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

Cliona O'Doherty, Rhodri Cusack
Trinity College Institute of Neuroscience, Trinity Coellege Dublin, Ireland

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Women in Computer Vision

CVPR VIRTUAL JUNE 19-25

Motivation

Evaluation of computer vision networks is often overly relient on classification accuracy. Are we lacking insight in how these networks understand?

Can we improve semantic knowledge within CNNs by implementing more naturalistic, brain-inspired mechanisms?

Will temporal co-occurrences of objects provide a signal for self-supervised learning of objects?

Method

Overarching goal

- To investigate if more semantic representations can be learned by a contrastive CNN using the associations present within video.
- To assemble a naturalistic dataset, enabling a richer opportunity for learning concepts and meaning.
- To investigate the timescale at which semantic knowledge can best be learned - we hypothesise that both too short and too long an interval will preclude semantic learning.

SemanticCMC

- **Contrastive Multiview Coding (CMC)** from Tian *et al.* (2019) was modified such that two views were two images separated by a specified **time lag**, encoded by the same AlexNet architecture.
- **Contrastive loss** was calculated in the latent space by selecting the positive pair of images from a distribution of negative pairs. Embeddings of frames at medium time scales would be learned to be represented as more semantically similar.
- SemanticCMC was performed as a fine tuning procedure on top of the published weights for CMC trained on an L vs. AB objective. The network was trained using movie images, and its **representations were evaluated** (Kriegeskorte *et al.* 2008)**.**

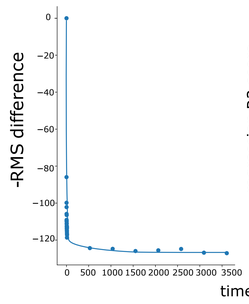
Evaluation

- The **semantic content of learned representations** was quantified using **representational similarity analysis** and without relying on improved object decoding accuracy for evaluation.
- We tested a variety of models and baselines against superordinate category and wordnet semantic models.

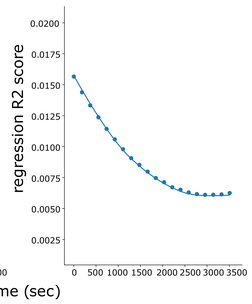
Naturalistic Dataset

- 158.4 hr of live action movies with naturalistic visual worlds.
- Automatically generated labels at sampling interval of 200 ms giving 2,851,272 sets of labels.
- Images taken every 1 sec, giving 572,949 images with preserved temporal relations.

Perceptual Similarity

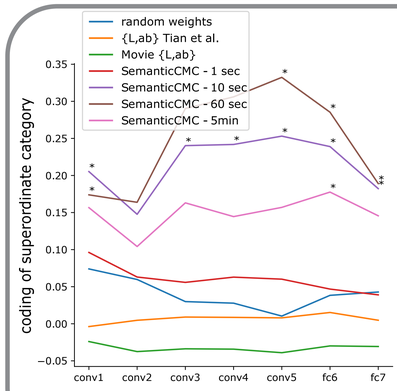


Object Associations



- Association patterns of objects persist for approx. 50 min.
- Perceptual similarity drops rapidly within seconds.
- This presents as a leveragable signal for learning more semantic representations.

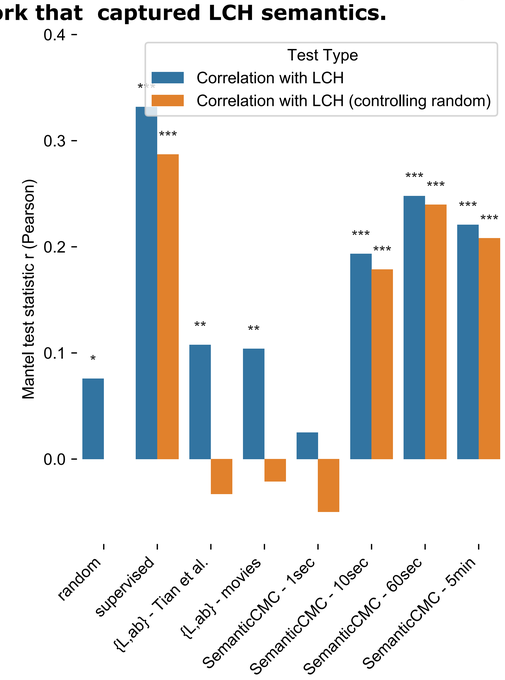
Learning superordinate category



- When trained on a contrastive loss objective across two images separated by 60 sec, SemanticCMC was most correlated to a superordinate category model.
- This coding of superordinate category was not found when the task was purely perceptual ({L,ab}) or when the interval between two images was either too short or too long.

Quantifying semantic content

- We tested whether pairs of classes that were more similar according to the **LCH WordNet measure** were also more similar in their neural network activations patterns.
- When controlling for perceptual content, only **SemanticCMC trained over a sufficient distance in time was the only unsupervised network that captured LCH semantics.**



- **Strong ImageNet classification performance was not indicative of semantic coding.**

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(***)

Conclusion

Naturalistic signals in a movie-derived dataset provide useful signals for learning object semantics, that are not always captured by successful classification performance

References

Tian, Y., Krishnan, D., & Isola, P. (2019). Contrastive multiview coding. arXiv preprint arXiv:1906.05849.
Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2, 4.

erc European Research Council
Established by the European Commission

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

Cliona O'Doherty, Rhodri Cusack
Trinity College Institute of Neuroscience, Trinity Coellege Dublin, Ireland

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Human infants learn what things are simply by observing the world around them. Little supervision is involved.

Can we improve semantic knowledge of ANNs by implementing more human-like training objectives?

Will temporal co-occurrences of objects provide a signal for self-supervision?

Method

Overarching goal

- To test if object semantics can be learned by a self-supervised network from using associations present within video.
- To assemble a naturalistic dataset, enabling a richer opportunity for learning concepts and meaning.
- To investigate the timescale at which semantic knowledge can best be learned - we hypothesise that both too short and too long an interval will preclude semantic learning.

SemanticCMC

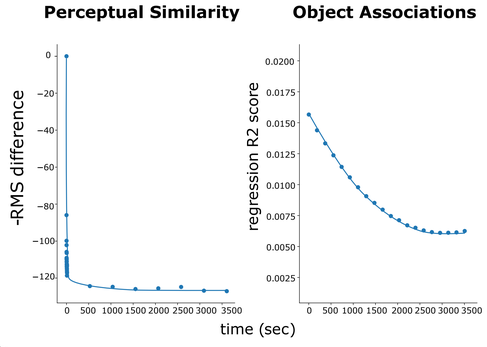
- **Contrastive Multiview Coding (CMC)** from Tian *et al.* (2019) was modified such that two views were two images separated by a specified **time lag**, encoded by the same AlexNet architecture.
- **Contrastive loss** was calculated in the latent space by selecting the positive pair of images from a distribution of negative pairs. Embeddings of frames at medium time scales would be learned to be represented as more semantically similar.
- SemanticCMC was performed as a fine tuning procedure on top of the published weights for CMC trained on an L vs. AB objective. The network was trained using our movie images, and its **representations were evaluated.**.

Evaluation

- The **semantic content of learned representations** was quantified using **representational similarity analysis** and without relying on improved object decoding accuracy for evaluation.
- We tested a variety of models and baselines against superordinate category and wordnet semantic models.

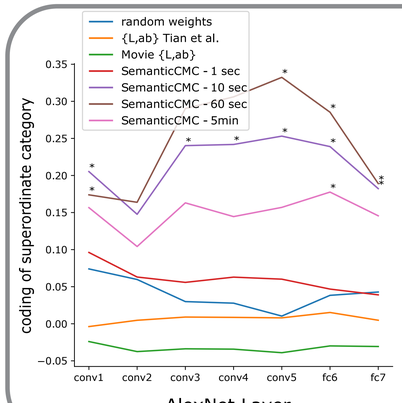
Naturalistic Dataset

- 158.4 hr of live action movies with naturalistic visual worlds.
- Automatically generated labels at sampling interval of 200 ms giving 2,851,272 sets of labels.
- Images taken every 1 sec, giving 572,949 images with preserved temporal relations.



- Association patterns of objects persist for approx. 50 min.
- Perceptual similarity drops rapidly within seconds.
- This presents as a leveragable signal for learning more semantic representations.

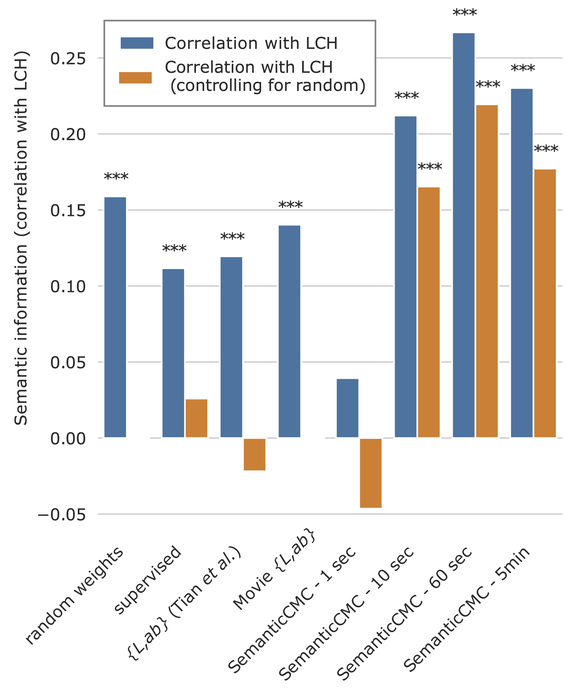
Learning superordinate category



- When trained on a contrastive loss objective across two images separated by 60 sec, SemanticCMC was most correlated to a superordinate category model.
- This coding of superordinate category was not found when the task was purely perceptual ({L,ab}) or when the interval between two images was either too short or too long.

Quantifying semantic content

- We tested whether pairs of classes that were more similar according to a language corpus-derived **LCH WordNet measure** were also more similar in their neural network activations patterns.
- Using a Mantel test with Pearson's correlation, semantic coding (LCH similarity) was quantified. A partial correlation that controlled for perceptual information (i.e. a random weights network) was then performed.
- Without controlling for perceptual features, all networks tested captured LCH semantic structure.
- When controlling for perceptual content**, only SemanticCMC trained over a sufficient distance in time captured LCH semantics.**



Conclusion

Temporal co-occurrence patterns of objects in a naturalistic dataset can provide a useful self-supervisory signal for building semantic representations. This may improve concept knowledge in artificial networks, and perhaps lead to computational models of infant learning.

References

Tian, Y., Krishnan, D., & Isola, P. (2019). Contrastive multiview coding. arXiv preprint arXiv:1906.05849.

erc
European Research Council
Established by the European Commission