SemanticCMC - improved semantic self-supervised learning with naturalistic temporal co-occurrences

Cliona O'Doherty
Trinity College Institute of Neuroscience
School of Psychology
Trinity College Dublin
odoherc1@tcd.ie

Rhodri Cusack
Trinity College Institute of Neuroscience
School of Psychology
Trinity College Dublin
cusackrh@tcd.ie

Abstract

Self-supervised learning is advancing state-of-the-art in machine learning methods, lessening our reliance on curated datasets. To understand why and how these models are performing well it is useful to examine the representational or cognitive foundations on which they execute their behaviours. Here, we present an application of Representational Similarity Analysis (RSA) to investigate the patterns of activations within a top-performing self-supervised network, Contrastive Multiview Coding (CMC). We illustrate that, despite enabling high ImageNet classification accuracy, purely perceptual auxiliary tasks prevent a self-supervised network such as CMC from capturing more high-level semantic structure. We present SemanticCMC, trained on a naturalistic movie dataset with meaningful temporal structure. We illustrate that this semantic task improves coding of concept semantics despite its attenuated classification accuracy. This preliminary analysis on a single self-supervised network highlights that reliance on object-level decoding does not always indicate that meaningful structure has been captured. By investigating these cognitive underpinnings of how artificial networks represent concepts, we can improve theoretical understanding of SSL and motivate its engineering progress for a breadth of applications.

1 Introduction

As SSL networks continue to beat computer vision classification benchmarks, it is important to understand how they represent and understand the objects they can recognise. Models that learn *via* self-supervision are exciting candidates for more naturalistic, human-like learning, perhaps capable of intelligent learning in the real world or as better models of human vision. Indeed, early insights suggest that such learning curricula can model neural and behavioural responses quite well and can be used to form hypotheses about the brain's own learning [Zhuang et al., 2019]. Yet, these models are often evaluated in a way that overlooks key issues such as their local feature bias and ignorance to more global, semantic structure [Brendel and Bethge, 2019]. Capturing this relational structure is important for unsupervised systems' accurate embedding of meaningful concepts [Roads and Love, 2020]; thus, it is worthwhile to investigate the true semantic quality of representations learned through self-supervision.

Taking a step back from performance metrics and instead employing the computational cognitive science method of representational similarity analysis (RSA) [Kriegeskorte et al., 2008] we use CMC [Tian et al., 2019] to explore how semantics are coded within a self-supervised representation. We hypothesised that the perceptual *{L,ab}* task described by Tian *et al.* would lack in global, interclass semantic structure and that this acquisition of semantic knowledge could be improved by using a

Self-Supervised Learning-Theory and Practice Workshop,34th Conference on Neural Information Processing Systems (NeurIPS 2020).

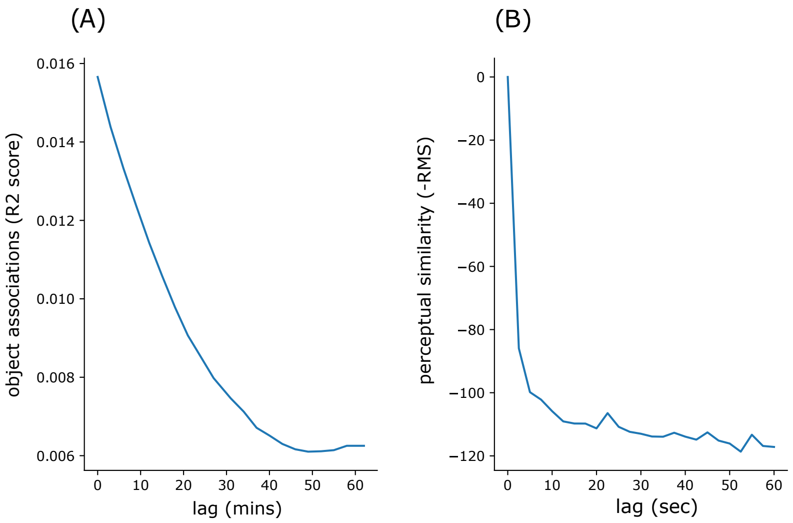

Figure 1: Temporal co-occurrence correlations of objects persist for an extended period of up to 40 min. In contrast, the perceptual similarity of inputs decrease rapidly within 20 sec. This difference in timescales presents as a potential signal for self-supervision. (A) R2 score of the autocorrelation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

more naturalistic auxiliary task. In the real world, objects in the visual environment have meaningful temporal co-occurrence patterns. For example, chair and lamp are visually distinct but tend to occur close together in space and time, facilitating their categorisation into the broader category of living room furniture. We tested whether these co-occurrences could be used to train CMC, and if the resultant representations were better correlated to the semantic measure of WordNet LCH similarity ratings. This method enables a cognitive investigation of SSL as potential grounds for improved theoretical insights into this realm of machine learning.

2 Method and results

2.1 Naturalistic movie dataset

A key hypothesis of this work is that the semantic quality CMC's representations can be improved by training on naturalistic images with meaningful temporal co-occurrence patterns. To test this, a new dataset was constructed. Feature length films were used as a proxy for the real-world environment. 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. No video data (e.g. motion or optical flow) were used. However, 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 meaningful way. Using autocorrelation (See Appendix A.1) to quantify the co-occurrence patterns of objects, it was found that labels belonging to the same broad, superordinate category (e.g. chair and closet) were more correlated across time. These temporal associations persisted for much longer (up to 40 min) than the perceptual similarity of two movie images (20 sec) (Figure 1). This difference in timescales between perceptual and semantic similarity in the movie dataset presents an opportunity for learning.

2.2 Training CMC on naturalistic images

Contrastive Multiview Coding (CMC) proposed by Tian et al. [2019] 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 et al. [2018], allowing mutual information to be learned across modalities or viewpoints. Noise-Contrastive Estimation (NCE) loss [Gutmann and Hyvärinen, 2010] 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 RSA [Kriegeskorte et al., 2008].

A total of eight CMC training regimes were analysed. First, we used the high-performing weights published by Tian *et al.* for CMC trained on a purely perceptual luminance vs. chrominance*{L,ab}* auxiliary task. We hypothesised that, although capable of reaching 42.6% top-1 accuracy on the 1000-way ImageNet classification task, the relational structure of these representations may be lacking in semantic similarity structure as it learns by focusing solely on 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 1000-way ImageNet classification task; 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 using 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. 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 Tian et al. [2019], motivated by the fact that semantic associations would likely be learned better having first found useful visual features for basic level recognition.

See Appendix A.3 for detailed implementation details. SemanticCMC was run on a range of values for Δt. 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 are different types of information to be learned from temporal signals, depending on lag distance. 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.

2.3 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 2.2 (random weights, supervised, *{L,ab}* task as published by Tian et al. [2019], *{L,ab}* on the movie dataset, SemanticCMC with a Δt of 1 sec, 10 sec, 60 sec and 5 min). The activations for each layer were calculated in response to 256 classes randomly selected from the full ImageNet databse (mean activation per layer, n=150 images per class). The pairwise distances between each class' activations were calculated and stored in a 256 X 256 RDM.

2.4 Naturalistic temporal associations provide better learning signals for semantic structure

To quantify semantic coding within each of the representations, the activation RDMs were correlated to a matrix containing 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, taking into account the depth of taxonomy, making it a suitable measure for quantifying semantic similarity. A 256 X 256 LCH RDM was constructed from the WordNet similarity scores using a Mantel test based on Pearson's product-moment correlation. In a pilot analysis on a different set of images and an alternate semantic measure (Appendix B.1, Figure 3) we found that AlexNet's convolutional layer 5 carried the strongest semantic information. Therefore, we used this layer for the current analysis. We found that the RDM of the random-weights network, which only captures perceptual features, correlated with the LCH RDM to some degree, suggesting that semantically related objects are more superficially visually similar. To measure the extra semantic information

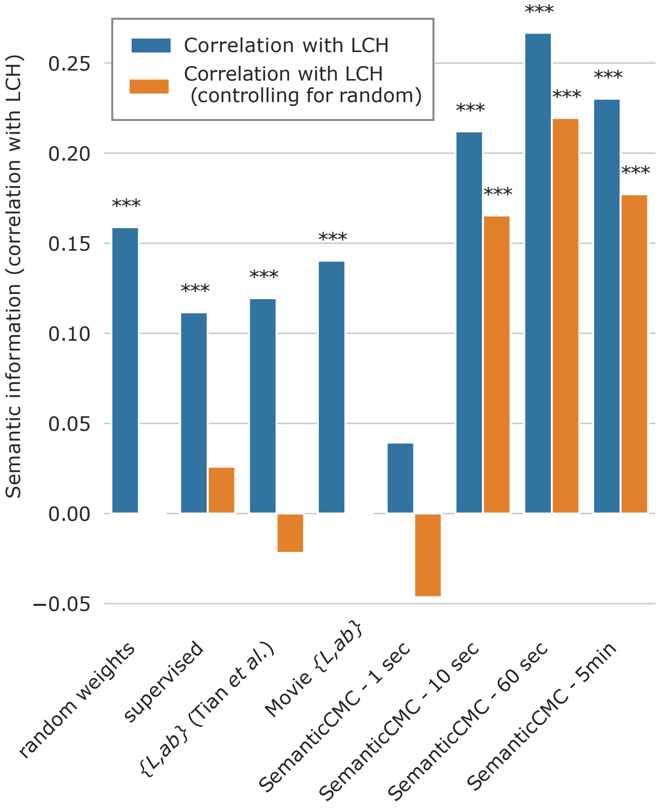

Figure 2: Results of Mantel test and partial Mantel test (Section 2.4, Pearson correlation). All results show correlation of the activation RDM for AlexNet Convolutional Layer 5 to the LCH RDM. Orange bars illustrate partial Mantel tests, controlling for perceptual cues using random weights. It was shown that only SemanticCMC at a sufficient temporal distance acquired additional semantic knowledge. (*** denotes significant correlation $p < 0.001$).

learned by a network, we therefore controlled the perceptual similarity using a partial Mantel test. This measured the correlation between a network's RDM and the LCH RDM, partialling out the random-weights RDM.

It was found that all networks bar one were significantly correlated to the LCH similarity network (Fig. 2, blue bars, $p < 0.001$). SemanticCMC-1sec was the exception, likely explained by the fact that it was probable for two movie images with $\Delta t = 1$ to be almost exactly the same. Hence, there is little signal from which to learn meaningful structure with the SemanticCMC auxiliary task. When controlling for random weights, i.e. perceptual content, the results were vastly different. We found that the only networks that were significantly correlated to LCH were those which had been trained on SemanticCMC at a sufficient value for Δt (10 sec, 60 sec and 5 min) (Fig. 2, orange bars, $p < 0.001$). The intermediate lag distance of 60 sec was optimal for learning improved semantic content. 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. Thus, it can be concluded that at intermediate temporal windows the temporal co-occurrences of objects provide a useful signal for self-supervised learning of semantic similarity structure, providing more meaningful concepts to the networks' representations.

3 Discussion and conclusion

We found that the temporal co-occurrence patterns within a naturalistic movie dataset provided a useful signal for self-supervised learning of object semantics. This improved semantic knowledge was not associated with strong object-level decoding performance, highlighting a key oversight in how many of these models are evaluated. Furthermore, the network trained on a perceptual-only auxiliary task was capable of reaching high ImageNet accuracy, but was not as correlated to semantic measures. Thus, we have shown that high classification accuracy does not always equate to meaningful representations of concepts, and that cognitive computational methods such as RSA help to shed light on the cognitive foundations and relational structure of SSL networks.

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 on encoder networks other than AlexNet. Furthermore, SemanticCMC was implemented as a finetuning procedure on top of pretrained weights. Although this choice was well motivated (see Appendix A.3) 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.

Increased efforts in understanding the cognitive and representational foundations of self-supervised networks will improve theoretical understanding of SSL. Here, we argue that naturalistic training improves semantic capabilities of these models. This may be useful for leveraging SSL for more naturalistic, human-like learning, and improving these systems' capabilities as more realistic models for human vision.

Acknowledgements

This work was generously funded by the ERC Advanced Grant ERC-2017-ADG, FOUNDCOG, 787981. Thank you to all reviewers who have helped to improve upon the evaluation and results presented here.

References

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

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

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.

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.

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

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.

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

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

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

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.

A Additional methods

A.1 Movie dataset regression analysis

Every 200 ms of video from Section 2.1 was automatically tagged using Amazon Rekognition, and the labels were used to quantify object co-occurrences across an increasing lag distance. To simplify the computation, the 150 most frequently occurring labels were calculated and a 2,851,272 X 150 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 to an increasing lag distance ($\Delta t = t_{lag} - t_0$) from 200 ms to 2 hr. Hierarchical clustering with Ward linkage was performed on the resulting 150 X 150 pairwise matrix of regression coefficients showing that objects which were more correlated across time could be grouped into superordinate, semantically meaningful categories such as furniture, clothing or electronics (Table 1).

A.2 Change with Δt of perceptual similarity vs. high-level associations

The mean R^2 score of each lags' regression from Appendix A.1 was plotted over increasing Δt, giving an indication of how well the regression model was performing with changing lag distance (Fig.1A). The pairwise temporal correlations of objects persisted for much longer than was expected. Associations decreased approximately linearly as Δt increased to ~40 minutes. This long window of temporal association for correlated objects was initially surprising. However, it likely reflects the fact that, in the movies as in life, the visual context or scene will persist for an extended period of time. At low values for Δt, temporal coherence can also be used to train a network for object-level decoding [Mobahi et al., 2009, Goroshin et al., 2015]. We hypothesised that the similarity of two images this close apart in time would dominate the learning signal and preclude the learning of global

Table 1: Frequent objects from the movie dataset that were also present in ImageNet. Objects are clustered according to Appendix A.1. The category label for each cluster was manually assigned.

Superordinate Category	Item	Superordinate Category	Item	Superordinate Category	Item
clothing	gown	computer	screen	furniture	lamp
	hair		computer		couch
	suit	dining	table		chair
	coat		food		closet
	tie		restaurant		piano
	shirt		glass		pillow
	sunglasses		alcohol		desk
	shoe		wine		bannister
					window

semantic relations. Furthermore, SemanticCMC on with low Δt would not learn good semantics even from perceptual cues, as the two images it loads are almost identical and therefore uninformative for contrastive learning. 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) showing that perceptual similarity decreased more rapidly than object associations, reaching a minimum within the first 20 seconds (Fig. 1B).

A.3 Implementation details

Mantel tests described in Section 2.4 were run using the 'vegan' package in R. All other 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 2.2, Movie *{L,ab}*) 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. [2019] a SplitBrain AlexNet architecture was used. 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 2.2) 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. When training the network from scratch on SemanticCMC, correlations to semantic category were not as strong, leading us to the conclusion that prior knowledge of object features is a useful basis for learning semantic structure, and motivating our choice of finetuning procedure. To prevent the network from cheating its learning based only on the colour histogram of the images, the colour distortion method described in Chen et al. [2020] 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 2.2 were only observed when inputs were transformed with the colour distortion method described in Chen et al. [2020]. This suggests that without colour distortion CMC was using an alternative perceptual cue, the colour histogram.

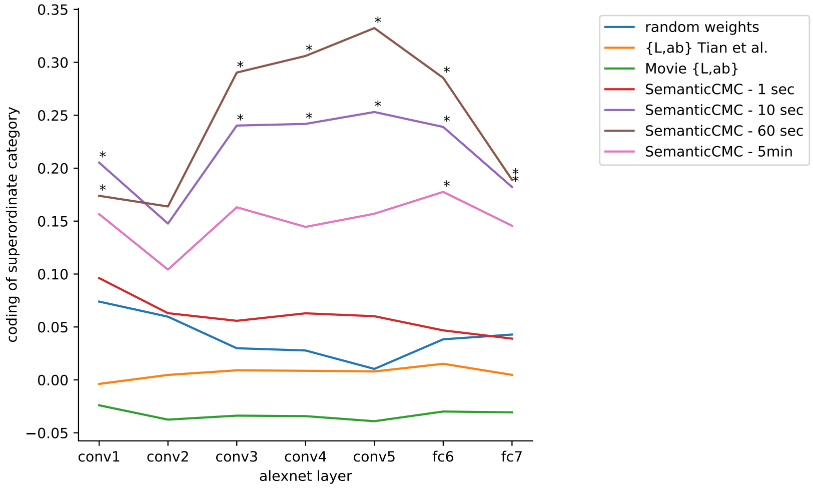

Figure 3: Coding of superordinate category in each training regime, across all AlexNet layers. Correlation was calculated using the Mantel test with Pearson correlation. An intermediate lag distance of 60 s best captured superordinate level categorisation. * denotes significant correlation ($p < 0.05$, Bonferroni corrected across 7 AlexNet layers).

B Supplementary results

B.1 Initial results for learning of superordinate semantics

Initial RSA experiments were conducted using the ImageNet categories in the superordinate clusters derived from the hierarchical clustering of regression coefficients in Appendix A.1. These are shown in Table 1. 50 randomly sampled ImageNet exemplars were show per class, and activation RDMs were constructed as described in Section 2.3. Activation RDMs were correlated to a binary category model RDM that coded for which pairs of objects occurred in the same category (e.g. wine and table received a value of 1 and table and hair was coded for with a 0). This modelled a scenario entirely explained by the superordinate clusters in Table 1. It was found that the SemanticCMC – 60 sec network best captured superordinate semantics, while perceptual only networks did not significantly correlate to the model RDM (see Figure 3).