Conjuring Semantic Similarity

Tian Yu Liu
tianyu@cs.ucla.edu

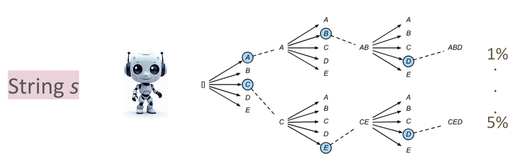
Stefano Soatto
soatto@cs.ucla.edu

Paper Website

ICLR

Abstract & Motivation

Distributional Semantics typically defines meanings of textual expressions via linguistic usage. For instance, the "meaning" of textual prompts can be represented in LLMs as a distribution over their future continuations:

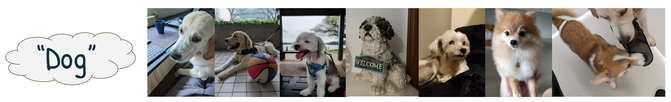


Liu et al., Meaning Representations from Trajectories in Autoregressive Models. ICLR 2024.

Here, we formulate a mathematically rigorous alternative: an expanded notion of meaning grounded in generative visual processes. In particular, we define meaning representation based on the distribution over images that a prompt generates, or "conjures".

Construction for Diffusion Models

Unlike for humans, generative models allow us to easily visualize and compare generated images, or their distribution, evoked by a textual prompt.



We show how our notion of meaning can be computationally realized through the class of diffusion models.

A (forward) diffusion process $\{\boldsymbol{x}(t)\}_{t=0}^{T}$ can be modeled as the solution to the following SDE:

$$d\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, t)dt + g(t)d\boldsymbol{w}_t$$

with drift coefficient *f* and (scalar) diffusion coefficient *g*, where *w* is standard Brownian motion.

Conditioning on a textual prompt *y*, a trained diffusion model $s_\theta(\boldsymbol{x}, t|y)$ can be used in the reverse-time SDE (the *generative process*) given by

$$d\boldsymbol{x} = [f(\boldsymbol{x}, t) - g(t)^2 s_\theta(\boldsymbol{x}, t|y)]dt + g(t)d\bar{\boldsymbol{w}}_t$$

where $\bar{\boldsymbol{w}}_t$ is the Brownian motion running backwards in time.

Conjuring Semantic Similarity

We propose a novel approach whereby the semantic similarity among textual expressions is based not on other expressions they can be rephrased as, but rather based on the imagery they evoke.

Core Proposition: Semantic similarity is the Jeffreys divergence between the reverse-time diffusion stochastic differential equations (SDEs) conjured by textual prompts, computable directly via Monte-Carlo sampling.

In particular, two different textual prompts y_1 and y_2 yields two separate diffusion SDEs in the space of images:

$$d\boldsymbol{x}_1 = \mu_\theta(\boldsymbol{x}_1, t, y_1)dt + g(t)d\bar{\boldsymbol{w}}_t$$
$$d\boldsymbol{x}_2 = \mu_\theta(\boldsymbol{x}_2, t, y_2)dt + g(t)d\bar{\boldsymbol{w}}_t$$

where $\mu_\theta(\boldsymbol{x}, t, y) := [f(\boldsymbol{x}, t) - g(t)^2 s_\theta(\boldsymbol{x}, t|y)]$

Then, to compute the semantic similarity of these two prompts, we can simply compute the divergences between these two SDEs.

Choosing the Jeffreys divergence as our distance function, we show that (*see paper for derivation*) this distance can be written in closed form via:

$$d_{ours}(y_1, y_2) = \mathbb{E}_{t\sim\text{unif}([0,T]),\boldsymbol{x}\sim\frac{1}{2}p_t(\boldsymbol{x}|y_1)+\frac{1}{2}p_t(\boldsymbol{x}|y_2)}\left[g(t)^2\|s_\theta(\boldsymbol{x}, t|y_1) - s_\theta(\boldsymbol{x}, t|y_2)\|_2^2\right]$$

which can be computed with Monte-Carlo sampling! This yields the following algorithm (below), which as an additional advantage, is visually interpretable (right column)

Algorithm 1 Conjuring Semantic Similarity

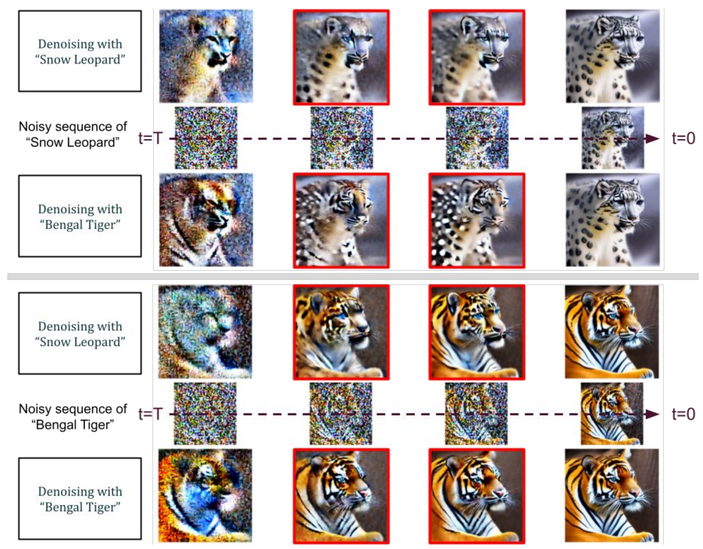
Require: Diffusion model s_θ, Prompts y_1, y_2, Monte-Carlo steps k
  Initialize $d = 0$
  for $i = 1 \dots k$ **do**
    $x_T \leftarrow$ Sample from initial distribution π
    $\hat{x}_T, \dots, \hat{x}_0 \leftarrow$ Denoise x_T conditioned on y_1
    $\tilde{x}_T, \dots, \tilde{x}_0 \leftarrow$ Denoise x_T conditioned on y_2
    $d \leftarrow d + \frac{1}{T}\sum_{t=1}^{T}\|s_\theta(\hat{x}_t, t|y_1) - s_\theta(\hat{x}_t, t|y_2)\|_2^2$
    $d \leftarrow d + \frac{1}{T}\sum_{t=1}^{T}\|s_\theta(\tilde{x}_t, t|y_1) - s_\theta(\tilde{x}_t, t|y_2)\|_2^2$
  end for
  return d/k ▷ Return similarity score

Meanings arise from Processes, not Models

In contrast to prior art, we view meanings not only a function of a trained model (weights, architecture etc.), but rather a function of the entire generative process. Different generative processes on the same model can yield completely different meaning representations for the same inputs.

This is a key feature of our definition. Indeed, A "well-trained" (text/image) generative model equipped with a random (diffusion/ next-token) sampler yields generative processes which clearly attributes "no meaning" to its inputs.

Empirical Validation



For the given prompts "Snow Leopard" and "Bengal Tiger", our method can be interpreted as integrating the distances between the resulting images in the two rows in each half of the above figure. In addition to a similarity score, our method provides a visual interpretation of the semantic differences between two prompts: Observing cells highlighted in red, we see the model converts pictures of Snow Leopards into Bengal Tigers by changing their characteristic spotted coats into stripes, and adding striped textures to the animal's face (top half of Figure), and conversely converts Bengal Tigers into Snow Leopards by changing their characteristic stripes into spotted coats (bottom half of Figure).

Model Category	STS-B	SICK-R	Avg (7 Datasets)
Contrastive-Trained Embeddings			
SimCSE-BERT	68.4	72.2	**76.3**
Autoregressive LLMs			
LLaMA-33B	71.5	73.0	**66.6**
Text-Conditioned Diffusion (Stable Diffusion)			
Init. Timestep Pred.	55.8	56.0	53.0
Direct Output Comp.	57.0	53.5	51.3
Ours (Conjuring Sim.)	70.3	66.0	**65.4**

Impressively, Zero-shot semantic alignment extracted purely from visual representations can rival embeddings derived from autoregressive LLMs and specially trained constrastive models, and significantly outperforms all other image diffusion model baselines we could think of.

Error Analysis: Analysis across part-of-speech (PoS) distributions (RG65, SimLex) reveals robust semantic alignment (with human annotators) of Noun semantics. However, although visually-grounded meanings from diffusion models tied to Verbs/Adjectives align less well with that of humans.