

Structure: Exposition

- Experimental Evaluation: Bias and Generalization in Deep Generative Models**
Nicholas Spin
MLD, Carnegie Mellon University
- Density Estimation Background**
 - Input space \mathcal{X}
 - True distribution $p(\mathbf{X})$ on \mathcal{X}
 - Dataset of training images $\mathcal{D} = \{\mathbf{x}^1, \dots, \mathbf{x}^n\}$ (i.i.d. from $p(\mathbf{X})$)
 - Goal: Using \mathcal{D} , estimate $q(\mathbf{X})$ such that $\mathbf{x} \sim q(\mathbf{X})$
- Motivation: Inductive bias in DGMs is not understood**
 - \mathcal{D} is exponentially small compared to \mathcal{X} , so assumptions are required
 - These assumptions (inductive bias) in explicit and/or implicit ways
 - Authors propose to systematically analyze this bias
 - Original paper used subject expertise as in-bias
 - Authors look at simplified feature space inspired by psychology (size, angle, color, luminance)
- Method: Choose specific dataset ...**
 - Authors use different:
 - Algorithms (GAN, GMM)
 - Datasets (PIG, pig stimuli)
 - Distributions over features for $p(\mathbf{X})$ (i.e., distribution of color patches)
- Method: ... and identify feature space behavior of $q(\mathbf{x})$**
 - Addition of n distribution over features to $\mathcal{D} \rightarrow \mathcal{X}$ (size A)
 - Direct look at non-dimensional distribution (after one feature)
 - Compare support of $p(\mathbf{X})$ and $q(\mathbf{X})$
 - Visualize 2D distribution for single combination

- General problem
- Purpose / stated goal(s)
- Experimental setup summary
- Result summary

- Results: DGMs generalize locally**
 - If single feature distribution centered around that feature with variance
 - If multiple separate modes, then distributions do average over these
 - If modes are near each other, create peak in mean (prototype enhancement)
 - Across multiple features, behavior is independent
- Results: Support of $q(\mathbf{x})$ increases faster than $p(\mathbf{x})$**
 - Active distributions are added by averaging them
 - These combinations are still increasingly generated
 - Number of unique, novel, and biological increases
 - Authors conclude: Generally faster to memorize >100 combinations
- Results: DGMs memorize when there are few modes**
 - If few combinations, then all activation combinations
 - If many combinations, then generalization outside of \mathcal{D} is rapid
- Some authors are from Stanford Dept. of Psych.**
 - Authors used 5, 10, and 20 with GANs and from Department of Psychology
 - Other two are from the Cognitive Science Department
 - Authors first verified that the **prototype enhancement effect** in psychology (the intermediate point between two color patches is correctly represented)
 - Authors first **reproduction** about how models need **generalization** when it may

Experimental Evaluation: Bias and Generalization in Deep Generative Models

Nicholay Topin
MLD, Carnegie Mellon University

*(NeurIPS 2018 paper from Stanford)

Density Estimation Background

- Input space \mathcal{X}
- True distribution $p(\mathbf{x})$ on \mathcal{X}
- Dataset of training points $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ (i.i.d. from $p(\mathbf{x})$)
- Goal: Using \mathcal{D} , calculate $q(\mathbf{x})$ over \mathcal{X} so it is close to $p(\mathbf{x})$

Example Algorithms:

- Generative Adversarial Networks (GANs)
- Variational Autoencoders (VAEs)

Motivation: Inductive bias in DEs is not understood

- \mathcal{D} is exponentially small compared to \mathcal{X} , so assumptions are required
- These assumptions (inductive bias) is implicit and not understood well
- Authors propose to systematically analyze this bias

- Original input and output spaces are too large (focus on images)
- Authors look at simplified feature space inspired by psychology (size, shape, color, numerosity)

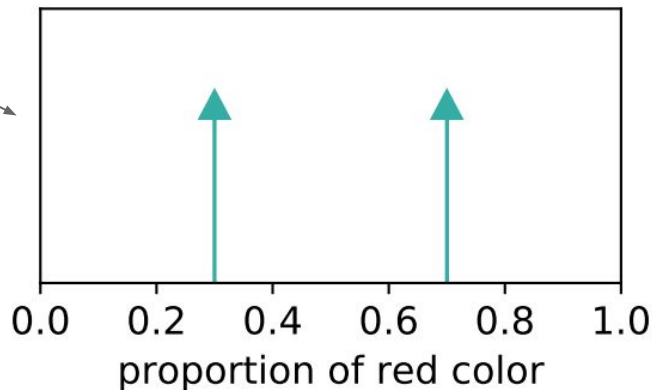
Method: Choose specific dataset ...

Authors use different:

- Algorithms
(VAE, GAN)

- Datasets
(e.g., pie charts)

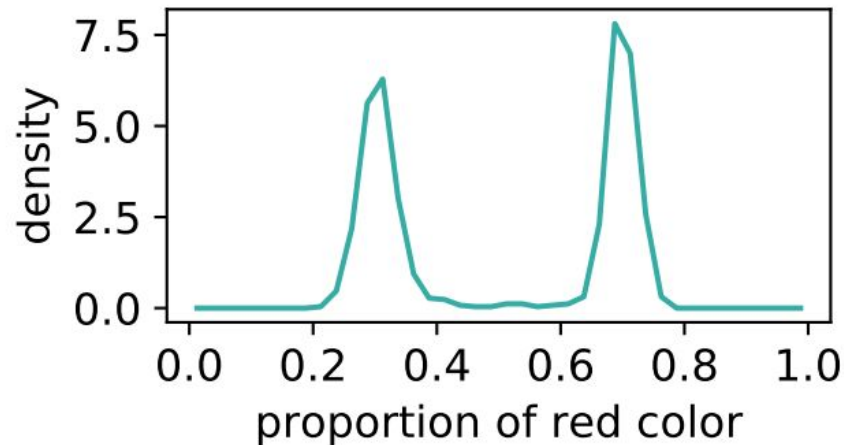
- Distributions over features for $p(\mathbf{x})$
(e.g., distribution of color portion)



Method: ... and identify feature space behavior of $q(\mathbf{x})$

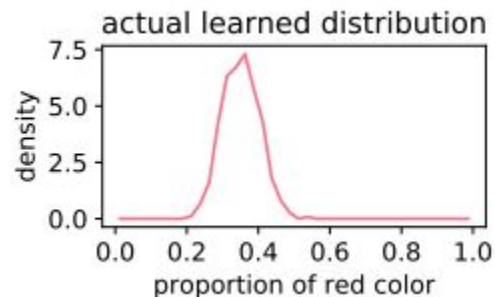
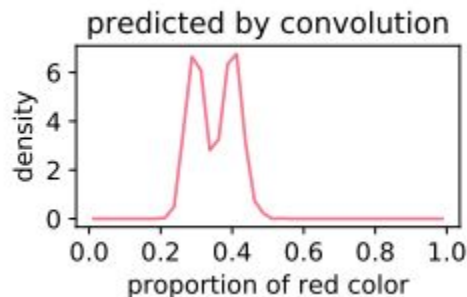
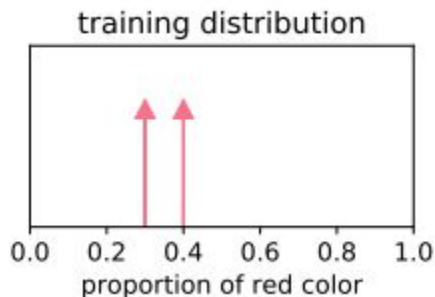
Authors look at distribution over features for $q(\mathbf{x})$ over \mathcal{X}

- Directly look at one-dimensional distribution (when one feature)
- Compare support of $p(\mathbf{x})$ and $q(\mathbf{x})$
- Visualize 2D distribution for single combination



Results: DEs generalize locally

- If single mode, distribution centered around mode but with variance
- If multiple separate modes, then distribution is average over these
- If modes are near each other, create peak at mean (“**prototype enhancement**”)
- Across multiple features, behavior is independent

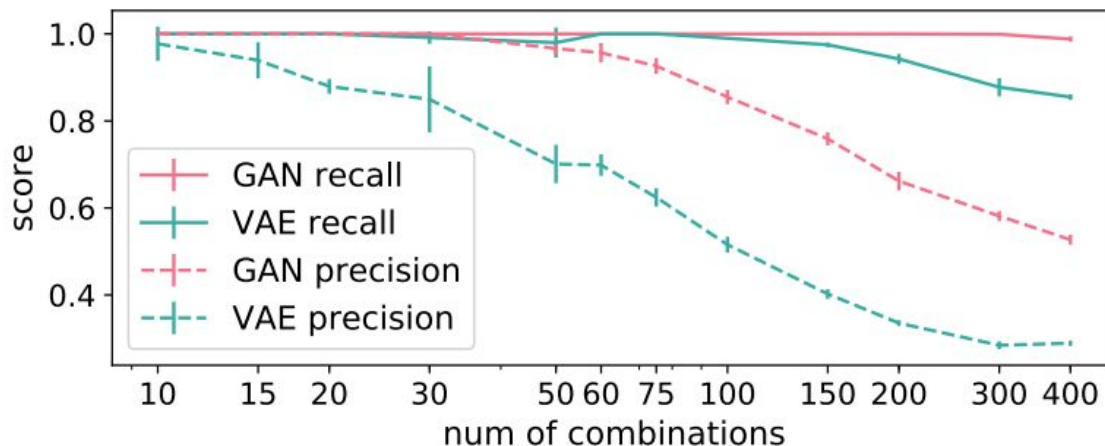


Results: Support of $q(x)$ increases faster than $p(x)$

As more combinations are added to training data:

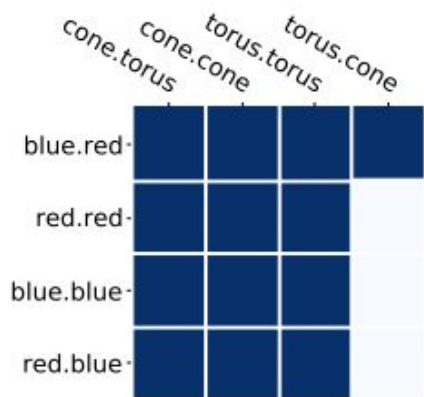
- These combinations are still consistently generated
- Number of unique, novel combinations increases

Authors conclude: Generally hard to memorize >100 combinations

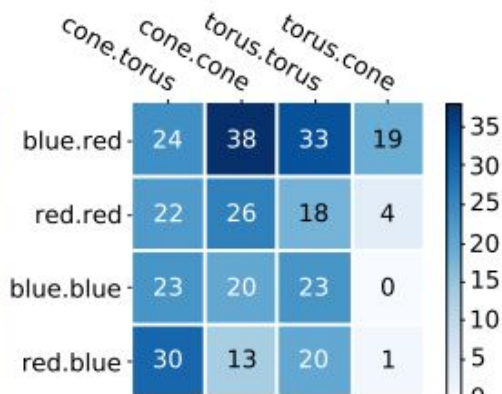


Results: DEs memorize when there are few modes

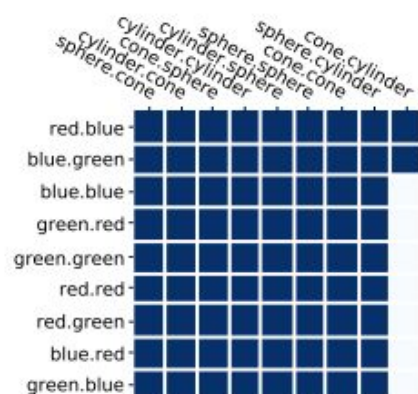
- If few combinations, then will memorize combinations
- If many combinations, then generalizes outside of $p(\mathbf{x})$ support



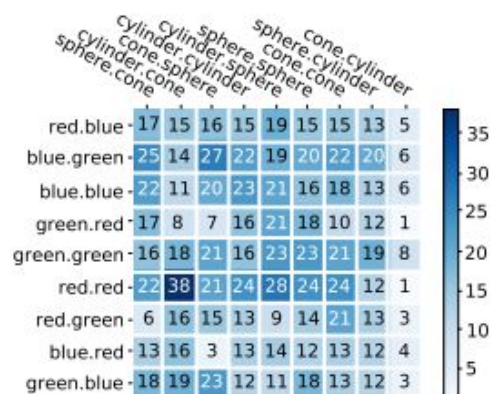
A: Training Distribution (4x4)



B: Generated Combinations (4x4)



C: Training Distribution (9x9)



D: Generated Combinations (9x9)

Some authors are from Stanford Dept. of Psych.

- Authors 3 and 5 (Yuan and Goodman) are from Department of Psychology
- Other four are from the Computer Science Department

- Authors find similarities to the **prototype enhancement effect** in psychology (the intermediate point between two close modes is strongly expressed)
- Authors find **memorization** when few modes and **generalization** when many

Structure: Critique

10

Critique: Do not explain psychology terms

- Prototype Abstraction: Learning a canonical representation for a category (membership of new items based on similarity to prototype)
- Exemplar Memorization: Learning a set of exemplars for a category (membership of new items based on similarity to all these exemplars)

11

Critique: Overlooked a related work

Related work in single on address:

- "Classification of Prototype Abstraction and Exemplar Memorization" (2015)

DINAM authors:

- Compare if Adv. and Lo. Mem. in **subdomains**
- Clearly effect of Adv. and Lo. Mem. (identifying critical work)
- Effect of Adv. effect only in training and Lo. Mem. effect in testing
- Effect of Adv. effect disappears when cognitive process well occurred
- Compare models with psychological studies and find their results (and psychological hypotheses in their system)

12

Critique: Psychology comparison was haphazard

DGM paper authors:

- Do not explain prototype abstraction or exemplar memorization
- Do not quantify IR effect (quantify generalization and memorization in a non-standard way)
- Do not look at behavior over course of training (only report for rest of training of Adv. regarding memorization conditions)
- Do not consider effect of category size (only consider case where 4 items are chosen as standard)
- Do not use hypothesis about IR relationship
- Do not compare in existing work in neural network IR

13

Critique: Try few hyperparameter settings

- Authors claim conclusions that fit different hyperparameters
- Appendix explains the authors only use one set per method (but not)


14

Critique: Broad generalization in memorization conclusion

Authors:

- Only consider random selection of probes
- Show generalization (broad support)
- Use the authors too controlling memorization is very difficult

Experiment set up to encourage generalization (see other slide to encourage memorization)




- Are appropriate baseline methods considered?
- Are appropriate evaluation metrics used?
- Is experiment design reasonable?
- Is uncertainty of data-driven approach accounted for?
- Are the results reproducible?
- Are conclusions corroborated by results?
- Are stated goals achieved?

15

Critique: Overgard factors aside from mode count


- Aim to test "when and how learning in mode generation novel situations"
- Conclude behavior as a function of number of modes (1-25 modes randomized and 4 trials based on generalization)



16

Critique: Overgard factors aside from mode count


- Aim to test "when and how learning in mode generation novel situations"
- Conclude behavior as a function of number of modes (1-25 modes randomized and 4 trials based on generalization)
- Acknowledge domain shift (only only quantify in support increases, but only use a factor of 4 for behavior increase and measure (broad support on Adv. case may lead to use in generalization behavior)
- Try to understand amount of time which may not depend on domain (time varying or not control may affect generalization behavior)



17

Critique: Leave unanswered questions


- In introduction, mention finding number of probes in training class before new conditions are generated, but do not do the analysis
- Do not address reproducibility across figures (see Figure 16) (Why are the results desirable or different?)



18

Critique: Leave unanswered questions

- In introduction, mention finding number of probes in training class before new test conditions are generated, but do not do the analysis
- Do not address reproducibility across figures (see Figure 16) (Why are the results desirable or different?)
- Clear results are the same for Adv., but the plots show something else



Critique: Do not explain psychology terms

- Prototype Abstraction: Learning a canonical representation for a category (membership of new items based on similarity to prototype)
- Exemplar Memorization: Learning a set of examples for a category (membership of new items based on similarity to all these examples)

Critique: Overlooked a related work

Related work in cognitive science:

“Development of Prototype Abstraction and Exemplar Memorization” (2010)

DPAEM authors:

- Consider P. Abs. and Ex. Mem. in **autoencoders**
- Quantify effect of P. Abs. and Ex. Mem. (following previous work)
- Find P. Abs. effect early in training and Ex. Mem. effect later in training
- Find P. Abs. effect diminished when categories are less well structured
- Compare results with psychological studies and find close match
(test psychological hypotheses in their system)

Critique: Psychology comparison was haphazard

DGM paper authors:

- Do not explain prototype abstraction or prototype enhancement
- Do not quantify PA effect
(quantify generalization and memorization in a non-standard way)
- Do not look at behavior over course of training
(only report for end of training without specifying termination condition)
- Do not consider effect of category structure
(only consider case where modes are chosen at random)
- Do not test hypotheses about PA relationship
- Do not compare to existing work in neural network PA

Critique: Try few hyperparameter settings

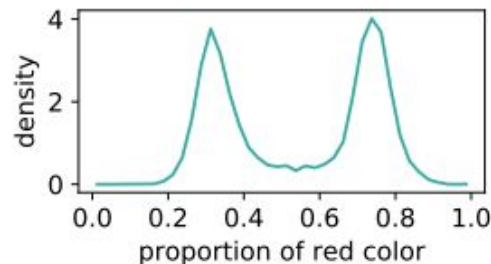
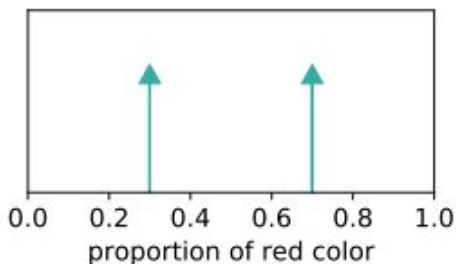
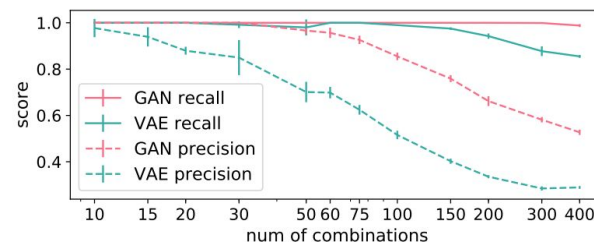
- Authors claim conclusions hold for different hyperparameters
- Appendix explains that authors only test one set per method (four total)

Critique: Broad generalization in memorization conclusion

Authors:

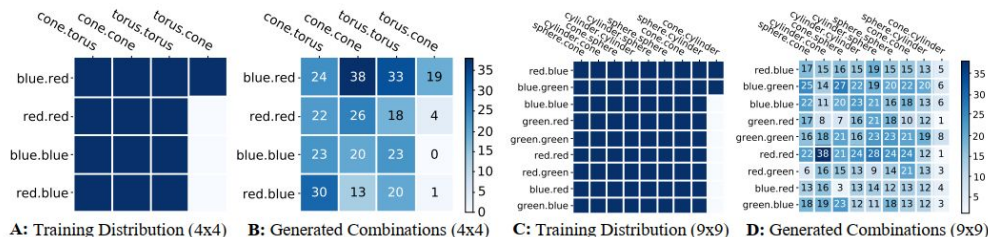
- Only consider random selection of modes
- Show generalization (increased support)
- Use as evidence that controlling memorization is very difficult

Experiment set up to encourage generalization
(no effort made to encourage memorization)



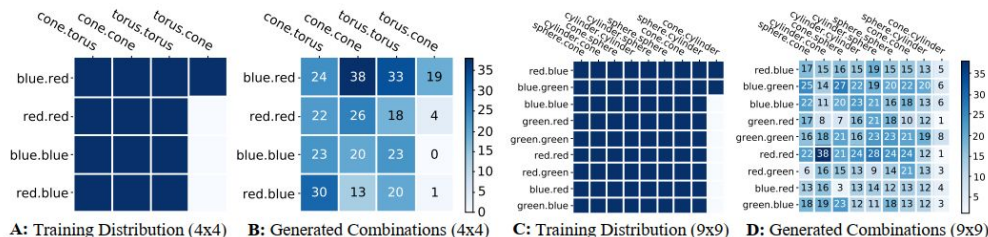
Critique: Disregard factors aside from mode count

- Aim to find “when and how existing models generate novel attributes”
- Conclude behavior is a function of number of modes
(< 20 modes memorized and > 80 modes lead to generalization)



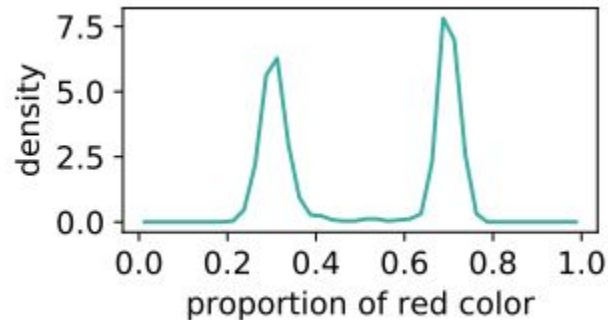
Critique: Disregard factors aside from mode count

- Aim to find “when and how existing models generate novel attributes”
- Conclude behavior is a function of number of modes
(< 20 modes memorized and > 80 modes lead to generalization)
- Acknowledge dataset must grow very quickly as support increases, but only use a factor of four between minimum and maximum
(fewer samples in 4x4 case may lead to same generalization behavior)
- Train for indeterminate amount of time which may not depend on dataset
(less training in 4x4 case may lead to same generalization behavior)



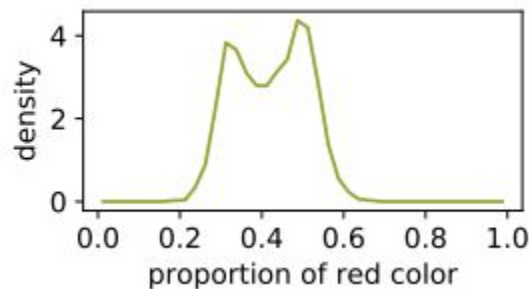
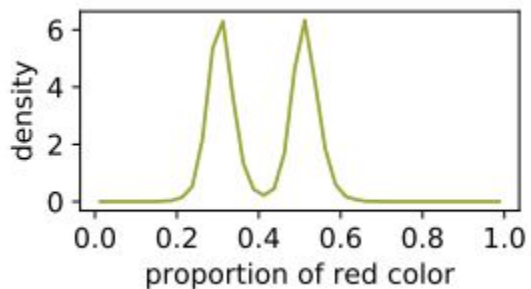
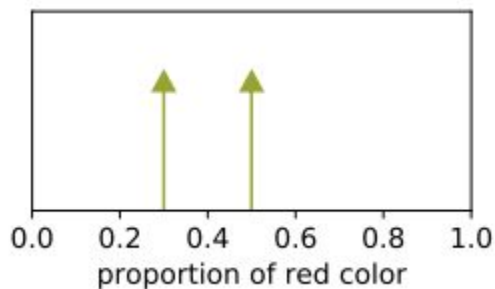
Critique: Leave unanswered questions

- In introduction, mention finding number of colors in training data before new combinations are generated, but do not do this analysis
- Do not address asymmetry in some figures (ex: Figure 10)
(Why are the mode densities so different?)



Critique: Leave unanswered questions

- In introduction, mention finding number of colors in training data before new combinations are generated, but do not do this analysis
- Do not address asymmetry in some figures (ex: Figure 10)
(Why are the mode densities so different?)
- Claim results are the same for VAE, but the plots show smoother trend



Critique: Psychology comparison was haphazard

DGM paper authors:

- Do not explain prototype abstraction or prototype enhancement
- Do not quantify PA effect
- Do not look at behavior over course of training
(only report for end of training without specifying termination condition)
- Do not consider effect of category structure
(only consider case where modes are chosen at random)
- Do not test hypotheses about PA relationship
- Do not compare to existing work in neural network PA

1

Experimental Evaluation: Bias and Generalization in Deep Generative Models

Nicholas Topin
MLD, Carnegie Mellon University

Source: [1] Experimental Evaluation

2

Density Estimation Background

- Input space: \mathcal{X}
- True distribution $p(\mathbf{X})$ on \mathcal{X}
- Dataset of training images $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ (i.i.d. from $p(\mathbf{X})$)
- Goal: learn \hat{p} (estimator) over \mathcal{D} , so it is close to $p(\mathbf{X})$

- Example Algorithms:
- Generative Adversarial Networks (GANs)
 - Variational Autoencoders (VAEs)

3

Motivation: Inductive bias in DGMs is not understood

- \mathcal{D} is exponentially small compared to \mathcal{X} , so assumptions are required
- These assumptions (inductive bias) is implicit, and not understood well
- Authors propose to systematically analyze this bias

- Original input and output spaces are too large
- Authors look at simplified feature space inspired by psychology (size, shape, color, numerosity)

IC 4

Method: Choose specific

- Authors use different:
- Algorithms (VAE, GAN)
 - Dataset (e.g., MNIST)
 - Distributions over features (e.g., distribution of color)

7

Ideally: Support of q

- As more combinations are added:
- These combinations are:
 - Number of unique, novel
- Authors conclude: Generally he



10

Critique: Do not

- Prototype Abstraction: (membership of no)
- Example: Identifying (membership of no)

13

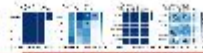
Critique: In

- Authors claim 0
- Appendix explains

16

Critique: Unregard factors aside from mode count

- Aim to find "when and how missing mode/generative novel distribution"
- Conclude behavior is a function of number of modes (≈ 20 modes re-emerged and ≈ 60 modes lead to generalization)
- Knowledge doesn't must grow very quickly as support increases, but only up a factor of four between minimum and maximum (their samples in 4-d case may lead to some generalization behavior)
- Look for intermediate amounts of bias which may not depend on dataset (bias existing in 4-d case may lead to some generalization behavior)



5

Method: ... and identify it

- Authors look at distribution over for:
- Clarity: look at one-dimensional
 - Compare support of $p(\mathbf{X})$
 - Visualize 2D distribution for a



8

Ideally: Use memory

- If few combinations, then will
- If many combinations, then ge



11

Critique: Overlap

- Related work in cognitive "Development of P" DGM authors:
- Consider P. Abs. i
 - Quantify effect of
 - Find P. Abs. effect
 - Find P. Abs. effect
 - Compare results (just psychology)

14

Critique: Bi

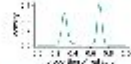
- Authors:
- Only cross
 - Show gaps
 - Use as out

Experiment as a (no effort)

17

Critique: Leave unanswered questions

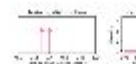
- In introduction, mention finding number of colors in training data before new combinations are generated, but do not do this analysis
- Do not address asymmetry in same figure (see Figure 18) (Why are the mode decisions so different?)



6

Ideally: Use generalization

- If single mode, distribution common
- If multiple separate modes, then c
- If modes are near each other, one
- Across multiple features, behavior



9

Some authors are fro

- Authors 3 and 5 (Huan and Gu)
- Other four are from the Comp

- Authors find similar links to the (the intermediate points bet
- Authors find generalization of

12

Critique: Psycho

- DGM paper authors:
- Do not explain pe
 - Do not quantify PE (quantity of generalization)
 - Do not look at test (only report for
 - Do not consider if (only consider
 - Do not test hypothesis
 - Do not compare p

15

Critique: De

- Aim to find "wh
- Conclude fairly (in 2D mode)

18

Critique: Leave unanswered questions

- In introduction, mention finding number of colors in training data before new combinations are generated, but do not do this analysis
- Do not address asymmetry in same figure (see Figure 18) (Why are the mode decisions so different?)
- Claim results are the same for VAE, but the plots show another trend

