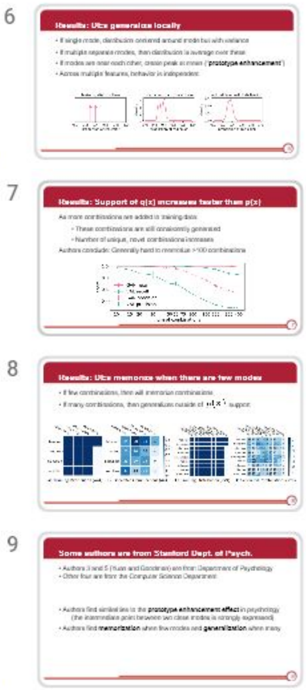
Structure: Exposition

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



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 $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ (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

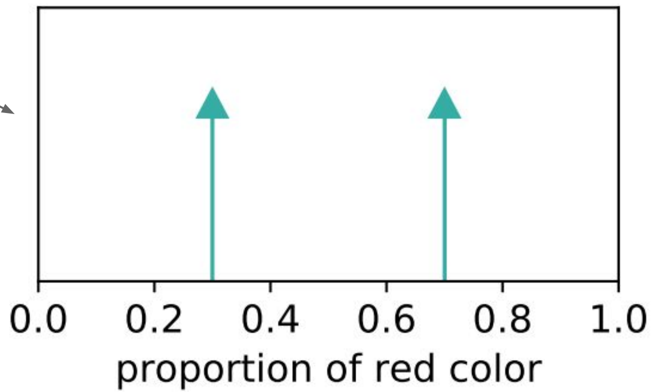
- $\mathcal{D}$ is exponentially small compared to $\mathcal{X}$, so assumptions are required
- These assumptions (inductive bias) is implicit and not understood well
- Authors propose to systematically analyze this bias

- 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:

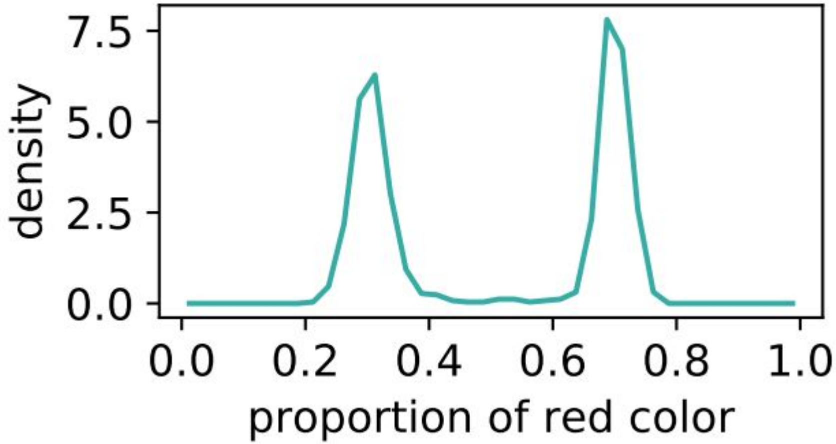
- Algorithms
  (VAE, GAN)
- Datasets
  (e.g., pie charts)
- Distributions over features for $p(\mathbf{x})$
  (e.g., distribution of color portion)



Method: … and identify feature space behavior of q(x)

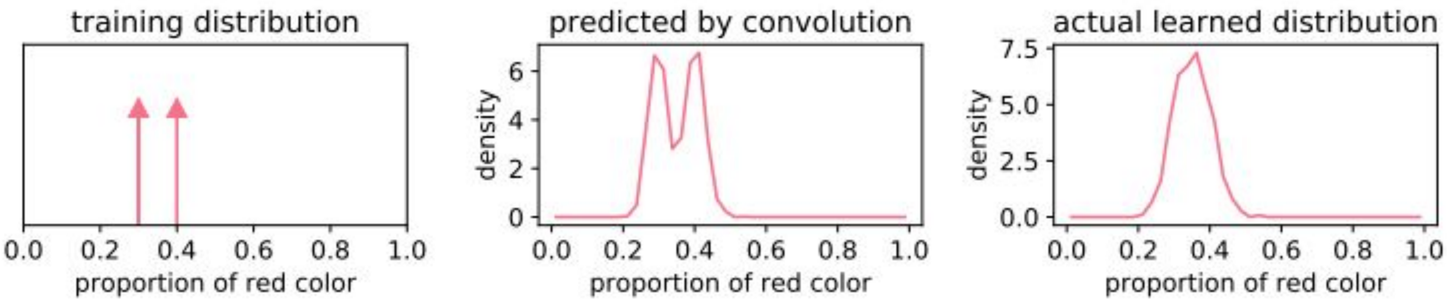
Authors look at distribution over features for $q(\mathbf{x})$ over $\mathcal{X}$

- 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

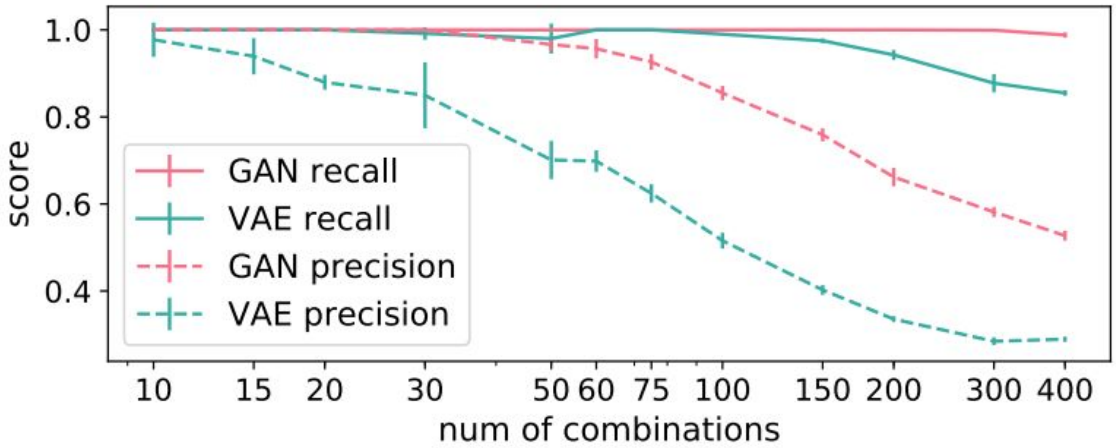


Results: Support of q(x) increases faster than p(x)

As more combinations are added to training data:

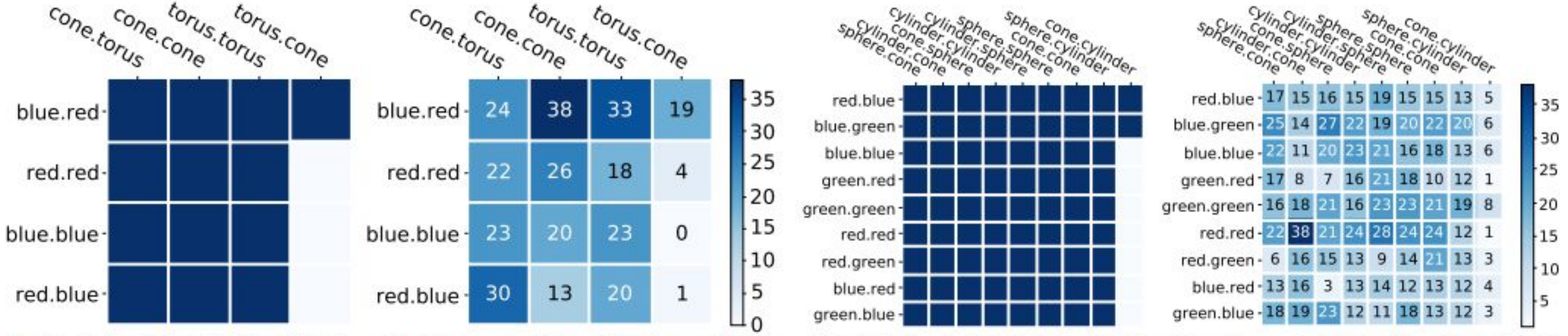
- These combinations are still consistently generated
- Number of unique, novel combinations increases

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.

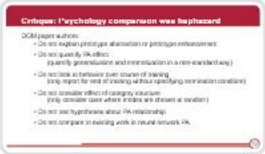
- Authors 3 and 5 (Yuan and Goodman) are from Department of Psychology
- Other four are from the Computer Science Department

- 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

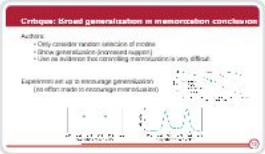
Structure: Critique

10

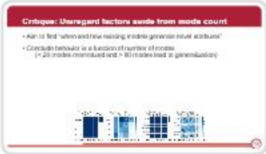

11

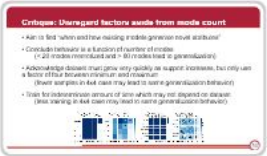

12


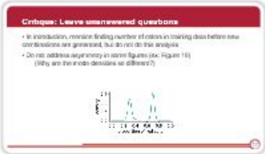
13

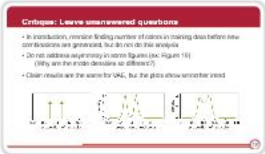

14


- 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


16


17


18


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

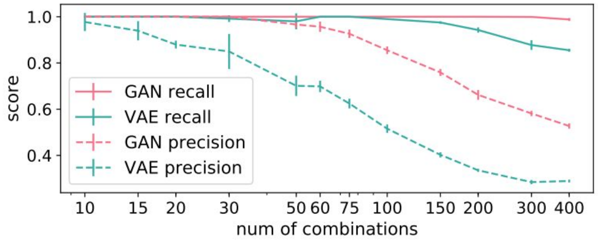
Critique: Try few hyperparameter settings

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

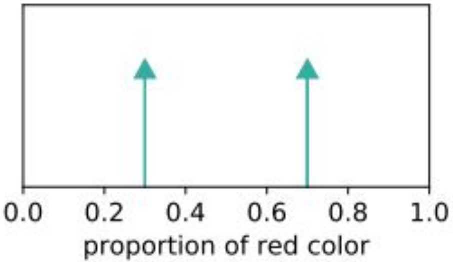
Critique: Broad generalization in memorization conclusion

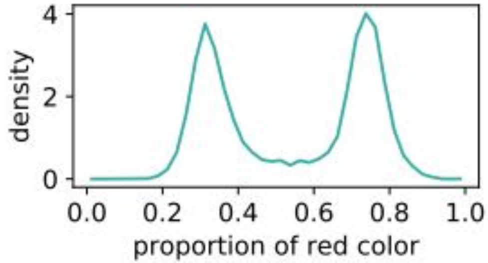
Authors:

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



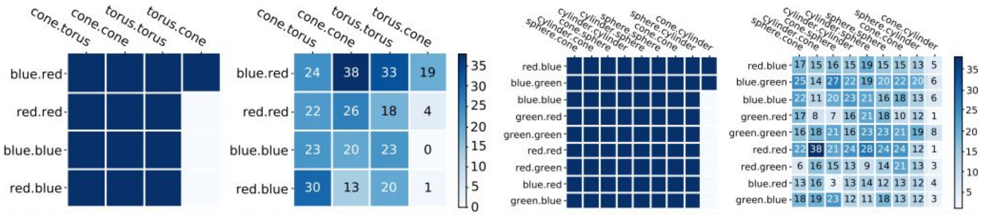
Experiment set up to encourage generalization
(no effort made to encourage memorization)





Critique: Disregard factors aside from mode count

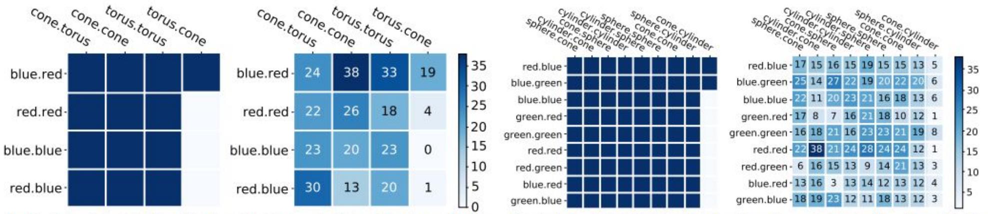
- Aim to find “when and how existing models generate novel attributes”
- Conclude behavior is a function of number of modes
  (< 20 modes memorized and > 80 modes lead to generalization)



A: Training Distribution (4x4) **B:** Generated Combinations (4x4) **C:** Training Distribution (9x9) **D:** Generated Combinations (9x9)

Critique: Disregard factors aside from mode count

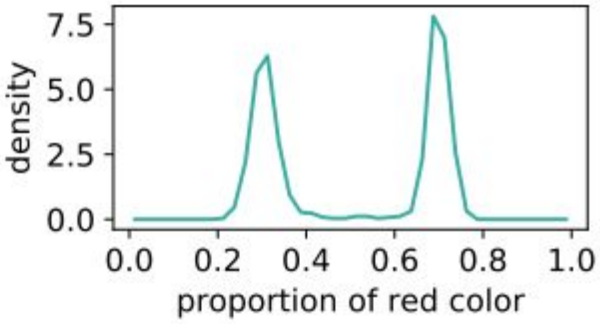
- Aim to find “when and how existing models generate novel attributes”
- Conclude behavior is a function of number of modes
  (< 20 modes memorized and > 80 modes lead to generalization)
- Acknowledge dataset must grow very quickly as support increases, but only use a factor of four between minimum and 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)



A: Training Distribution (4x4) **B:** Generated Combinations (4x4) **C:** Training Distribution (9x9) **D:** Generated Combinations (9x9)

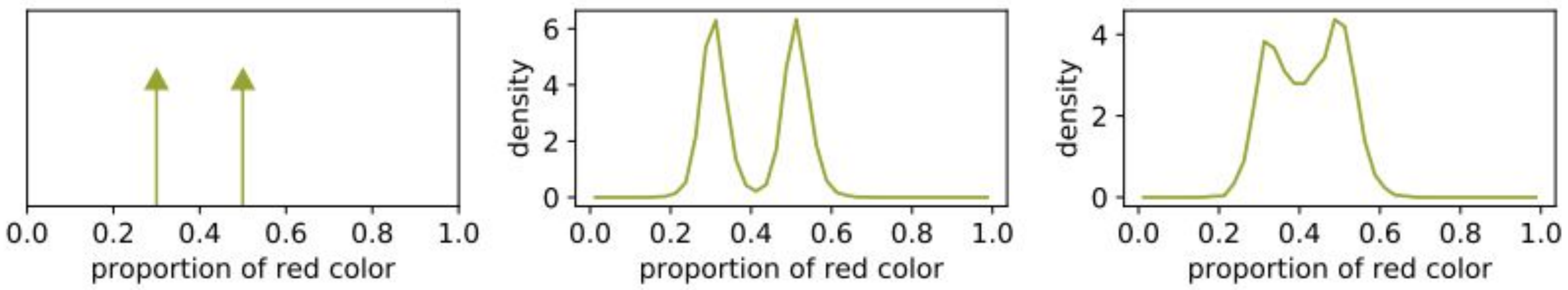
Critique: Leave unanswered questions

- 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

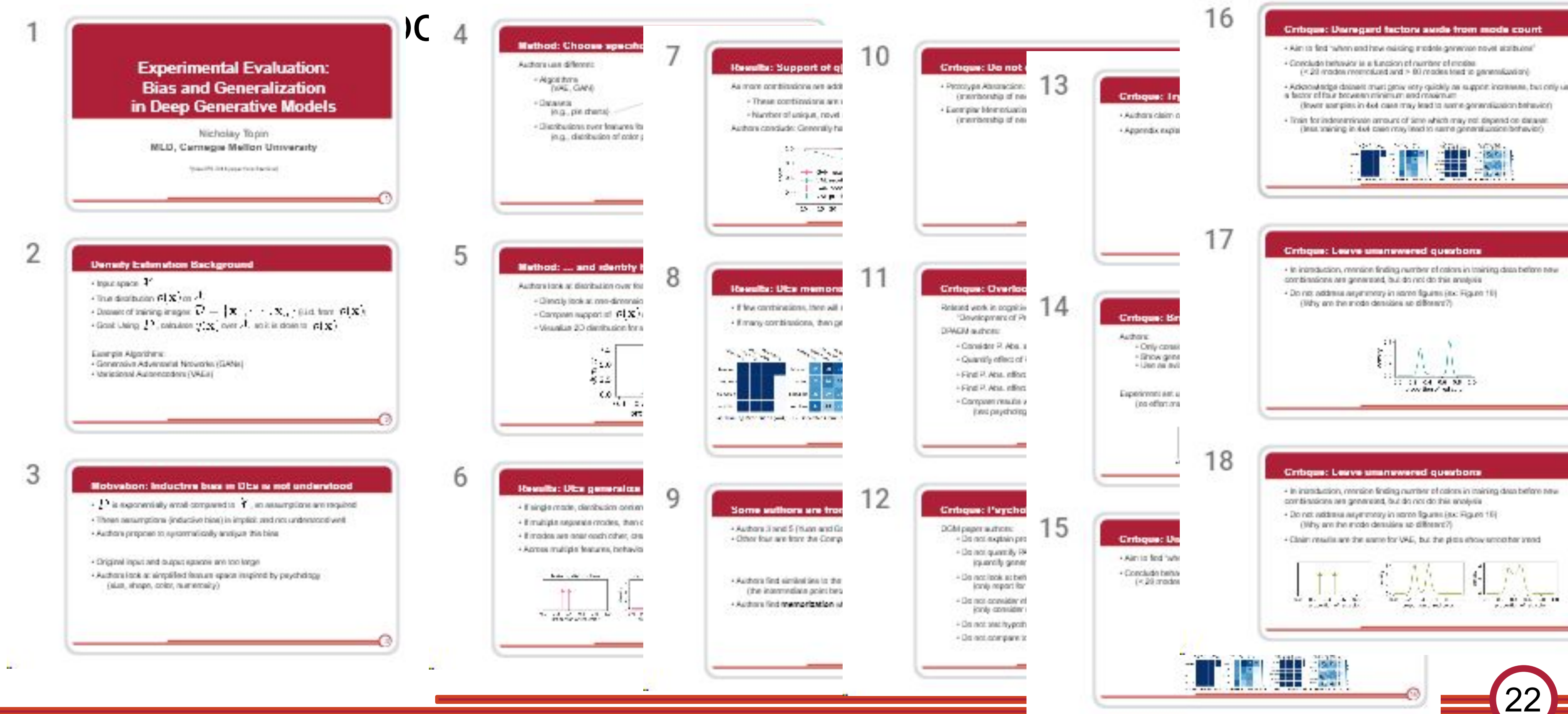
- In introduction, mention finding number of colors in training data before new combinations are generated, but do not do this analysis
- Do not address asymmetry in some figures (ex: Figure 10)
  (Why are the mode densities so different?)
- Claim results are the same for VAE, but the plots show smoother 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
Nicholay Topin
MLD, Carnegie Mellon University
2
Density Estimation Background
3
Motivation: Inductive bias in DGMs is not understood
)C
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18