SemanticCMC: Contrastive Learning of Meaningful Object Associations from Temporal Co-occurrence Patterns in Naturalistic Movies

Cliona O'Doherty odoherc1@tcd.ie
Rhodri Cusack cusackrh@tcd.ie
Trinity College Institute of Neuroscience
Trinity College Dublin

Abstract

Deep convolutional neural networks continue to prevail at the forefront of innovation in computer vision. New, contrastive methods of self-supervised learning are lessening our reliance on curated datasets and hold potential for more robust and generalisable learning. While much of the focus in computer vision is on classification accuracy and if a network is performing well, it is equally useful to understand how these models are learning. By probing the representational underpinnings of DNN behaviour, we can better understand the foundations of success in computer vision. Here, we present an application of representational similarity analysis to investigate the patterns of activations within the self-supervised network Contrastive Multiview Coding (CMC). We illustrate that, despite enabling high ImageNet classification accuracy, purely perceptual tasks prevent CMC from capturing more high-level semantic structure that would easily be learned by a human. Building from this, we present SemanticCMC. Trained on a naturalistic movie dataset with meaningful temporal co-occurrence patterns, we illustrate that this alternate task improves coding of concept semantics despite attenuated classification accuracy. This preliminary analysis on a single self-supervised network highlights that reliance on object-level decoding does not always indicate meaningful concepts have been captured. By investigating the nature of the content learned by DNNs, we can improve our understanding of their similarities and differences to human vision and progress towards naturalistic visual machine learning in the real world.

1. Introduction

Supervised DNNs excel at classification tasks, applicable to a wide range of problems in engineering, science and technology. However, the nature of their training and the dependence of the current state-of-the-art on highly-curated, labelled datasets limits the potential of modern computer vision to emulate learning in the real world. There is currently a failure in the computer vision community to look into the why and how of these networks' learning mechanisms - something which is imperative if we are to progress towards more robust and generalisable machine learning.

Models that learn *via* self-supervision are exciting candidates for more naturalistic computer vision, perhaps capable of intelligent learning in the quotidian environment or as better models of human vision. Indeed, early insights suggest that such learning curricula can model neural and behavioural responses well and are even useful for forming hypotheses about the brain's own learning [18]. Unlike DNNs, the human visual experience is not solely reliant on statistical regularities of local image features [1], but accounts for more global structure between objects in the world. This helps to build semantic meaning into foundational concepts that are learned through vision; so much of our early experience as a pre-verbal infant involves learning what things are from diverse patterns in our visual inputs. The current literature fails to fully account for the presence or absence of this global, high-level information in computer vision models.

Here, we take a step back from performance metrics and employ techniques from cognitive science to explore the nature of the representations learned by a self-supervised network [14]. Roads and Love show in [13] that when the relational structure of concepts is accurately captured across a range of modalities, their idiosyncratic signals within each system can be leveraged to align the concepts and form a more integrated, distributional account of their meaning; one cannot fully understand what something is unless one understands how it relates to other things. By investigating and building this level of understanding in computer vision models, perhaps a more complete representation of concepts can be learned.

Echoing this emphasis on relational structure, there is evidence that humans and DNNs can learn semantics from

the typical context in which an object is found [11]. Text-based distributional semantic models (DSMs) can perform complex analogy tasks by learning vector representations for words from their context in a large corpus of text [9], a more global view that opposes the feature-level focus seen in visual DNNs. We propose that a similar method of learning from context is possible in pixel-based algorithms.

In naturalistic scenarios, objects have meaningful co-occurrence patterns across space and time. For example, a chair and a lamp are visually distinct but tend to occur in similar contexts facilitating their categorisation into the broader concept of living room furniture. While common scene datasets such as COCO [7] better account for spatial co-occurrences, no image dataset properly accounts for the temporal associations of objects that would be present in a naturalistic setting. We tested whether these temporal co-occurrences could be used to train a self-supervised network and, importantly, whether the information present in these co-occurrences would lead to a more semantic representation.

2. Approach

Our experiment uses images taken from naturalistic movies to train CMC [14], a self-supervised network. We prioritise a self-supervised approach given its more cognitively plausible learning mechanism as it does not rely on a large number of labels. The contrastive task itself builds a representation that is useful for object recognition by discriminating between observed data and simulated noise, thereby maximising mutual information between two images [5].

Video data are very rich, and a notable jump in unsupervised learning performance was observed when their sequential information was incorporated into training methods [15]. We use pairs of still movie frames which are separated by a fixed lag, but we do not use perceptual video data such as motion or optical flow. This preserves the task as one of image-based representation learning, but simply introduces a new level of information in the temporal relations between images.

The idea of temporal coherence has previously been used as a training signal for deep learning [4, 10]. This describes how neighbouring frames in a video are likely to be similar, and can be used to learn features which are invariant to slight shifts in the inputs over time [16]. We build upon this but explore the presence of a much slower type of association: the long-range co-occurrence patterns of objects in the visual environment and the potential semantic structure that could emerge from such patterns. In fact, we suspect that the temporal coherence of perceptual features at shorter timescales will be too high to reveal the high-level associative structure, as perceptual similarity will dominate over more global co-occurrence patterns.

We hypothesise that CMC will be able to learn semantic structure by predicting an image which is related by a lagged interval. This is similar to the temporal prediction idea described in Contrastive Predictive Coding (CPC) [12] which inspired the network used here. However, we do not give the network sequential information of patches from an image, instead leveraging the long-range associations in a naturalistic dataset. Our hypothesis is partly inspired by text-based DSM models such as word2vec which learn word meaning from context in a large corpus of text [9]. Instead, we ask if object meaning can be learned from visual context across time. Despite its considerations of temporal data, we chose not to use CPC but rather its more malleable successor, as CMC is inspired by cross view representation learning in humans and is very flexible with regards to its input views or modalities.

3. Dataset description

A key hypothesis of this work is that the semantic quality of CMC's representations can be improved by training on naturalistic images with meaningful temporal co-occurrence patterns that would appear in the real world. We therefore needed to diverge from training on important but highly-curated computer vision datasets such as ImageNet. We constructed a new dataset using feature length films as a proxy for the real-world environment. One might expect, in an imperfect analogy, that a pre-verbal infant experiences its visual world in snippets of contexts similar to those portrayed in movies with realist genres. Of course, using films is not a perfect model for the world but it is certainly more appropriate for our purposes than ImageNet. Note that efforts to provide datasets which are more ecologically valid as well as being suitable for large-scale deep learning have been shown to lead to improved computer vision models of human vision [8].

158.4 hr of video were chosen whose worlds were deemed to have naturalistic visual scenarios (*e.g.* *Bridget Jones Diary*, 2001 or *The Social Network*, 2010). A unique image dataset was created from the movies by taking a still image every 1 sec, giving 572,949 naturalistic images. Importantly, the temporal structure of the videos was preserved in their sequence such that two images separated by a specified time lag were related in a manner that would hold meaning in the natural world.

3.1. Movie dataset regression analysis

We aimed to use the co-occurrence patterns of objects in these images to train a DNN on a contrastive learning task. Thus, the presence of such patterns and the structure of the object relationships within the movie dataset were first examined.

Every 200 ms of video from the 158.4 hr of video was automatically tagged using Amazon's Rekognition service.

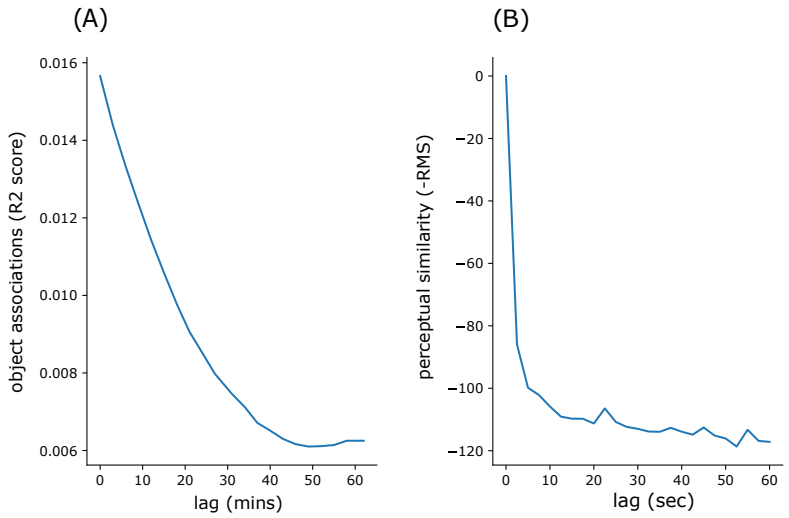

Figure 1. Temporal co-occurrence correlations of objects persist for an extended period of up to 40 min. In contrast, the perceptual similarity of inputs decrease rapidly within 20 sec. This difference in timescales presents as a potential signal for self-supervision. (A) R2 score of the autocorrelation of objects in the movie dataset vs. lag distance (Section 3.1), indicative of temporal associations. (B) The pixel wise -RMS difference of two movie images over increasing lag distance. The negative RMS similarity is plotted for ease of comparison to (A).

The labels were used to quantify object co-occurrences across an increasing lag distance, thereby analysing whether objects that occurred closer together in time were more strongly associated. Note that these labels were not used in subsequent deep learning experiments, but simply to examine the structure in the dataset. To simplify the computation, the 150 most frequently occurring labels in the movie images were calculated and a 2,851,272x150 matrix was constructed with each row representing a 200 ms movie frame and each column an object. A binary encoding indicated the presence or absence of an object at a timepoint.

Using an autoregressive ridge regression model (α=1.0), the probability of appearance of the 150 objects at t_0 was predicted from the set of objects present at a lagged interval earlier (t_{lag}). Across models, this interval ranged from 1 lag of 200 ms to an increasing lag distance ($\Delta t = t_{lag} – t_0$) of up to 1 hr. The coefficient of determination (R^2 score) of the ridge model was plot over an increasing lag distance (Fig. 1A). It was expected that object associations would be stronger at shorter lag distances, illustrated by a stronger model fit and that this would decrease as the time between two object occurrences was increased. This would be explained by the fact that, at longer Δt, objects would appear in quite different contexts or environments which would lead to weaker correlations.

It was found that object associations persisted for much longer than was expected, decreasing approximately linearly as Δt increased to ˜40 minutes. This long window of temporal association for correlated objects was initially surprising. However, it likely reflects that in the movies, as in life, the visual context or scene will persist for an extended period of time as one lingers in the same place or situation. As discussed in Section 2, at low values for Δt the temporal coherence of nearby frames can provide a useful signal for unsupervised training of object decoding networks [10, 4] but this purely perceptual cue is exactly the type of signal we aimed to reach beyond. We hypothesised that the strong perceptual similarity of two images at low Δt would dominate the learning signal and preclude CMC from building representations with more semantic meaning.

To quantify the change in perceptual similarity of images over time, the mean pixel wise root mean square (RMS) difference of two images was calculated over a range of values for Δt (n=1000 images per lag). It was found that perceptual similarity decreased much more rapidly than object associations, reaching a minimum within the first 20 seconds (Fig. 1B). This difference in timescales between perceptual and semantic similarity in the movie dataset indeed provides a signal that can be leveraged for self-supervised learning.

Finally, the nature of the object associations found from the autocorrelation analysis were examined. Hierarchical clustering with Ward linkage was performed on the pairwise matrix of object regression coefficients. The emergent clusters were semantically interpretable and could be manually assigned categorical labels such as furniture, clothing or electronics; objects belonging to the same broad, superordinate category such as chair and closet, were more correlated across time. Given this, we went on to test if training CMC on the signal illustrated in Fig. 1 would enable these semantic clusters to be captured in its representations.

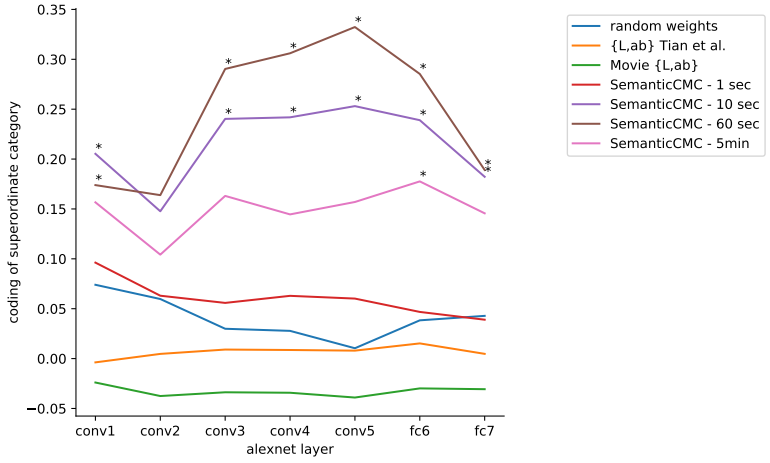

Figure 2. Coding of superordinate category in each training regime, across all AlexNet layers. Correlation was calculated using the Mantel test with Pearson correlation. An intermediate lag distance of 60 s best captured superordinate level categorisation. * denotes significant correlation ($p < 0.05$, Bonferroni corrected across 7 AlexNet layers).

4. Method

Contrastive Multiview Coding (CMC) proposed by Tian *et al*. [14] is inspired by the brain's view-invariant representation encoding as researched in cognitive science and neuroscience. It leverages co-occurrence patterns across multiple views of data similar to the approach described by Oord, Li and Vinyals in [12], allowing mutual information to be learned across modalities or viewpoints. Noise-Contrastive Estimation (NCE) loss [5] is calculated in the latent space. CMC offers great flexibility and cognitive plausibility in its modality or viewpoints *via* choice of its encoding network and definition of auxiliary task. We introduce SemanticCMC, trained on our naturalistic movie dataset, and explore how this new task affects the quality of concepts in CMC's learned representations. Weights from a variety of training regimes with CMC on an AlexNet architecture were examined using Representational Similarity Analysis (RSA) [6].

4.1. Training CMC on naturalistic images

A total of eight CMC training regimes were analysed. First, we used the high-performing weights published by Tian *et al*. for CMC-AlexNet trained on a purely perceptual luminance vs. chrominance auxiliary task ($\{L,ab\}$). We hypothesised that, although capable of reaching 42.6% top-1 accuracy on the 1000-way ImageNet classification transfer learning task [14], the relational structure of these representations may be lacking in semantic similarity structure due to its reliance on solely perceptual cues.

As a control for our new dataset, we trained CMC using the $\{L,ab\}$ auxiliary task on the movie images. This reached 32.38% top-1 and 54.3% top-5 accuracy on the ImageNet classification transfer learning; an interesting display of the utility of this self-supervised framework for successful object decoding without relying on training with the highly-curated ImageNet. Two further baseline trainings were examined by initialising AlexNet with random weights and with supervised weights loaded from PyTorch.

Next, CMC was trained on our SemanticCMC task with the movie dataset. Using one full-sized AlexNet as the encoding network, two images which were separated by a specified time lag (Δt) were loaded and passed through the same encoder. This contrasts to the $\{L,ab\}$ task in which one image was split across its channels and passed through half of an AlexNet encoder [17]. Contrastive loss was calculated in the latent space to identify the positive pair (i.e. the two images connected by Δt) from other randomly selected negative samples. SemanticCMC was implemented as a finetuning procedure on top of the published weights from [14], motivated by initial experimentation and the fact that semantic associations would likely be learned better having first found useful visual features for basic level recognition.

SemanticCMC was run on a range of values for Δt (1 sec, 10 sec, 60 sec, 5 min) chosen to reflect intervals at which object associations would be high, but perceptual similarity is decreasing (Fig. 1). Interestingly, as Δt increased, the value to which loss converged increased (1 s loss, 6.20; 10 s loss, 9.29; 60 s loss, 10.97; 5 min loss, 11.29) giving a first indication that there was different information to be learned from temporal signals in the movies, depending on the magnitude of Δt. For comparison to the transfer learning capabilities of the $\{L,ab\}$ trained networks, ImageNet validation was used to test

SemanticCMC-60sec, resulting in extremely poor validation accuracy (1.89% top-1, 5.77% top-5). Despite this poor classification performance, we went on to test whether the representational structure of SemanticCMC was in fact meaningful, and how it compared to the better performing networks.

4.2. Implementation details

With the exception of the mantel tests described in Section 4.3 which were run using the 'vegan' package in R, all analyses were coded in Python 3.7 using PyTorch 1.4.0 with CUDA v10.2, and run on RTX 6000 GPUs each with 24 GB in a Lambda quad workstation with 28 CPU cores and 128 GB of RAM. Pretraining with the movie image dataset on the $\{L,ab\}$ task (Section 4.1) ran to convergence at 200 epochs with a batch size of 128 and a learning rate of 0.03 with decay by 0.1 at epochs 120 and 160 using a SGD optimiser with 0.9 momentum. As originally described by Tian *et al*. [14] the encoder was a SplitBrain AlexNet architecture. Batch normalisation was used and images were transformed into the $\{L,ab\}$ space with random resized crops and random horizontal flipping. Linear decoding was performed on top of AlexNet convolutional layer 5 for 60 epochs using an SGD optimiser with an initial learning rate of 0.1 and decay by 0.2 at epochs 30, 40 and 50.

Temporal training (Section 4.1) was performed using one full-sized AlexNet as the encoder network. AlexNet was initialised with the published weights for CMC, as trained by the $\{L,ab\}$-ImageNet task and then finetuned on our SemanticCMC objective. We tested whether training the network from scratch on SemanticCMC was worthwhile, but found that the time expense did not justify the result. Moreover, semantic analyses (as in Section 4.3) revealed that when training from scratch, coding of semantic relations was not as strong as with the fine tuning procedure (although the trend in the results reported below persisted). This led us to the conclusion that prior knowledge of object features is a useful basis for learning semantic structure, and motivated our choice of finetuning procedure.

To prevent the network from cheating its learning based only on the colour histogram of the images, SimCLR the colour distortion method described in Chen *et al*. [3] was used instead of an *Lab* transform, as well as random resized crops and random horizontal flipping. Temporal finetuning was run for 80 epochs, with a batch size of 128 and a learning rate of 0.03 with decay by 0.1 at epochs 30, 50 and 70. Batch normalisation was used, and a SGD optimiser with momentum of 0.9. The differences in loss values reported in Section 4.1 were only observed when inputs were transformed with the colour distortion method described in [3]. This suggests that without colour distortion CMC was using an alternative perceptual cue, the colour histogram.

4.3. Representational similarity analysis

RSA characterises a representation within a system by the distance matrix of the response patterns elicited by a set of stimuli. A two dimensional representational dissimilarity matrix (RDM) is constructed from the pairwise distances between patterns of activations in vector space. This gives insight into how similar or dissimilar objects are 'thought' to be by the system; there is a greater distance between two unrelated object vectors and a shorter distance between two similar objects. RDMs were constructed from AlexNet loaded with the learned weights from the eight training regimes described in Section 4.1 (random weights, supervised, $\{L,ab\}$ task as published by [14], $\{L,ab\}$ on the movie dataset, SemanticCMC with a Δt of 1 sec, 10 sec, 60 sec and 5 min).

As an initial experiment, network activations were calculated in response to the ImageNet categories overlapping with the 150 frequent labels used in the regression analysis described in Section 3.1 (n=25 classes). 50 randomly sampled ImageNet exemplars per class were passed through each frozen weights network and the pairwise distances between each class's activations were structured into a matrix. These activation RDMs were correlated to a binary category model matrix that coded for which pairs of objects occurred in the same category cluster returned by the hierarchical clustering; *e.g*. wine and table were clustered together, and so they received a value of 1 while table and hair were not clustered so were coded for with a 0. This binary matrix therefore modelled a scenario entirely explained by the superordinate clusters derived from the label regression analysis. Using a Mantel test to correlate the activation RDMs to the superordinate category model, it was found that SemanticCMC–60sec best captured semantic content, while networks trained on a purely perceptual task did not significantly correlate to the model RDM (Fig. 2). Convolutional layer 5 was most correlated to the semantic model.

Although this analysis was indicative that semantic content could be learned from SemanticCMC training at an appropriate Δt, it was flawed in that it used a small number of ImageNet categories (25); it made use of an ill-defined semantic measure (the clustering results); and it was based on a regression that only examined the 150 most frequent objects in the movie dataset. Thus, we extended the evaluation to construct RDMs from the mean activation to 256 randomly selected ImageNet classes (n=150 images per class). The pairwise distances between each class's activations in the frozen weights networks were calculated and stored in a 256x256 RDM.

To further improve the evaluation, the quantification of semantic content was improved using the WordNet Leacock Chodorow (LCH) similarity scores for every pair of the ImageNet classes tested. LCH quantifies the shortest distance between two classes in the WordNet hierarchy tak-

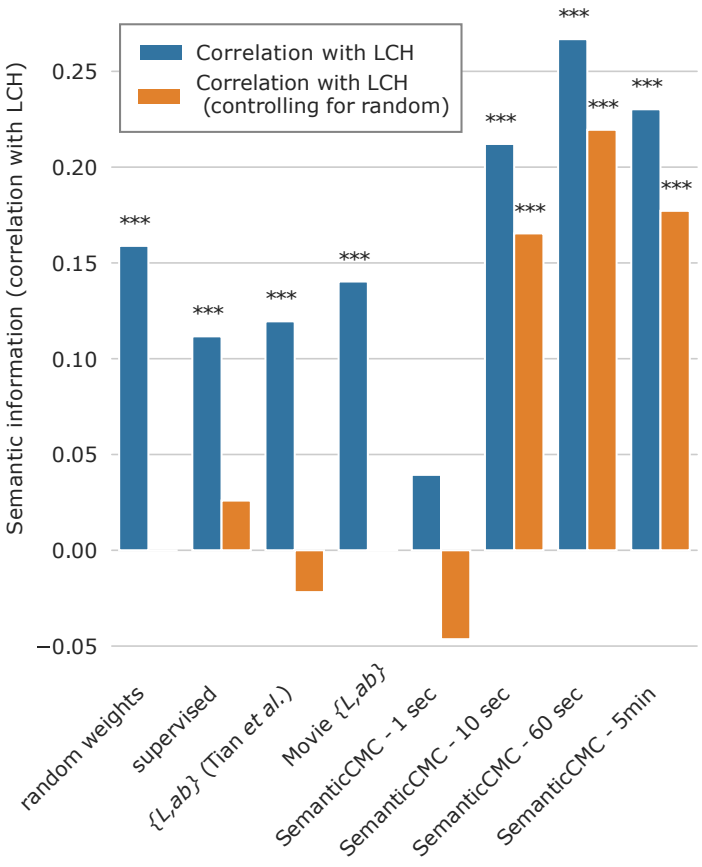

Figure 3. Results of Mantel test and partial Mantel test (Section 4.3, Pearson correlation, 999 permutations). All results show correlation of the activation RDM for AlexNet Convolutional Layer 5 to the LCH RDM. Orange bars illustrate partial Mantel tests, controlling for perceptual cues using activations from the randomly initiated network. It was shown that only SemanticCMC at a sufficient temporal distance acquired additional semantic knowledge. (*** denotes significant correlation p <0.001).

ing into account the depth of taxonomy, making it a suitable measure for quantifying semantic similarity. A 256x256 pairwise semantic LCH model was constructed from the WordNet similarity scores. Using a Mantel test with Pearson's product moment correlation (number of permutations = 999) the relationship between LCH and network activations were found.

5. Results and discussion

As previously reported, ImageNet training using the $\{L,ab\}$ task for CMC with an AlexNet base architecture is capable of high object-level decoding accuracy (42.6% top-1) [14]. Note that this result can be improved with ResNet architectures. We have shown that CMC, when used with a completely different dataset generated from naturalistic feature-length movies, can still achieve moderate classification performance at 32.38% top-1. This result in itself is an interesting display of the utility of self-supervised networks for learning on less curated, computer vision specific resources. However, the aim of this paper is to explore the representational and semantic quality of the DNNs' learned embeddings.

5.1. Associations at an intermediate lag distance capture semantic information

Initial results shown in Fig. 2 revealed that a DNN trained to maximise mutual information between two naturalistic images separated by an intermediate lag distance of 60 sec was better able to capture relational semantic structure. A purely perceptual task did not succeed in learning this high-level information. Similarly, when the distance between two images was either too long or too short, semantic learning was attenuated. This indicates that there is an optimal value for the temporal distance between objects that can be leveraged for learning high-level object associations. Subsequent RSA evaluations develop and improve upon these initial findings.

5.2. Training on temporal patterns in naturalistic images forms more semantically relevant representations

The results of the RSA evaluation described in Section 4.3 are illustrated in Fig. 3. The extent to which each network's representations captured semantic information is quantified by the magnitude of the Pearson product moment correlation with the LCH model.

It was found that all networks except for one were significantly correlated to the semantic model (Fig. 3, blue bars, $p<0.001$). Each mode of training - be it randomly initiated, supervised, perceptual CMC or SemanticCMC - did capture a representation that correlated with the semantic relations described by the LCH model. SemanticCMC-1sec was the exception, explained by the fact that it was probable for two movie images with $\Delta t = 1$ sec to be almost exactly the same. This means that there was very little signal from which to learn meaningful structure with the SemanticCMC auxiliary task; images were essentially identical and therefore there was little to gain by contrasting them.

We found that the RDM of the random-weights network correlated with the LCH RDM to some degree. This randomly initiated network could only concern perceptual features, extracted from the convolutional architecture of the AlexNet encoder that was used in all networks tested. This suggests that semantically related objects hold superficial visually similarities. To measure the *extra* semantic information learned by each network, we controlled for perceptual similarity using a partial Mantel test. This measured the correlation between a network's RDM and the LCH RDM, while partialling out the random-weights pairwise activation patterns.

When controlling for perceptual similarity in this manner, the results were vastly different. We found that the only networks that preserved their significant correlation to the LCH semantic model were those trained on SemanticCMC at a sufficient value for Δt (10 sec, 60 sec and 5 min), with an intermediate lag distance once again being preferable for capturing semantic structure (Fig. 3, orange bars, $p<0.001$). It can be inferred that at short distances in a naturalistic visual dataset, temporal co-occurrence patterns of objects are not informative for forming an accurate representation of semantics. Similarly, at too long a distance the correlations of objects begin to weaken and learning is not as effective. These results therefore indicate that at intermediate temporal windows the co-occurrences of objects provide a useful signal for self-supervised learning of visual representations, leading to more semantically meaningful concepts being embedded into the networks' representations.

5.3. High object-level decoding performance does not imply that meaning has been captured

As displayed in Table 1, high top-1 accuracy on an ImageNet transfer learning classification task was not indicative of high correlation to a semantic model. Representations formed by CMC trained on the perceptual task were indeed useful for classification, but were not correlated with the semantic model once perceptual similarity had been controlled for. In contrast, SemanticCMC-60sec was not capable of ImageNet classification, but was most strongly correlated to the semantic model even when perceptual similarity had been controlled.

This result highlights a key oversight when interpreting most evaluations of computer vision models. Often, success is narrowly defined as percentage increases in performance benchmarks. The innovation presented by researchers in the computer vision community is unparalleled, yet there is little effort to understand why or how a DNN is behaving the way it is. While this "is it working" approach has merits for myriad applications in engineering or commercial practice, a deeper understanding of the models being published will enable greater success in future innovation. The efficacy of having correct relational structure in unsupervised systems has been shown [13], and huge progress has been made in natural language processing with models that accurately capture semantic biases [2]. While this goal may not be applicable to every computer vision problem, one can envision a future where such models must better capture concepts' meaning to perform robustly in the real world.

5.4. Limitations and future direction

There are notable limitations to the work presented here. We have only tested one self-supervised system, CMC. This was motivated by CMC's inspiration from cognitive science and its learning across multiple viewpoints of a stimulus; it was natural to investigate learning from temporal co-occurrence patterns by taking two viewpoints as two images separated by a time lag. Further work will extend these findings to other self-supervised frameworks as well as testing encoder networks other than AlexNet. Furthermore, Se-

Network (AlexNet)	**Task and Dataset**	**Top-1 Accuracy**	**Semantic Correlation**	
			Mantel Test	Partial Mantel
CMC	L vs. ab - ImageNet	42.60%	0.1196 (***)	-0.02184 (n.s.)
CMC	L vs. ab - Movies	32.38%	0.1403 (***)	0.0001595 (n.s.)
SemanticCMC-60s	t_0 vs. t_{lag} - Movies	1.89%	0.2668 (***)	0.2195(***)

Table 1. Classification accuracy vs. semantic content. Mantel results report Pearson's product moment correlations to the LCH semantic matrix. Partial Mantel tests controlled for the randomly initialised network activations i.e. perceptual content. It was found that high object-level decoding accuracy did not indicate that semantic content was captured by the network.

manticCMC was implemented as a finetuning procedure on top of pretrained weights. Although this choice was well motivated (Section 4.2) a system that concurrently learns to recognise and relate objects would be preferable. Finally, although improving ImageNet classification accuracy was not our aim, a system that is capable of capturing semantic concepts as well as performing successful transfer learning to state-of-the-art benchmarks is a worthy big picture goal. It will be useful and worthwhile to apply the present analyses to more SOTA computer vision networks with the aim of finding a solution to both the problem of high classification and semantically meaningful representations. Additionally, it will be interesting to test if the improved semantic concepts captured by SemanticCMC better correlate to human behavioural and neural data.

6. Conclusion

The results presented here show that, at intermediate intervals, the temporal co-occurrence patterns within a naturalistic movie dataset provide a signal for self-supervised learning of more semantic visual representations. This result was not indicated by better object-level decoding performance, highlighting a key pitfall in relying solely on classification benchmarks for the evaluation of computer vision models. By doing so, we fail to account for the quality of concepts within the learned representations. Furthermore, we have shown that a naturalistic dataset which was not expertly designed for training purposes can be used in self-supervised learning scenarios. These results provide preliminary stepping stones towards more robust, naturalistic and human-like computer vision.

Acknowledgements

This work was generously funded by the ERC Advanced Grant ERC-2017-ADG, FOUNDCOG, 787981.

References

[1] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. 1

[2] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 7

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 5

[4] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 4086–4093, 2015. 2, 3

[5] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010. 2, 4

[6] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008. 4

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[8] Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8), 2021. 2

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[10] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Proceedings*

of the 26th Annual International Conference on Machine Learning, pages 737–744, 2009. 2, 3

[11] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007. 2

[12] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

[13] Brett D Roads and Bradley C Love. Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, pages 1–7, 2020. 1, 7

[14] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 1, 2, 4, 5, 6

[15] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015. 2

[16] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. 2

[17] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 4

[18] Chengxu Zhuang, Siming Yan, Aran Nayebi, and Daniel Yamins. Self-supervised neural network models of higher visual cortex development. In *2019 Conference on Cognitive Computational Neuroscience*, pages 566–569, 2019. 1