-recognize objects [20]. Here, we show that PFC precedes ITC for representing high-level semantic information, and feedbacks this representation to visual cortex, which fits quite well in the

mentioned literature. Therefore, we suggest that the dynamic interactions between ITC and PFC are required for a complete perception of visual input.

From another perspective, our results explain the behavioral observations on perceptual sequence of object features [4,5]. Object recognition is not a one-stage process, but rather a sequential one [4,5]. Certain features of an object are recognized earlier, while others take longer to be perceived [4,5]; the more abstract, such as animate or not, and more memory-dependent, like identity, attributes are typically processed later, while salient characteristics are detected rapidly [4,5]. Parallel to behavior, the same sequence is represented in ITC population activity [7]. While the early representation of mid-level information in ITC is a wide belief [7,23-25] and solves part of the behavioral hierarchy [7], the late perception of more abstract information remained unresolved. Considering the PFC as the area responsible for this degree of abstraction, fills the mentioned gap. It is reasonable to assume that this delayed perception could be due to the time required for information to reach PFC and the internal cognitive processing mechanisms within PFC to solve the problem.

Is this sequential progressive abstraction only a property of biological vision? Computer vision systems are far less capable than primates for categorization and image recognition tasks [40-42]; besides the etiological bases of these phenomena, we are also lacking methodological knowledge to improve their performance. One reasonable approach to construct efficient networks for such problems is to simulate biological vision, for which primate visual system is a remarkable candidate [29]. But a major problem in this case would be our very limited knowledge of the primate vision itself. Object recognition has long been attributed to the ventral visual stream [2,11], while a number of studies point to substantial roles of PFC as a major contributor to these processes [8,19,20]. Interestingly in this case, the roles of PFC become crucial for more difficult categorization problems [20], which could most probably overlap with the situations in which current deep models of vision fail. With the data presented here, we suggest that since obviously PFC activity is crucial for object recognition, semantic processing, and abstract categorizations in primates [8,19,20], considering a stage for accomplishing PFC duties will probably improve the performance of computer vision systems.

In conclusion, we provide mechanistic insights for why and how the collaboration of visual and prefrontal cortices is required to form robust semantic perceptions of the visual input in primates. From a general perspective, these results lay the foundations for a more networked view of visual processing, contrary to the traditional modular processing theories. Also, we suggest approaches to improve the artificial visual systems in object recognitions tasks.

Methods

Animals and surgery

Two male rhesus macaque monkeys (monkey F/V; Macaca mulatta, weight: 9.4/8.8 kg, age: 10/9 years old) entered the study. Experiments were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and were

approved by the internal ethics committee at Royan Institute. In the beginning, monkeys underwent MRI imaging to help design recording chambers and head posts, which were subsequently implanted through surgery. The head post was located midline and monkey F/V had one recording chamber on left/right hemisphere. In a second surgery, the craniotomy was performed over the area covering both ITC and PFC. Finally, a CT scan was acquired to help correctly localize the regions of interest, in combination with the MRI.

Stimuli, task, and behavior

Visual stimuli were isolated natural and artificial objects in grayscale and were shown on gray background (Fig. 1A). The stimulus set comprised of several basic categories required to capture the diversity of real-world objects underneath the most abstract level, that is animacy (Fig. 1B). Specifically, it contained faces and bodies of humans and monkeys, four-limb animals, reptiles, fishes, birds, and insects for the animate category and flowers, fruits, chairs, cars, houses, clocks, and tools as inanimates. Spatial frequency profile was matched among different categories, using SHINE toolbox [43].

Monkeys were trained to perform a fixation task, to receive juice reward following periods of continuous fixation. As factors like prior experience to categorizations affect the timing of perceptual information [4,17], subjects were kept naïve to any categorization or identification tasks. This is ideal for the present study purpose, since it helps purely capture the basic representational sequences within the brain. All experiments, including training, were performed in one experimental rig, and the task was run in PsychToolbox v3.0.18. Stimuli were presented in the central 7° of the animal's visual field on a BenQ monitor with a resolution of 1920 × 1080 and a refresh rate of 144 Hz. Monkeys were positioned 50 cm distant from the center of the monitor. Simultaneously, the eye position was tracked using an infrared eye-tracking device (Zist Kankash Toos, Mashhad, Iran) with a sampling frequency of 200 Hz. Each visual stimulus was shown 5-10 times, in different recording sessions (same number of repetitions for all stimuli in every recording session).

Visual feature extraction

We used OpevCV v4.10.0 library in Python to extract physical features of objects. After preprocessing, *contourArea*, *arcLength*, and *boundingRect* functions were used to extract surface area (A_{obj}), perimeter (P_{obj}), and bounding rectangle (i.e., the smallest upright rectangle that fully encloses the object) of non-background pixels of each object's contour, respectively. Subsequently, the circularity and elongation were calculated as the following:

$$Circularity = \frac{4\pi \times A_{obj}}{P_{obj}^2}$$

$$Elongation = 1 - \frac{d_{min}}{d_{max}}$$

where d_{min} and d_{max} are the shorter and longer dimensions the object's bounding rectangle, respectively. Spikiness was computed using the A_{obj} and the area of the object's convex hull (A_{conv}; computed by OpenCV *convexHull* function), as the following:

$$Spikiness = 1 - \frac{A_{obj}}{A_{conv}}$$

Contrast was defined as the standard deviation of the object's pixel values in grayscale.

Electrophysiological recording

In each session, head-fixed animal sat in the monkey chair and viewed the visual stimuli at the center of the screen. Neural recordings were performed through grids uniquely designed for each subject's chamber with 1.5/1 mm spacing between centers of the neighboring holes for monkey F/V. Tungsten electrodes (FHC, 130 mm length; Bowdoin, ME, USA) and the covering stainless steel guide tubes were mounted on a Motorized Electrode Manipulator (MEM)™ (Thomas Recording; Gießen, Hessen, Germany) and were lowered to cross the dura, at AP and ML coordinates related to ITC and ventrolateral PFC. After passing the dura, the electrodes were cautiously inserted into the brain using the mentioned micro-driver. Neural data was recorded using a recording device (Blackrock Neurotech; Salt Lake City, UT, USA) in a sampling rate of 30 kHz. A total of 88/68 recording sessions were performed from monkey F/V. Most of the sessions were dually recorded, from both ITC and PFC, while a few sessions contained the neural response of one region. Thus, we had 78/59 ITC neural sites and 57/63 PFC neural sites for monkey F/V. Data from online-detected neural sites with auto-thresholding were used for subsequent analyses.

Neural data analysis

Offline data analyses were performed in MATLAB 2022b and Python v3.11.7. Neuronal responses were time-locked to the stimulus presentation onset. For all analyses, each unit's response was z-scored to 80 ms time window prior to stimulus onset, which was subsequently smoothed with averaging in consecutive 20 ms-long windows (with step size of 1 ms) to form the final peri-stimulus time histogram (PSTH). Responses were averaged for all trials of the same stimulus. All onset times were considered as the first moment of time that the time-course of response passed the following threshold:

$$Value_{threshold} = \text{baseline average} + 3 \times \text{baseline std}$$

where baseline was [-50,50] ms relative to stimulus presentation onset. Repetitions (for population data analyses) and units (for encoding models) with either onset or peak time outside the window of [50,300] ms relative to stimulus presentation onset were considered unreliable and excluded from subsequent analysis. Solution time was calculated as the time difference between onset and peak times.

Classification

All classifications were performed using LDA method (MATLAB *fitcdiscr.m* and *predict.m*) on ITC and PFC population neural data for different levels of the semantic hierarchy (Fig. 1 G-J and Supplementary Fig. 2). Specifically, for every timepoint an LDA classifier was trained to either detect animates from inanimates (high-level abstraction) or faces from bodies (mid-level abstraction). In all cases, %70/%30 of the data were used to the train/test the model. After forming confusion matrices, the average of within-class accuracies was used as the representative accuracy. This procedure was repeated 200 times. Subsequently, critical times, i.e., onset and peak, and solution time were computed as described above. Classification of animate from inanimate objects using visual features (Supplementary Fig. 3C) was done in Python using scikit-learn *LinearDiscriminantAnalysis* function. In every run of 200 iterations, physical features of %70/%30 of the samples were used to train/test the model.

Representational similarity analysis

For RSA, first the ground truth RSMs of each case were created, which theoretically is 0 for no similarity and 1 when perfect similarity is expected. At every timepoint in each region, the cosine-similarity (using scikit-learn *cosine_similarity* function) was computed between the vectors of neural response to every possible pair of the stimuli, as the following:

$$\cos(\theta) = \frac{Resp_{stim_i} \cdot Resp_{stim_j}}{||Resp_{stim_i}||\,||Resp_{stim_j}||}$$

where $Resp_{stim_i}$ and $Resp_{stim_j}$ are the vectors of neural response to i_{th} and j_{th} stimuli, respectively, and θ is the angle between these two vectors in the high-dimensional neural space. The greater the $\cos(\theta)$, the more similar the two vectors are. Of note, only the neural responses to face and body objects were used to construct face-body RSMs. This process would create the instantaneous N×N regional RSMs for face-body and animate-inanimate conditions, which would have the following structure:

$$\begin{bmatrix} Similarity_{stim_{1,1}} & \cdots & Similarity_{stim_{1,N}} \\ \vdots & \ddots & \vdots \\ Similarity_{stim_{N,1}} & \cdots & Similarity_{stim_{N,N}} \end{bmatrix}$$

where N is total number of objects for both categories and $Similarity_{stim_{i,j}}$ is the cosine-similarity between vectors of neural response to $stim_i$ and $stim_j$. Next, the correlation between each data-derived RSM and the ground truth RSM was calculated with Kendall's tau correlation (using scipy *kendalltau* function). This procedure was repeated 200 times and, in every run, 20/50 stimuli per each class (a total of 40/100 objects) were randomly selected to form face-body/animate-inanimate RSMs. Subsequently, onset, peak, and solution times were computed as described above.

Granger causality

To quantify information transfer rate over time, we used the RSMs created during the RSA (see above), in a way similar to Granger causality [27,28]. First, the RSMs were created for each region separately over time; this would result in N×N×T matrices for each region, where N is the number of stimuli and T is number of timepoints. Next, we formed two regression models to predict the RSM at timepoint t (RSM_t) of the theoretical receiver area from the RSMs of timepoints [t-75,t-25] in the following manner: (1) the first regression model only had the past RSMs of the receiver area to predict the RSM_t, while (2) the second model had to predict the RSM_t from the past RSMs of both areas. Hypothetically, if the theoretical sender area transfers information to the theoretical receiver area the residual of the second model should be lower than that of the first model. To compare the residuals, we used the following formulation:

$$Predictivity = \log\frac{SSR_{model(2)}}{SSR_{model(1)}}$$

where $SSR_{model(1)}$ and $SSR_{model(2)}$ are the sum squared residuals for model (1) and model (2), respectively. PFC and ITC were the receiver areas for the feedforward and feedback directions, respectively. This procedure was repeated 200 times and, in every run, 20/50 stimuli per each class (a total of 40/100 objects) were randomly selected for face-body/animate-inanimate information transfer. Subsequently, the initiation of information transfer was calculated for every case and direction, as described above.

Generalized linear models

Encoding models, specifically GLMs, were formed to predict the neural response from a set of semantic and physical features of each object. Specifically, the semantic features were the following variables in binary format: animate, face, body, human, monkey; each regressor for each object was either True or False. Physical features were circularity, elongation, spikiness, and contrast of each object, as floating-point numbers. One model was formed for every unit at every timepoint (using Statsmodels *GLM* function). The full model was as the following:

$$y_t = \beta_0 + \beta_1 \times X_{animacy} + \beta_2 \times X_{face} + \beta_3 \times X_{body} + \beta_4 \times X_{human} + \beta_5 \times X_{monkey} + \beta_6 \times X_{circularity} + \beta_7 \times X_{elongation} + \beta_8 \times X_{spikiness} + \beta_9 \times X_{contrast}$$

y_t is the baseline z-scored, stimulus-averaged, and smoothed (moving average with 20-ms window and 1-ms step) neural response at timepoint t. To prevent the units with negative and positive coefficient values at each timepoint cancel-out each other, we used the absolute β values for subsequent analyses. The time-course of face, body, and animate coefficients of these models were extracted, and onset, peak, and solution times of each coefficient were computed as described above.

Deep learning models of vision

Neural network evaluations were performed in PyTorch v2.3.0 and Torchvision v0.18.0. Layers of interest for each network are listed in Table 1. For each network, first, the activations of

each layer of interest were extracted in the feedforward direction. Next, the layer-wise RSMs were created and compared to the related ground truth matrices similar to the procedure employed for neural data (Fig. 5A-D). To compute the brain-network similarities (Fig. 5E-J and Supplementary Fig. 4), the RSMs of the last layer of interest of each network was used instead of the ground truth matrices used in Fig. 2D,H. Both procedures were repeated 200 times and, in every run, 20/50 stimuli per each class (a total of 40/100 objects) were randomly selected for face-body/animate-inanimate condition. Peak similarity value of every repetition entered the statistical comparisons.

Table 1. Layers of interest for deep models of vision

Network	Layers
CORnet-S CORnet-RT CORnet-Z	"module.V1.output", "module.V2.output", "module.V4.output", "module.IT.output"
AlexNet	"features.1", "features.4", "features.7", "features.9", "features.12"
SqueezeNet	"features.0", "features.1", "features.2", "features.3.cat", "features.4.cat", "features.5", "features.6.cat", "features.7.cat", "features.8", "features.9.cat", "features.10.cat", "features.11.cat", "features.12.cat"
ResNet	"layer1.2.relu_2", "layer2.3.relu_2", "layer3.5.relu_2", "layer4.2.relu_2"
DenseNet	"features.transition1.pool", "features.transition2.pool", "features.transition3.pool", "flatten"
Inceptionv3	"maxpool1", "maxpool2", "Mixed_5d.cat", "Mixed_6e.cat", "avgpool"
EfficientNet	"features.1.1.add", "features.2.2.add", "features.3.2.add", "features.4.3.add", "features.5.3.add", "features.6.4.add", "features.7.1.add", "avgpool"
VGG-16	"features.4", "features.9", "features.16", "features.23", "features.30"
VGG-19	"features.4", "features.9", "features.18", "features.27", "features.36"
MobileNet	"features.0", "features.1.add", "features.2.block.2", "features.3.add", "features.4.block.3", "features.5.add", "features.6.add", "features.7.block.2", "features.8.add", "features.9.add", "features.10.add", "features.11.block.3", "features.12.add", "features.13.block.3", "features.14.add", "features.15.add", "features.16"

Statistical analyses

Statistical and machine learning analyses were performed in Python v3.11.7, using scikit-learn v1.2.2, Statsmodels v0.14.0, and SciPy v1.13.0 libraries, and MATLAB 2022b. Details of the statistical tests used for each comparison are described wherever appropriate throughout the text. All tests were two-tailed and p-values less than 0.05 were considered as statistically significant. d' for two data distributions with the means μ_1 and μ_2 and the SDs σ_1 and σ_2 was computed as the following:

$$d' = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$$

References

1 DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends in cognitive sciences* **11**, 333-341 (2007).

2 DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415-434 (2012).

3 Serre, T., Oliva, A. & Poggio, T. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences* **104**, 6424-6429 (2007).

4 Mack, M. L. & Palmeri, T. J. The timing of visual object categorization. *Frontiers in Psychology* **2**, 165 (2011).

5 Mack, M. L. & Palmeri, T. J. The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General* **144**, 551 (2015).

6 Riesenhuber, M. & Poggio, T. Models of object recognition. *Nature neuroscience* **3**, 1199-1204 (2000).

7 Dehaqani, M.-R. A. *et al.* Temporal dynamics of visual category representation in the macaque inferior temporal cortex. *Journal of neurophysiology* **116**, 587-601 (2016).

8 Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K. & Poggio, T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology* **100**, 1407-1419 (2008).

9 Rainer, G., Rao, S. C. & Miller, E. K. Prospective coding for objects in primate prefrontal cortex. *Journal of Neuroscience* **19**, 5493-5505 (1999).

10 Wallis, J. D., Anderson, K. C. & Miller, E. K. Single neurons in prefrontal cortex encode abstract rules. *Nature* **411**, 953-956 (2001).

11 Lehky, S. R. & Tanaka, K. Neural representation for object recognition in inferotemporal cortex. *Current opinion in neurobiology* **37**, 23-35 (2016).

12 Vinken, K., Prince, J. S., Konkle, T. & Livingstone, M. S. The neural code for "face cells" is not face-specific. *Science Advances* **9**, eadg1736 (2023).

13 Bardon, A., Xiao, W., Ponce, C. R., Livingstone, M. S. & Kreiman, G. Face neurons encode nonsemantic features. *Proceedings of the national academy of sciences* **119**, e2118705119 (2022).

14 Arcaro, M. J., Ponce, C. & Livingstone, M. The neurons that mistook a hat for a face. *Elife* **9**, e53798 (2020).

15 Baldassi, C. *et al.* Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS computational biology* **9**, e1003167 (2013).

16 Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annual review of neuroscience* **24**, 167-202 (2001).

17 Tanaka, J. W. & Taylor, M. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology* **23**, 457-482 (1991).

18 Johnson, K. E. & Mervis, C. B. Effects of varying levels of expertise on the basic level of categorization. *Journal of experimental psychology: General* **126**, 248 (1997).

19 Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312-316 (2001).

20 Kar, K. & DiCarlo, J. J. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron* **109**, 164-176. e165 (2021).

21 Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience* **22**, 974-983 (2019).
22 Toosi, R. *et al.* The Spatial Frequency Representation Predicts Category Coding in the Inferior Temporal Cortex. *bioRxiv*, 2023.2011. 2007.566068 (2023).
23 Shi, Y. *et al.* Rapid, concerted switching of the neural code in inferotemporal cortex. *bioRxiv*, 2023.2012. 2006.570341 (2023).
24 Landi, S. M., Viswanathan, P., Serene, S. & Freiwald, W. A. A fast link between face perception and memory in the temporal pole. *Science* **373**, 581-585 (2021).
25 Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126-1141 (2008).
26 Miller, E. K. The prefontral cortex and cognitive control. *Nature reviews neuroscience* **1**, 59-65 (2000).
27 Bernasconi, C. & KoÈnig, P. On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. *Biological cybernetics* **81**, 199-210 (1999).
28 Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424-438 (1969).
29 Kubilius, J. *et al.* Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in neural information processing systems* **32** (2019).
30 Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012).
31 Iandola, F. N. *et al.* SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and. *arXiv preprint arXiv:1602.07360* (2016).
32 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770-778.
33 Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4700-4708.
34 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2818-2826.
35 Tan, M. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019).
36 Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
37 Howard, A. G. *et al.* MobileNets: efficient convolutional neural networks for mobile vision applications (2017). *arXiv preprint arXiv:1704.04861* (2017).
38 Rose, O. & Ponce, C. R. A concentration of visual cortex-like neurons in prefrontal cortex. *Nature Communications* **15**, 7002 (2024).
39 Nonaka, S., Majima, K., Aoki, S. C. & Kamitani, Y. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience* **24** (2021).
40 Rajalingham, R. *et al.* Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience* **38**, 7255-7269 (2018).
41 Schrimpf, M. *et al.* Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007 (2018).
42 Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* **19**, 356-365 (2016).
43 Willenbockel, V. *et al.* Controlling low-level image properties: the SHINE toolbox. *Behavior research methods* **42**, 671-684 (2010).

Acknowledgements

The authors would like to thank Mohammad Rabiei for his technical assistance.

Author Contributions

MRAD & FS conceptualized the study. MM, MZ, & MQ collected the data. MM analyzed the data, performed neural network evaluations, performed visualizations, and wrote the manuscript. MJ & MRAD supervised the study.

Data Availability

Data will be made available upon reasonable request to the corresponding authors.

Code Availability

MATLAB scripts and functions as well as Python notebooks will be made publicly available at https://github.com/mooziri/Paper_VisualSemanticProcessing following publication of the study.

Competing Interests

None declared.

Funding

This work was funded by --- to MRAD (grant number: ---).

Figure Legends

Figure 1. Experimental design, theoretical framework, and regional information representation. (A) Schematic of the experiment; the animal was trained to watch the visual objects at the center of the screen, to receive juice rewards. In every trial, the stimulus appeared in the central 7^{o} of the animal's visual field for 80 ms, which was followed by 400 ms of blank screen. Simultaneously, the neuronal activities of ITC and vlPFC was recorded for offline analyses. Lower left panel shows a schematic of the recording locations. (B) Semantic hierarchy of visual stimuli used in this study, displaying the categorical distinctions: animate vs. inanimate represent high-level and face vs. body represent mid-level abstraction. (C-F) PSTHs of sample units with greater response for face (C,E, upper panels), body (C,E, lower panels), animate (D,F, upper panels), inanimate (D,F, lower panels) objects in ITC (C,D) and PFC (E,F). Solid lines and shaded areas indicate the mean values and SEM of the instantaneous firing rates, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus

presentation. (G,I) Time-course of decoding accuracy for classifiers trained and tested on neural data from ITC (G) and PFC (I) to distinguish face vs. body or animate vs. inanimate. Solid lines and shaded areas indicate the mean values and SD of the classifiers' accuracies, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines denoted with arrows and arrowheads are mean values of onset and peak times, respectively. (H,J) Statistical comparison of onset (upper panels) and peak (lower panels) times for the regional classifiers in G,I for ITC (H) and PFC (J). Bars and error bars indicate mean values and SD, respectively. Statistical significance measured by Mann-Whitney test. **p < 0.01, ***p < 0.001. ITC, inferior temporal cortex; PSTH, peri-stimulus time histogram; vlPFC, ventrolateral prefrontal cortex.

Figure 2. Temporal dynamics of semantic information in the ITC-PFC circuit. (A,B) Schematic illustrations of the theoretical frameworks describing random walk model (A) and solution time. (C) Statistical comparison of solution times between different levels of semantic hierarchy in ITC (upper panel) and PFC (lower panel). Bars and error bars indicate mean values and SD, respectively. Statistical significance measured by Mann-Whitney test. (D,H, bottom) Time-course of similarity to ground truth for the population activities of ITC and PFC for face vs. body (A) or animate vs. inanimate (G) distinctions. Solid lines and shaded areas indicate the mean values and SD of the similarities, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines denoted with arrows and arrowheads are median values of onset and peak times, respectively. (D,H, top) Peak similarity RSMs of ITC (leftmost panels) and PFC (rightmost panels) alongside the ground truth (middle panels) for face-body (A) and animate-inanimate (E) conditions. Warmer colors indicate greater values of cosine-similarity. (E-G,I-K) Statistical comparison of onset (E,I), peak (F,J), and solution (G,K) times for the similarities in D,H for face-body (E-F) and animate-inanimate (I-K) separations. Bars and error bars indicate median values and SEM, respectively. Statistical significance measured by Mann-Whitney test. **p < 0.01, ***p < 0.001. ITC, inferior temporal cortex; PFC, prefrontal cortex; RSM, representational similarity matrix.

Figure 3. Temporal dynamics of encoding models' semantic coefficients in the ITC-PFC circuit. (A,E,I) Time-course of encoding models' β_{face} (A), β_{body} (E), and $\beta_{animacy}$ (I) for neuronal activities of ITC and PFC. Solid lines and shaded areas indicate the mean values and SD of the coefficients, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines denoted with arrows and arrowheads are median values of onset and peak times, respectively. (B-D,F-H,J-L) Statistical comparison of onset (B,F,J), peak (C,G,K), and solution (D,H,L) times for the coefficients in A,E,I. Bars and error bars indicate median values and SD, respectively. Statistical significance measured by Mann-Whitney test for 200 times subsampling with replacement from each distribution. *p < 0.05, ***p < 0.001. ITC, inferior temporal cortex; PFC, prefrontal cortex.

Figure 4. Temporal dynamics of information transfer in the ITC-PFC circuit. (A,C) Time-course of Granger causality in the ITC-to-PFC and PFC-to-ITC directions for face vs. body (A) or animate vs. inanimate (C) distinctions. Solid lines and shaded areas indicate the mean values and SD of the predictivity, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines denoted with arrows are median values of information transfer initiation. (B,D) Histogram of temporal delay of information transmission for "PFC-to-ITC – ITC-to-PFC" in face-body (B) and animate-inanimate (D) conditions. Positive time difference values indicate greater raw values for PFC-to-ITC direction, which means delayed initiation of transfer in this direction. Dashed black lines indicate 0 ms time difference, while dashed red lines show the median value of each distribution. Statistical significance measured by one-sample Wilcoxon signed rank test from a theoretical null value of 0 ms delay. ITC, inferior temporal cortex; PFC, prefrontal cortex.

Figure 5. Semantic processing in deep models of vision. (A,B, top) From left, RSMs of ground truth expectations and last layer of the studied networks for mid- (A) and high-level (B) semanticness. Warmer colors indicate greater values of cosine-similarity. (A,B, bottom) Each panel depicts the similarity of all studied layers of a network to ground truth RSM, shown in the upper rows, for face-body (A) and animate-inanimate (B) conditions. Solid lines and shaded areas indicate the mean values and SD of similarity, respectively. (C,D) Comparison of all networks' last layer similarity to ground truth RSMs for face-body (C) and animate-inanimate (D) distinctions. Bars and error bars indicate mean values and SD, respectively. (E,G) Time-course of similarity of ITC and PFC population representations to the last layer of all studied networks for face-body (E) or animate-inanimate (G) separations. Solid lines and shaded areas indicate the mean values and SD of the similarities, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. (F,H) Scatter plot for the networks' peak similarities to ITC and PFC population representations shown in E,G. Each point is one network, and the bars along the x and y axes are the SD of the network similarity to PFC and ITC, respectively (n = 12 networks). Histograms on the top right corners illustrate the similarity difference between NN~ITC and NN~PFC raw values. Dashed red lines in the histograms are the median similarity difference in each case. Statistical significance measured by one-sample Wilcoxon signed rank test from a theoretical null value of 0 for similarity difference. (I,J) Comparison of all networks' last layer similarity peak to ITC (left panels) and PFC (right panels) population representations for face-body (I) and animate-inanimate (J) distinctions. Bars and error bars indicate mean values and SD, respectively. ITC, inferior temporal cortex; PFC, prefrontal cortex; RSM, representational similarity matrix.

Supplementary Figure 1. Matching spatial frequency among categories. Spatial frequency profile for the visual stimuli in each category. Solid lines and shaded areas indicate the mean values and SD of the amplitude, respectively. Statistical significance measured by Mann-Whitney test between category pairs at every frequency.

Supplementary Figure 2. Temporal evolution of semantic information classification in the ITC-PFC circuit. (A,E) Time-course of decoding accuracy for classifiers trained and tested on neural data from ITC and PFC to distinguish face vs. body (A) or animate vs. inanimate (E). Solid lines and shaded areas indicate the mean values and SD of the classifiers' accuracies, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. Dashed lines denoted with arrows and arrowheads are mean values of onset and peak times, respectively. (B-D,F-H) Statistical comparison of onset (B,F), peak (C,G), and solution (D,H) times for the classifiers in A,E for face vs. body (B-D) and animate vs. inanimate (F-H). Bars and error bars indicate mean values and SD, respectively. Statistical significance measured by Mann-Whitney test. **p < 0.01, ***p < 0.001. ITC, inferior temporal cortex; PFC, prefrontal cortex.

Supplementary Figure 3. Visual feature differences between high-level semantic categories. (A,B) Illustration of top 20 most circular (A) and spiky (B) objects in the entire stimulus-set. (C) Histogram of the classifier's accuracy to distinguish animate vs. inanimate using four basic physical features, i.e., circularity, spikiness, elongation, and contrast. Dashed black line indicates the chance level accuracy, i.e., %50, while dashed red line shows the median value of the performance distribution. Statistical significance measured by one-sample Wilcoxon signed rank test from a theoretical chance level of %50.

Supplementary Figure 4. Semantic similarity of CORnet-S ITC to primate ITC and PFC. (A,C) Time-course of similarity to CORnet-S ITC layer for the population activities of ITC and PFC for face-body (A) or animate-inanimate (C) distinctions. Solid lines and shaded areas indicate the mean values and SD of the similarities, respectively. Gray rectangle at 0-80 ms represents the time window of stimulus presentation. (B,D) Histogram of peak similarity values in A,C. Dashed lines show the median value of each distribution. Statistical significance measured by Mann-Whitney test. ITC, inferior temporal cortex; PFC, prefrontal cortex.