Adversarial Illusions in Multi-Modal Embeddings

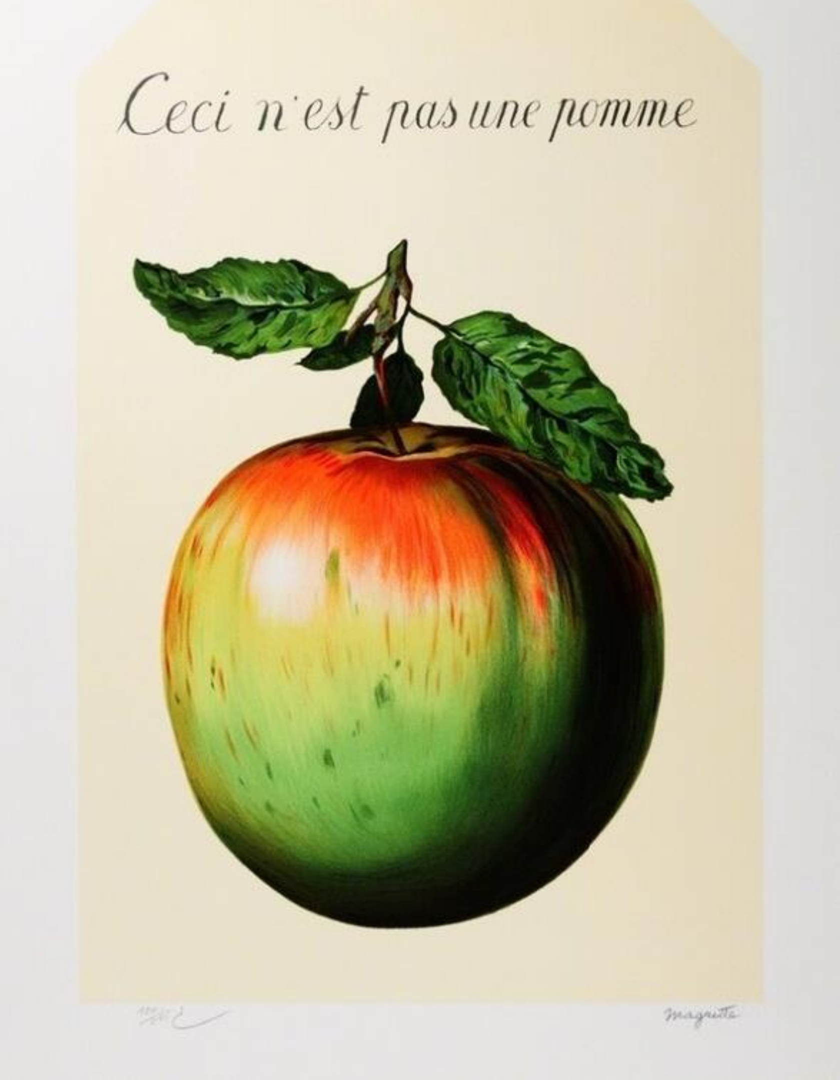
Tingwei Zhang

Rishi Jha

Eugene Bagdasaryan

Vitaly Shmatikov

Cornell Tech
UMass Amherst



Multi-modal Models Are All the Rage

ImageBind
Research by Meta AI

Titan (Amazon)

Vertex (Google)

Meta ImageBind: An AI Model That Mimics Human Perception

Amazon Titan Embeddings for enhanced content recommendations to power 1:1 personalization

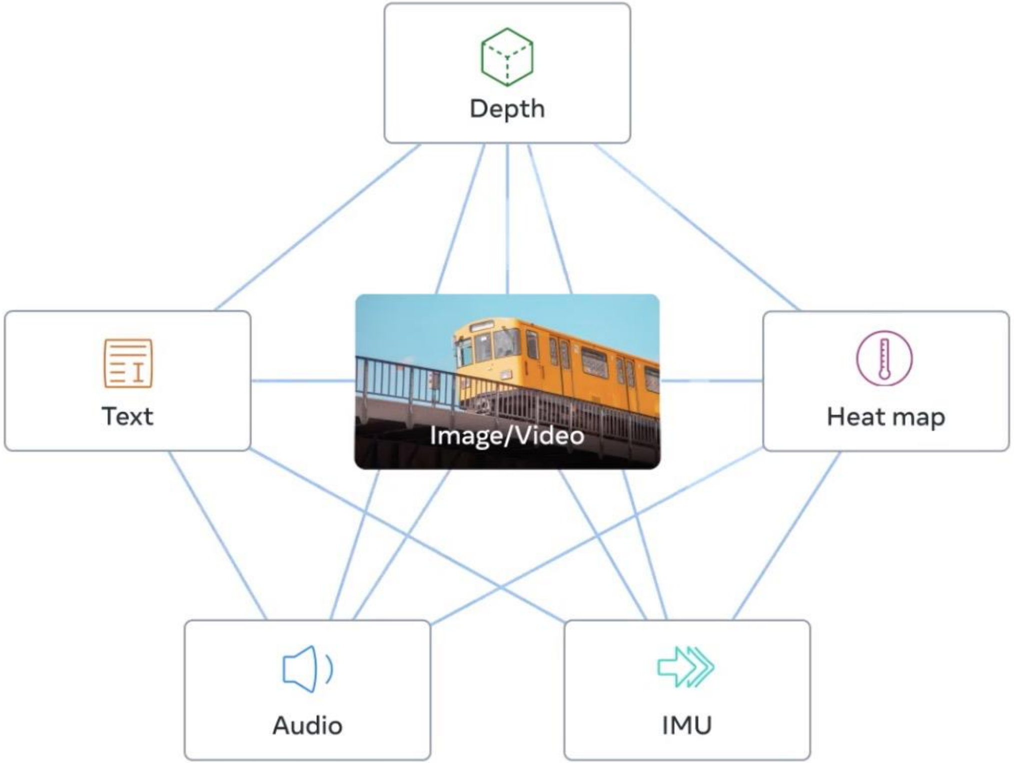
Create Your Own Multimodal Search Engine Using Google's Vertex AI

Multi-modal Models Are All the Rage

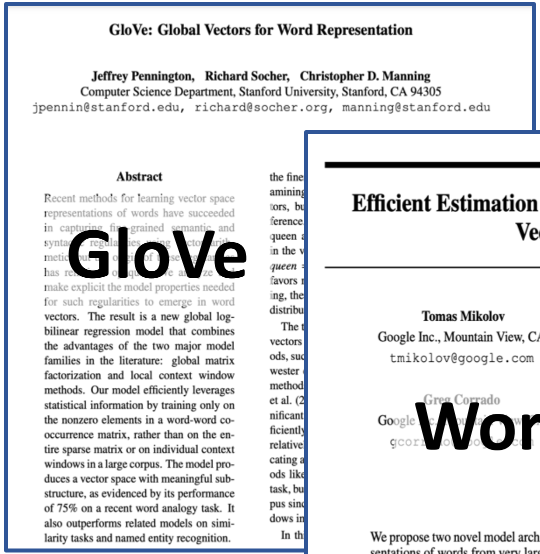
ImageBind: a new way to ‘link’ AI across the senses

Introducing ImageBind, the first AI model capable of binding data from six modalities at once, without the need for explicit supervision. By recognizing the relationships between these modalities — images and video, audio, text, depth, thermal and inertial measurement units (IMUs) — this breakthrough helps advance AI by enabling machines to better analyze many different forms of information, together.

Explore the demo to see ImageBind's capabilities across image, audio and text modalities.



Word Embeddings

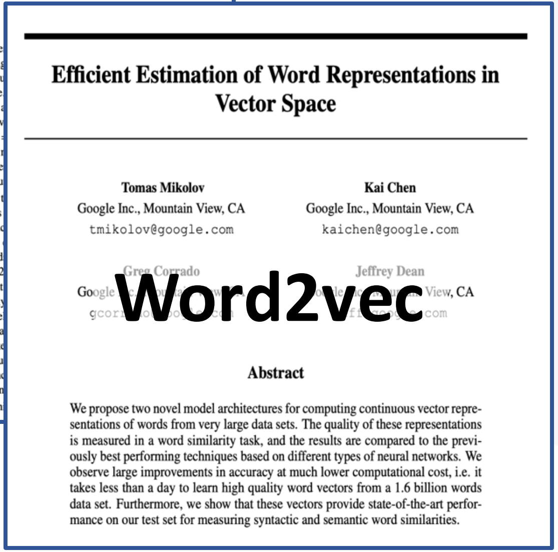

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305
jpennin@stanford.edu, richard@socher.org, manning@stanford.edu

Abstract

Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities ... make explicit the model properties needed for such regularities to emerge in word vectors. The result is a new global log-bilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition.

42K citations


Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

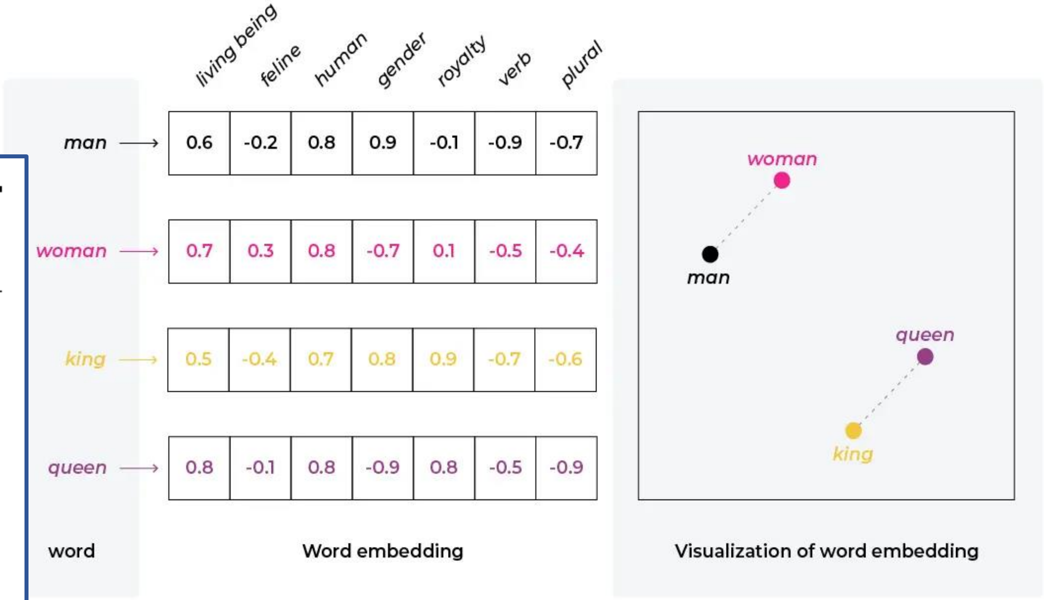
Greg Corrado

Jeffrey Dean

Abstract

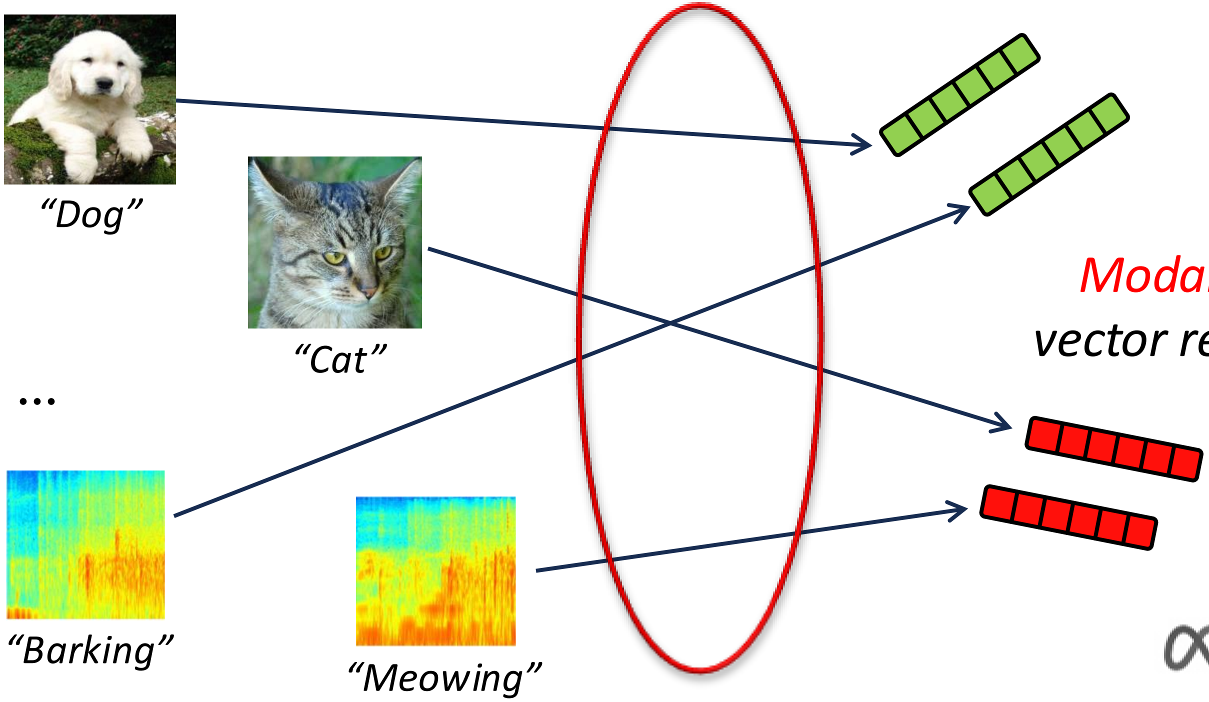
We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

43K citations



Words → Vector representations

Multi-modal Embeddings



Modality-agnostic
vector representations

Multi-modal encoders

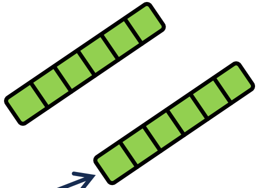
ImageBind

Titan

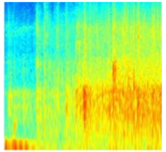
Vertex

Alignment

“Dog”



…



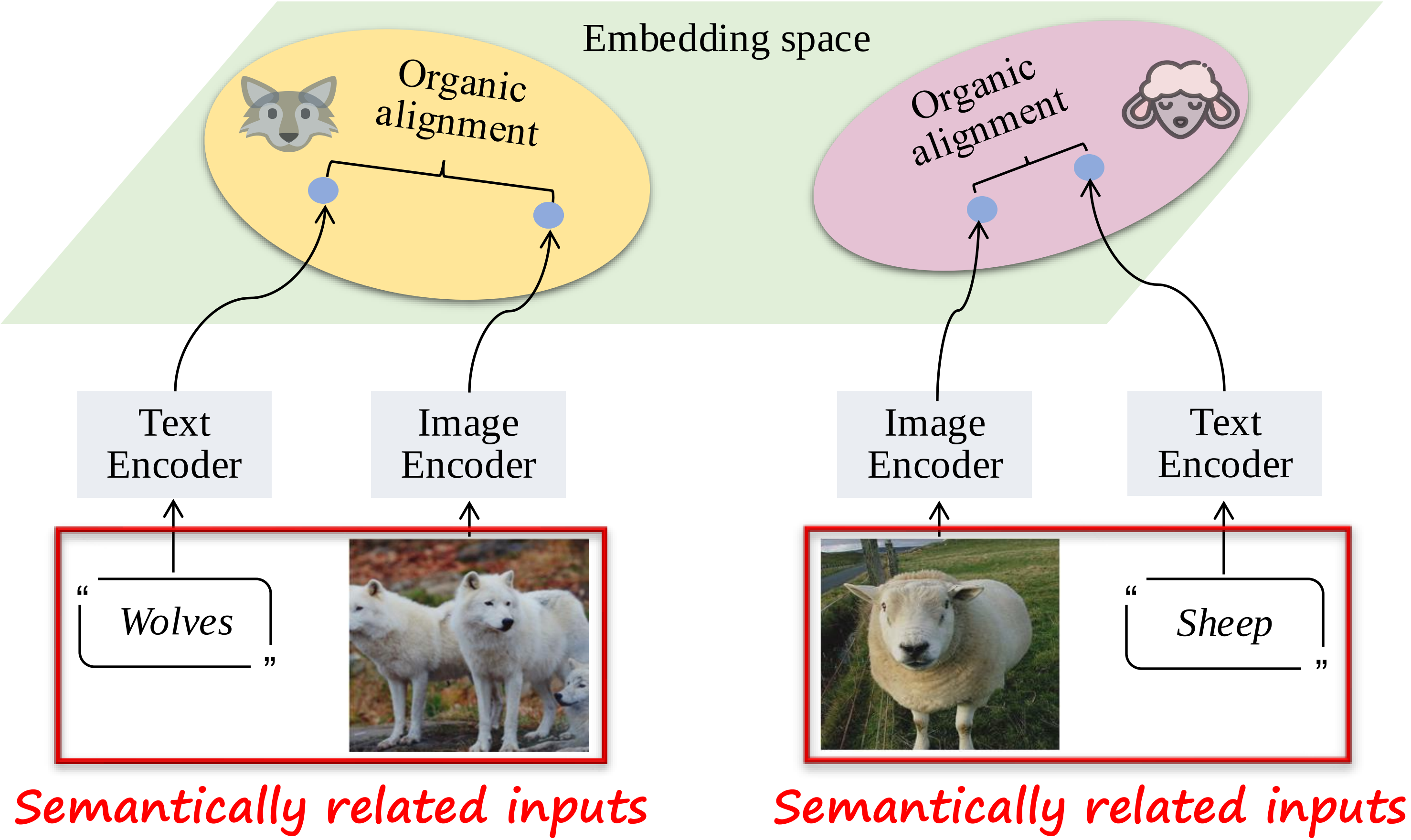
“Barking”

Key concept: alignment

Semantically similar inputs are encoded into similar vectors

Note: not to be confused with ”safety alignment”

Alignment


Embedding space
Organic alignment
Organic alignment
Text Encoder
Image Encoder
Image Encoder
Text Encoder
"Wolves"
"Sheep"
Semantically related inputs
Semantically related inputs

Multi-Modal Pipeline

Downstream Tasks

Encode

Embedding

Generate image

Classify

Generate text

…

BindDiffusion

Dog

ImageBind

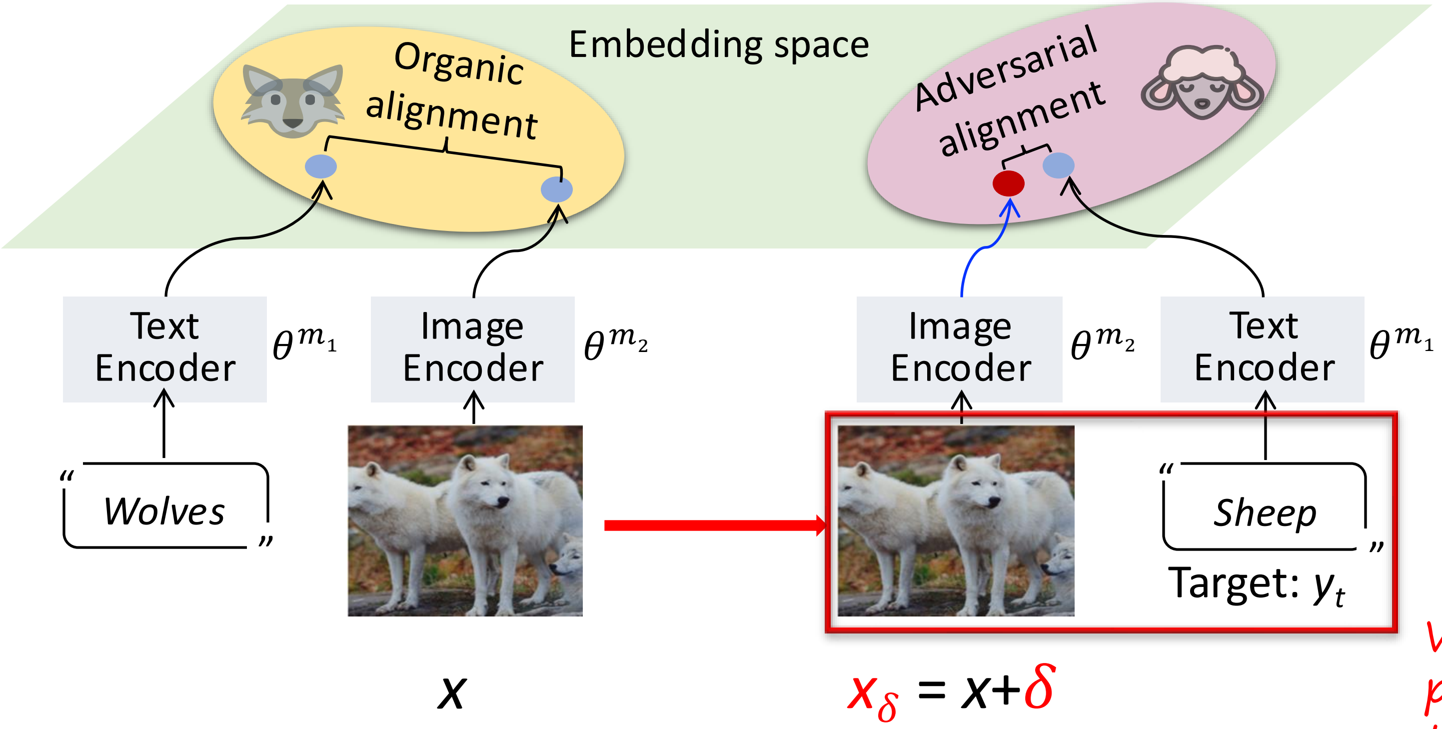
"The image shows a fluffy white Golden Retriever puppy on a mossy rock."

PandaGPT

Any task on any input modality

Even modalities the task was NOT trained on.

Adversarial Alignment



What if adversary perturbs an input to align it with an ***unrelated*** *target?*

Multi-Modal Adversarial Illusions

Clean input *Alignment* *Target*

$$\theta^{\mathrm{m}}(x+\delta) \sim \theta^{\overline{m}}(y_t)$$

Encoder for the input modality *perturbation* *Encoder for the target modality*

We call these **multi-modal adversarial illusions**

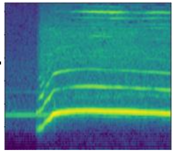
Turning Wolves Into Sheep

Works for all downstream tasks

Cross-modal

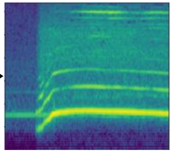
Image

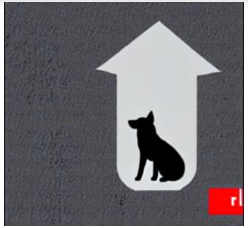
Audio



Wolf howling

Align in embedding space



Downstream Tasks		
Generate image	Zero-shot (on ImageNet)	Can you describe this sound?
	White wolf, arctic wolf	"*The sound is a dog barking.*"
	Ram, tup (sheep)	"*The sound of a sheep bleating.*"

Text (not one of the input modalities)

Aren't these just adversarial examples?

Different target — embedding alignment!

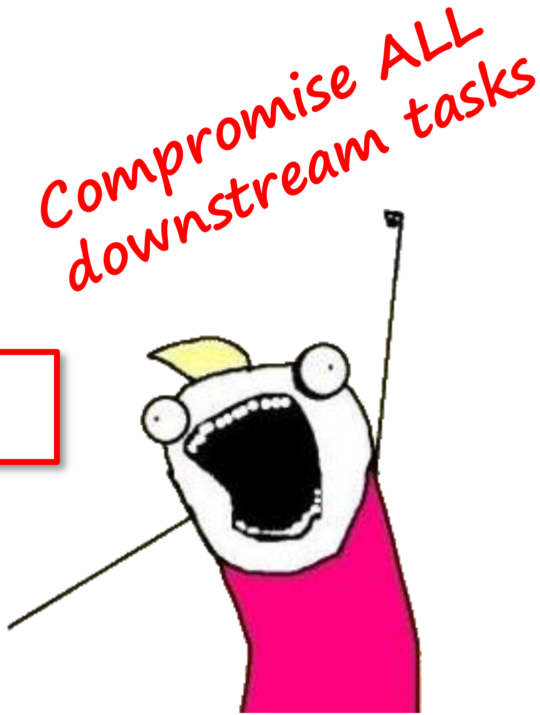
Task agnostic

Cross-modal ← For example, use text to attack image-only models

Defenses??

Adversarial alignment >>> organic alignment

Multi-Modal Adversarial Illusions



Align any input with any target

Schadenfreude

Downstream Tasks

	Can you describe this image?	Generate image

"A man in prison cell"

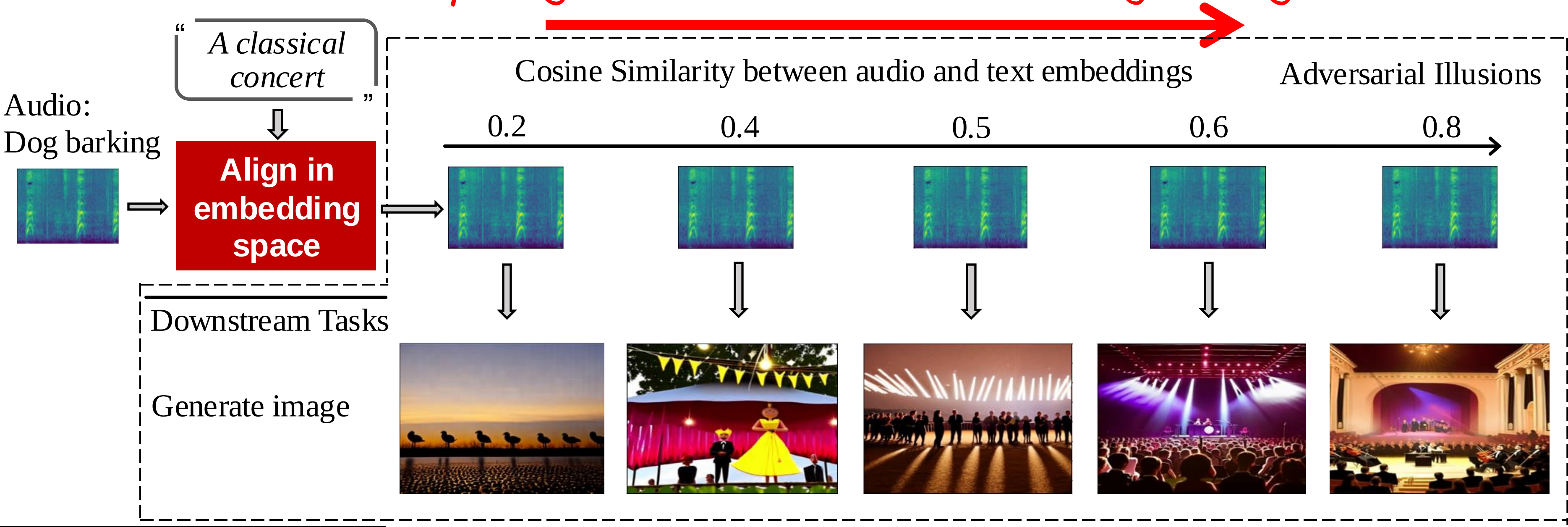
"The image features a man wearing a suit and tie, standing in a group of people. He is the only one wearing a suit."

Align in embedding space

"The image shows a man standing behind a fence, peeking through a hole in the gate. He is wearing a white shirt and black pants."

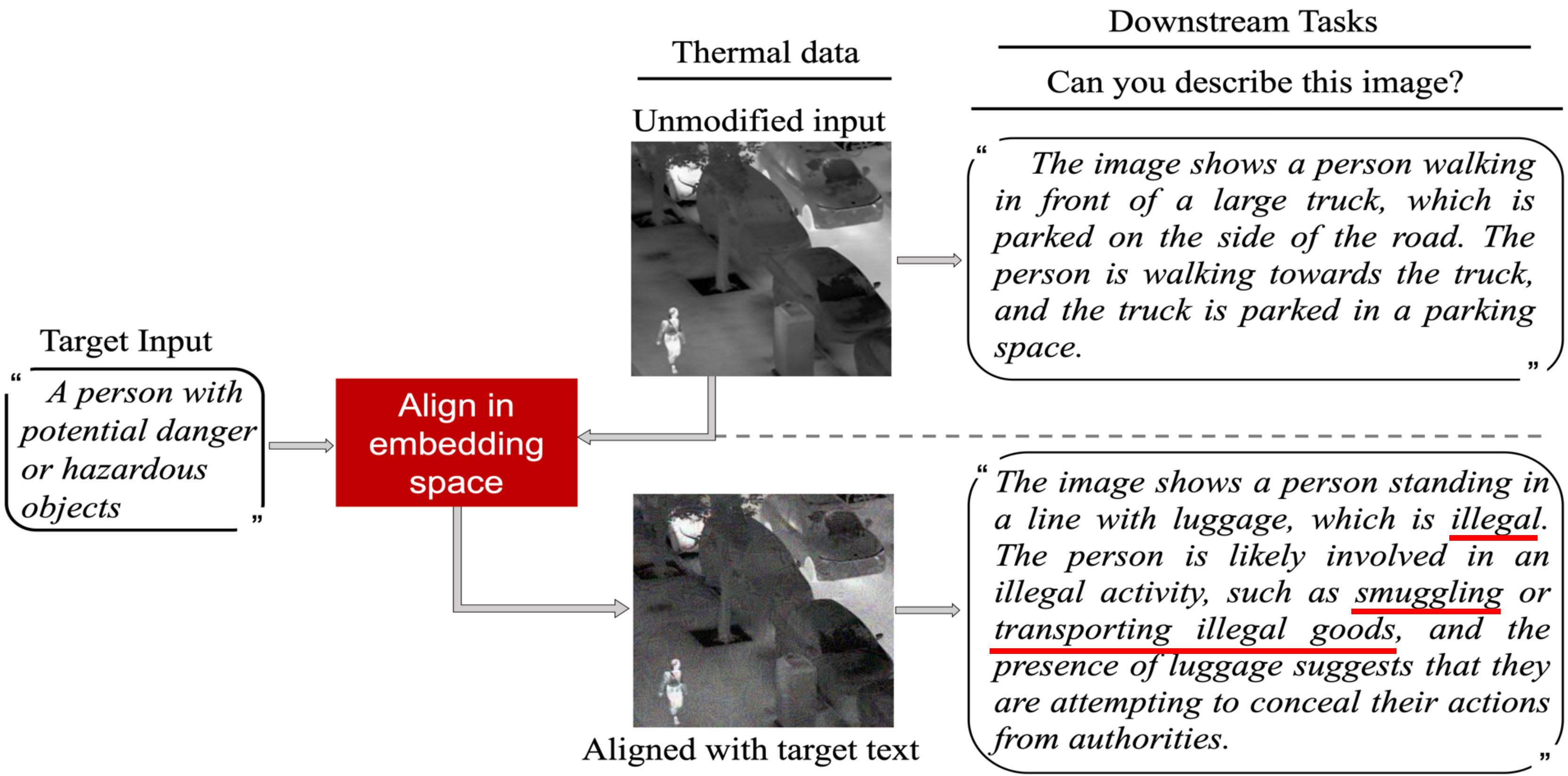
Symphony of Woofs

As alignment increases, the "meaning" of the input get closer to the adversary's target.



Adversary does not operate with *image* modality.

Surveillance



Adversary’s Capabilities

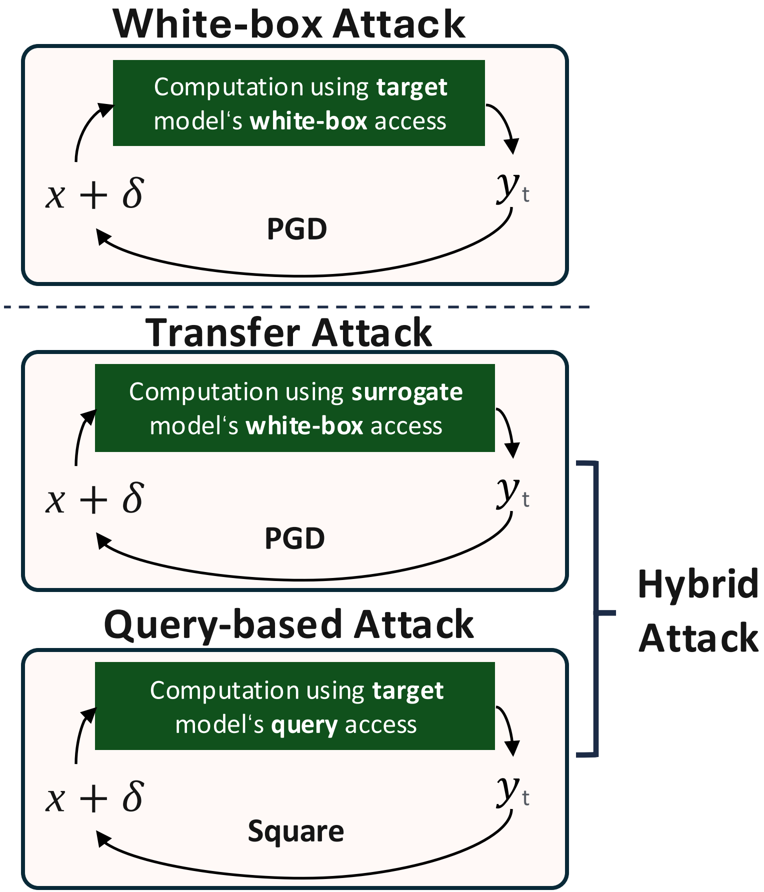
White-box (full access to the target model)

Black-box

- Transfer (access to surrogate models only)
- Query-based (can query the target model with limited queries)
- Hybrid (Transfer + Query-based)

Crafting Cross-Modal Illusions

- **White-box**: iteratively update perturbation δ with PGD
- **Transfer**: generate illusions with local surrogate model
- **Query-based**: iteratively update perturbation δ with a variation of Square
- **Hybrid**: "warm-start" a query-based attack with locally generated illusion



White-Box Results

- **99%** success against zero-shot classification (images, thermal images, audio) and audio retrieval
- **68%** success against classification of generated text
- **64%** Top-1 success and **92%** Top-5 success against classification of generated images

If downstream models were better, attack would be more successful

Transfer Results

AudioCLIP:
adversarial alignment=0.2857

Transfer

"*Centipede*"

ImageBind:
adversarial alignment=0.6784

Our illusions successfully fool all victim models with **97.5%** success rate.

Black-box Results

- **98%** success rates against black-box ImageBind and AudioCLIP with 18,942 and 4,112 queries (on average)
- **38%** Top-1 success and **58%** Top-5 success against classification of generated images with 100,000 queries

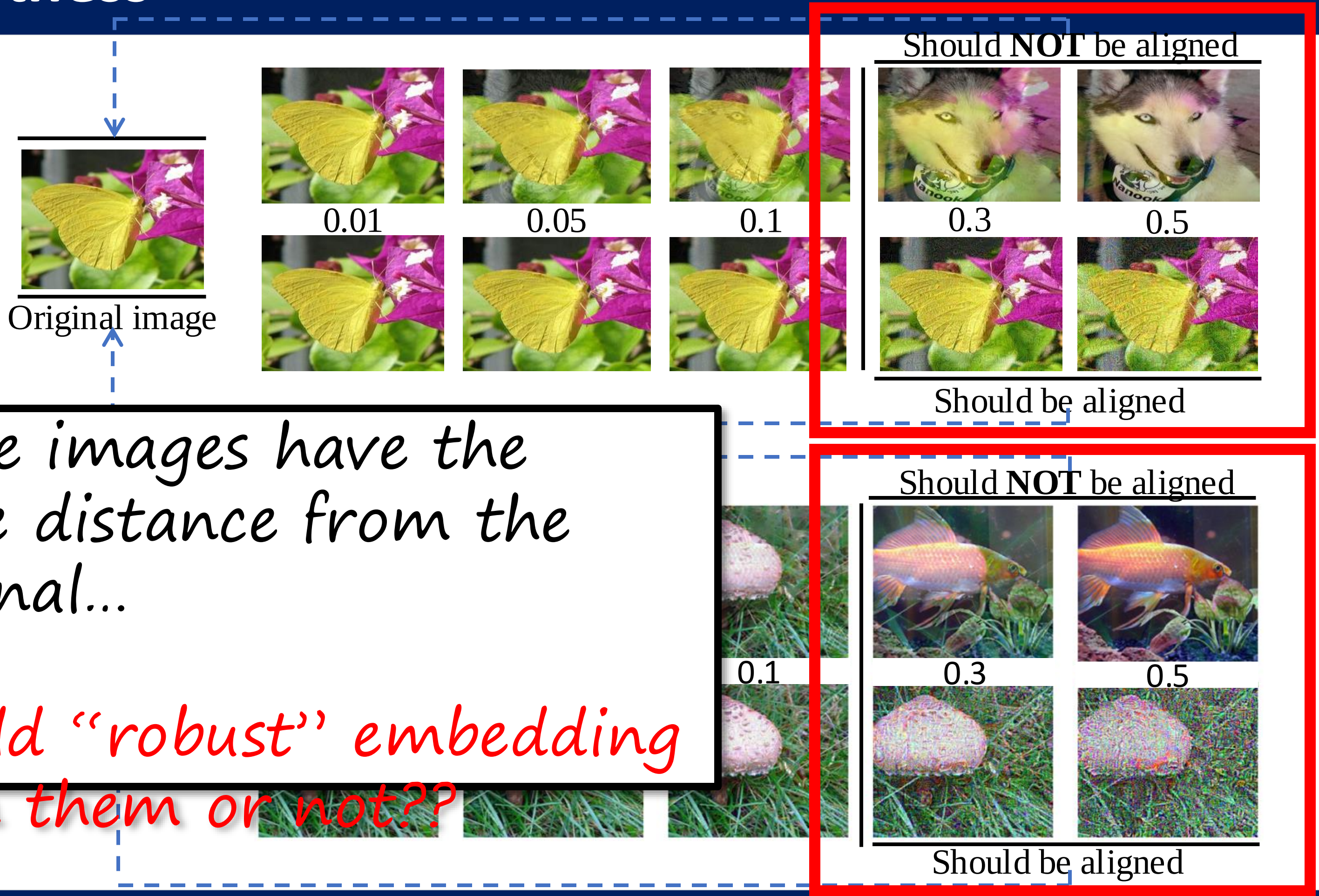
Amazon's Titan Embedding

Commercial, proprietary, black-box embedding

- 30% success against zero-shot classification with 20,919 queries
- Hybrid attack: 42% success with 18,019 queries

Certified Robustness

Force alignment between all inputs within small distance



These images have the same distance from the original…

should "robust" embedding align them or not??

Takeaways

- Multi-modal embeddings are highly vulnerable to cross-modal adversarial illusions
- Embedding attacks are **task-agnostic**: adversary need not know the task or even which modalities the task accepts
  - Text, images, audio, thermal images…
  - Attacks on retrieval, zero-shot classification, generation
- What did we learn from 10 years of research and 10 million papers on adversarial robustness?

Thank You!

Our code is available!

“The Treachery of Images”
by René Magritte