Emergence of Hidden Capabilities: Exploring Learning Dynamics in Concept Space

Park et al. NeurIPS 2024 Spotlight

Thomas Melistas
University of Athens

CV & Robotics reading group
10 February 2025

Learning Dynamics of Generative Models

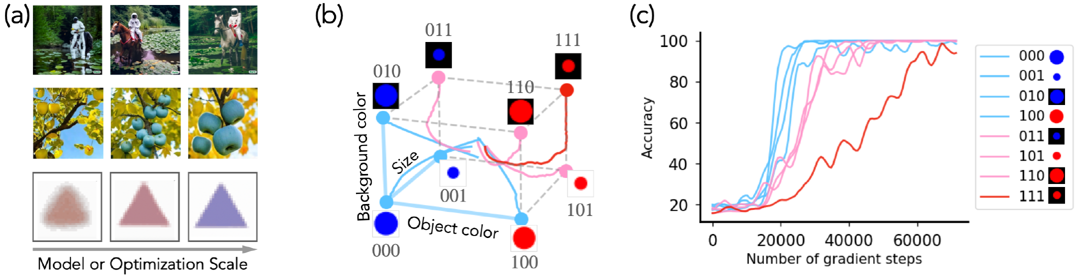
Generative models are able to **generalize** out-of-distribution (OOD) and combine **concepts** in novel ways, not seen during training, by:

- internalizing data-generating process
- disentangling concepts (latent factors of variation) underlying it

Q: What determines whether the model will disentangle a concept and learn to manipulate it? Are all concepts learned at the same time?

Problem Setting

Class of interest: A generative model F, trained using conditioning information h to produce images y



Contributions

- Introduce *Concept Space* to analyze a model's learning

ARCHIMEDES

Contributions

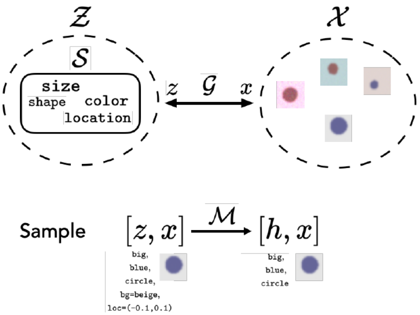
- Introduce *Concept Space* to analyze a model's learning
- Show that *Concept Signal* dictates the order of concept learning

Contributions

- Introduce *Concept Space* to analyze a model's learning
- Show that *Concept Signal* dictates the order of concept learning
- Learning of concepts happens in two phases:
  - (P1) learning of a hidden capability
  - (P2) learning to generate the desired output from the input space

Concept Space

Definition 1. ***(Concept Space.)*** *Consider an invertible data-generating process $\mathcal{G} : \mathcal{Z} \to \mathcal{X}$ that samples vectors $z \sim P(\mathcal{Z})$ from a vector space $\mathcal{Z} \subset \mathbb{R}^d$ and maps them to the observation space $\mathcal{X} \in \mathbb{R}^n$. We assume the sampling prior is factorizable, i.e., $P(z \in \mathcal{Z}) = \Pi_{i=1}^d P(z_i)$, and individual dimensions of $\mathcal{Z}$ correspond to semantically meaningful concepts. Then, a concept space $\mathcal{S}$ is defined as the multidimensional space composed of all possible concept vectors z, i.e., $\mathcal{S} := \{z \mid z \sim P(\mathcal{Z})\}$*



On Capability

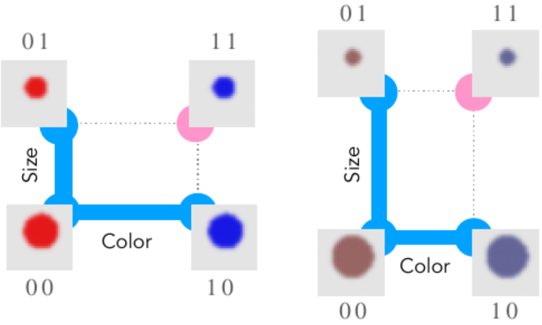
Definition 2. ***(Capability.)*** *A concept class $\mathcal{C}$ denotes the set of concept vectors $z_{\mathcal{C}}$ such that a subset of dimensions of these vectors are fixed to predefined values. Classes $\mathcal{C}$ and $\mathcal{C}'$ are said to differ in the k^{th} concept if $\forall z \in z_{\mathcal{C}}$, there exists $z' \in z_{\mathcal{C}'}$ with $z[k] \neq z'[k]$ and $z[i] = z'[i]$ for $i \neq k$. We say a model possesses the "capability to alter the k^{th} concept" if for any class $\mathcal{C}$ whose samples were seen during training, the model can produce samples from class $\mathcal{C}'$ that differs from $\mathcal{C}$ in the k^{th} concept.*

- we do not need to use the conditioning h used for training
- other techniques can be used (over-prompting, latent interventions)

Concept Signal

Definition 3. *(Concept Signal.) The concept signal σ_i for a concept z_i measures the sensitivity of the data-generating process to change in the value of a concept variable, i.e., $\sigma_i := \left|\partial\mathcal{G}(z)/\partial z_i\right|$.*

Intuitively, concept signal indicates how much the model would benefit from learning a concept



ARCHIMEDES

Experimental Setup

Models:

- Variational Diffusion Model [1]
- Generate 3×32×32 (& 3×64×64) images conditioned on h

Datasets:

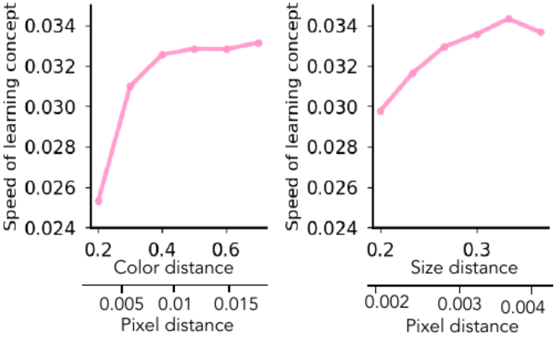
- Synthetic toy 2D objects with controlled concepts
- CelebA

Evaluation:

- Classifier probes for individual concepts (U-Net)
- Using same training set

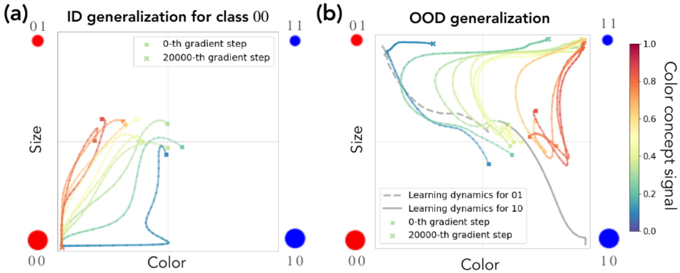
Concept Signal Determines Learning Speed

- Changing the level of concept signal in the training data
- $h := z$
- speed of learning: inverse of the number of gradient steps required to reach 80% accuracy on OOD class



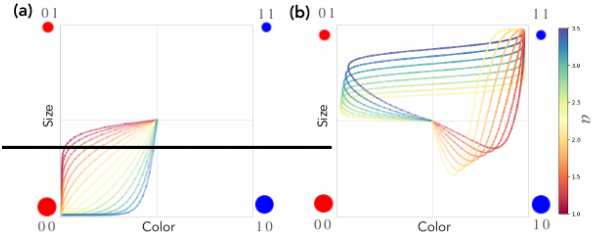
Concept Signal Governs Generalization Dynamics

- Concept memorization: OOD generations biased towards class with strongest concept signal
- Problem when early stopping text-to-image models
- Unseen conditioning associated to nearest concept class



Landscape Theory of Learning Dynamics

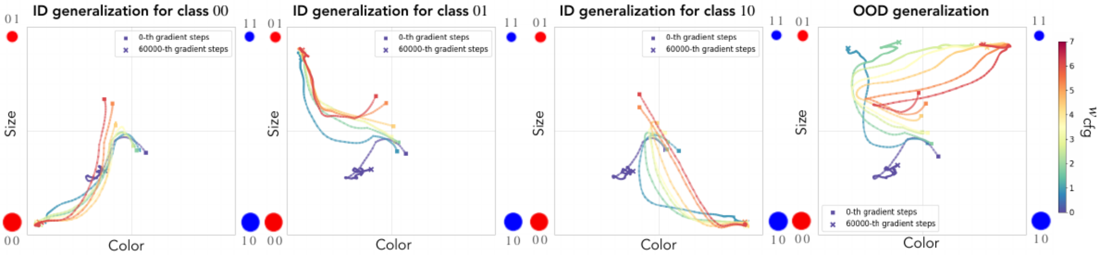
- There is a sudden turn from concept memorization to OOD generalization
- Learning dynamics can be decomposed into two stages
- *Hypothesis:* there is a phase change, in which the model learns to alter concepts



Sudden Transitions in Learning Dynamics

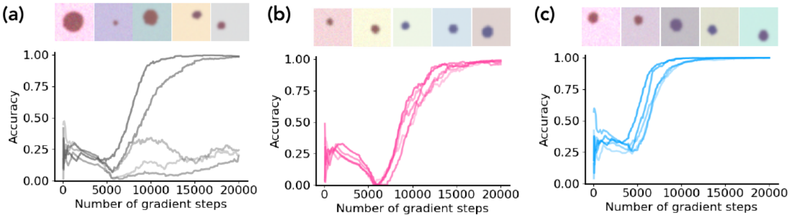
- There is a phase in which the model is capable of disentangling concepts, but still produces incorrect images
- Naive input prompting is insufficient to elicit these capabilities and generate samples from OOD classes
- Second phase in learning dynamics: an alignment between the input space and concept representations is learned

Different Guidance Scales



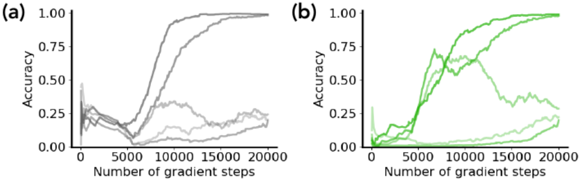
Techniques to elicit hidden capabilities

1. **Activation Space:** *Linear Latent Intervention*
2. **Input Space:** *Overprompting*



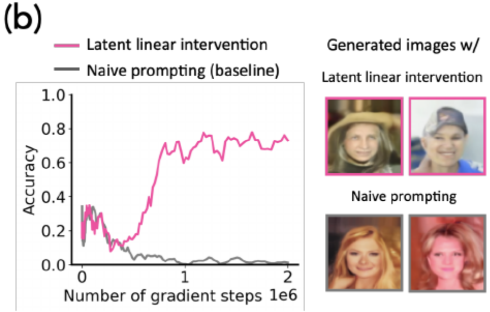
Patching Embeddings

1. Take the embedding module from final checkpoint
2. Patch it to an intermediate U-Net checkpoint
3. Naive prompting works as well as previous techniques



- *Second phase aligns input space to intermediate representations*
- *Embedding module disentangles concepts*
- *U-Net generates a representation for each*

Results on CelebA



Effect of Underspecification

In the previous experiments $h := z$, *what if not?*

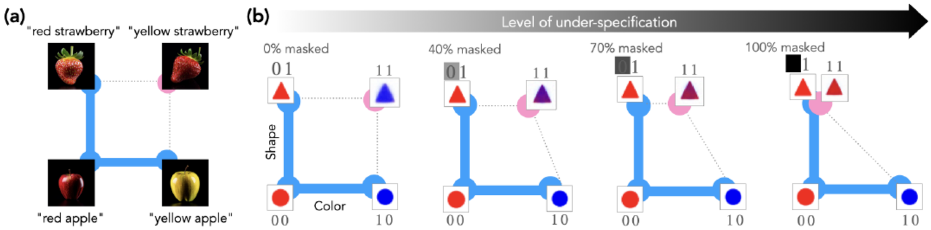
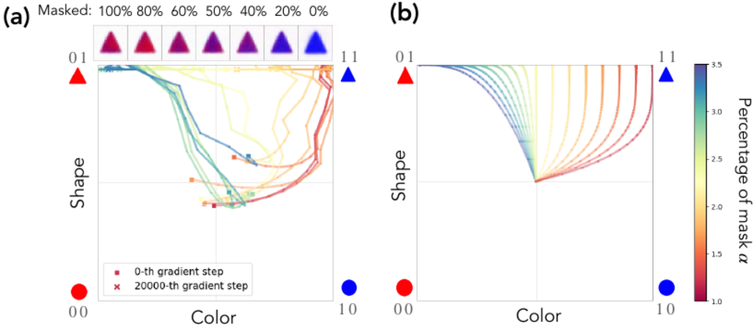


Figure: Images of a strawberry are often correlated with the color red

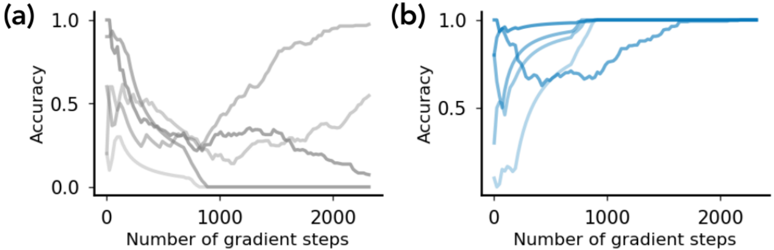
- Simulate underspecification by randomly masking (e.g. ~~red~~ triangle)

Underspecification hinders OOD generalization



When prompts are masked, the model's understanding of shape triangle becomes intertwined with color red, **even when blue is specified**

Overprompting and Underspecification



Capability can develop prior to observable behavior, even in cases of underspecification.

Conclusion

- Concept Space may be useful to understand learning in generative models
- Concept Signal Dictates Speed of Learning
- Generative models learn to manipulate concepts earlier than exhibited

Limitations:

- Real-world data are more complex (not always compositional)
- Concepts are not always linearly embedded in the vector space $\mathscr{Z}$

References

[1] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, NeurIPS 2021