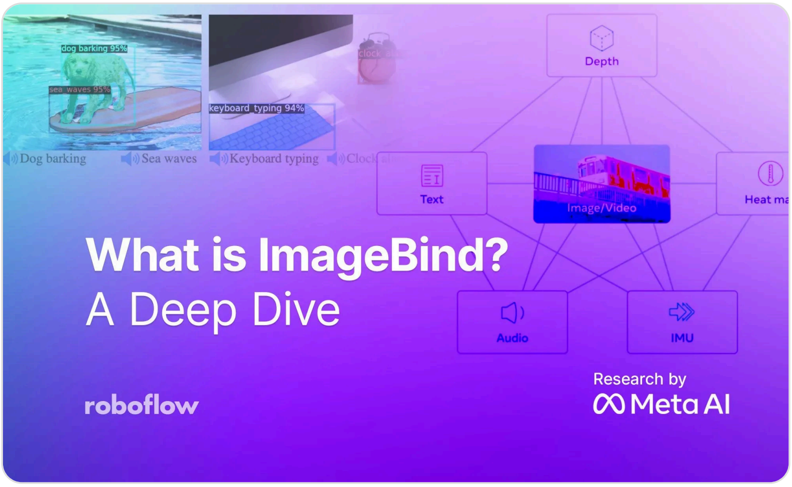
roboflow Product Solutions Resources Pricing Docs Sign In Sign Up

ROBOFLOW UNIVERSE ROBOFLOW DEPLOY MODEL DEPLOYMENT

What is ImageBind? A Deep Dive

Written by
James Gallagher

MAY 12, 2023 | 6 MIN READ



James Gallagher
MAY 12, 2023 | 6 MIN READ

Search blog

TABLE OF CONTENTS

Over the last few years, AI and computer vision researchers and practitioners have built powerful model architectures that work across a single modality. For instance, models like YOLO take in an image and generate predictions from the image; GPT-3.5 takes in text and generates text.

There is interest in building models that can combine multiple different modalities (like audio and images), known as multimodal models. We have taken steps in this direction with models like DALL-E, which can accept a text prompt and generate an image, or Grounding DINO, which takes a text prompt and generates bounding boxes. With that said, these models cross two modalities; in the aforementioned examples, text and image modalities are combined.

ImageBind, a new embedding model by Meta Research, is pioneering an approach to AI embeddings that allows you to encode information across many different modalities, from images to text to audio to depth information. These modalities, which would traditionally be in their own models, are combined into a single space, enabling various use cases from advanced semantic search to new ways of interacting with models.

In this guide, we're going to discuss:

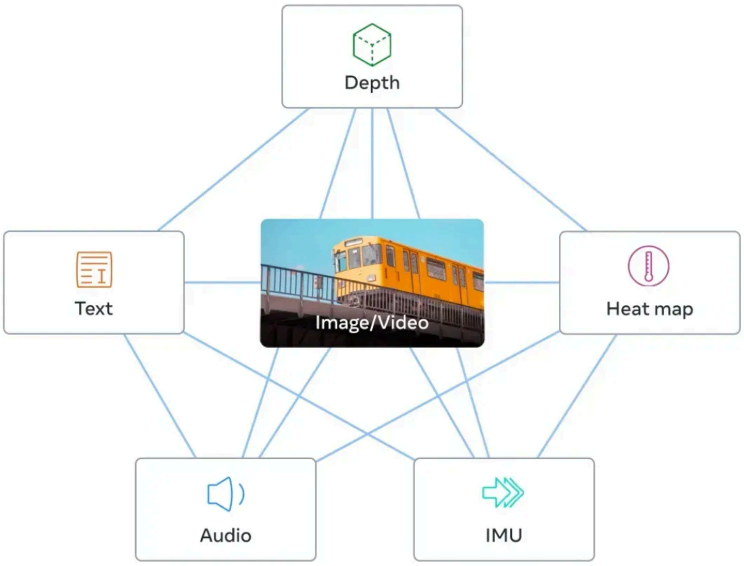
1. What is ImageBind?
2. How does ImageBind work?
3. What can you do with ImageBind?
4. How you can get started with ImageBind

Without further ado, let's get started!

What is ImageBind?

ImageBind, released in May 2023 by Meta Research, is an embedding model that combines data from six modalities: images and video, text, audio, thermal imaging, depth, and IMUs, which contain sensors including accelerometers and orientation monitors.

Using ImageBind, you can provide data in one modality – for example, audio – and find related documents in different modalities, such as video.



Combining data across modalities. Source: Meta Research.

Through ImageBind, Meta Research has shown that data from many modalities can be combined in the same embedding space, allowing richer embeddings. This is in contrast to previous approaches, where an embedding space may include data from one or two modalities. Later in this post, we'll talk about the practical applications of the ImageBind embeddings.

As of writing this post, ImageBind follows a series of pioneering new open source models released by Meta Research with uses in computer vision. This includes the Segment Anything Model that set a new standard for zero-shot image segmentation, and DINOv2, another zero-shot computer vision model.

How Does ImageBind Work?

ImageBind was trained with pairs of data. Each pair mapped image data – including videos – to another modality, and the combined data was used to train a large embedding model. For instance, image-audio pairings and image-thermal pairings were used. ImageBind found that features for different modalities could be learned using the image data used in their training.

A notable conclusion from ImageBind is that pairing images with another modality, then combining the results in the same embedding space, is sufficient to create a multi-modal embedding model. Previously, one would need to have separate models that mapped different modalities together.

The embeddings from ImageBind can be combined with other models to directly leverage generative AI models alongside ImageBind. In the ImageBind paper, Meta Research notes that they use a pre-trained DALLE-2 diffusion model (private) and replaced the prompt embeddings with audio embeddings from ImageBind. This enabled the researchers to generate images using DALLE-2 directly with speech, without an intermediary model (i.e. speech-to-text).

What Can You Do With ImageBind?

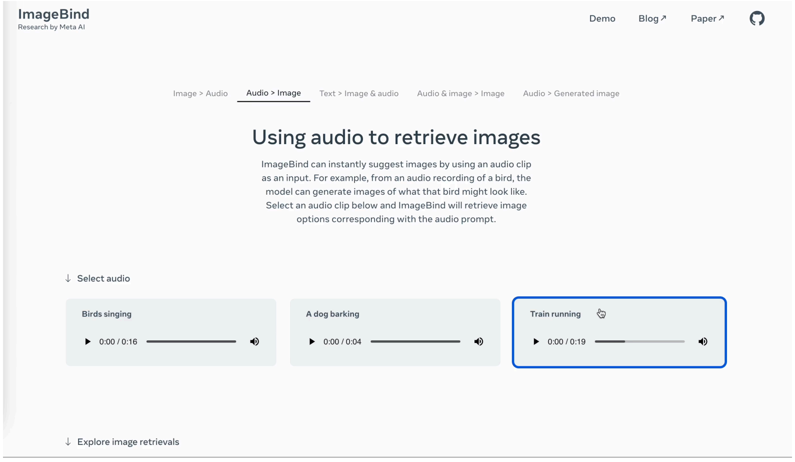
Like all embedding models, there are many potential use cases for ImageBind. In this section, we're going to talk about three primary use cases for ImageBind: information retrieval, zero-shot classification, and connecting the output of ImageBind to other models.

Information Retrieval

One can build an information retrieval system that traverses modalities with ImageBind. To do so, one would embed data in supported modalities – such as video, depth data, and audio – and then create a search system that embeds a query in any modality and retrieves related documents.

You could have a search engine that lets you upload a photo and shows you all of the audio materials associated with that image. One example scenario where this would be useful is in birding. A nature enthusiast could enter a bird call that they hear and the search engine could return the closest image documents it has stored. Inversely, the enthusiast could take a photo of a bird and retrieve an audio clip with its call.

Meta research published an **interactive playground** to accompany the paper that shows information retrieval across different modalities.



Using audio to retrieve images

ImageBind can instantly suggest images by using an audio clip as an input. For example, from an audio recording of a bird, the model can generate images of what that bird might look like. Select an audio clip below and ImageBind will retrieve image options corresponding with the audio prompt.

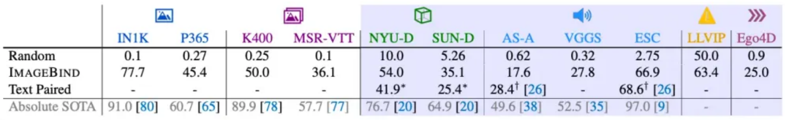
Classification

ImageBind embeddings can be used for zero-shot classification. In zero-shot classification, a piece of data is embedded and fed to the model to retrieve a label that corresponds with the contents of the data. In the case of ImageBind, you can classify audio, images, and information in the other supported modalities.

Further, ImageBind supports few-shot classification, where a few examples of data can be sent to the model before classification is run to achieve better performance on a specific task.

ImageBind realized “gains of approximately 40 percent accuracy in top-1 accuracy on ≤four-shot classification” when compared to Meta’s self-supervised and supervised AudioMAE models, according to **Meta’s summary blog post** published to accompany the project.

Here are the results of ImageBind’s zero-shot classification capabilities:

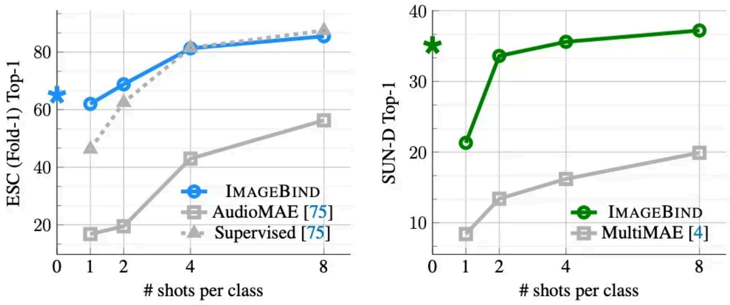


	IN1K	P365	K400	MSR-VTT	NYU-D	SUN-D	AS-A	VGGS	ESC	LLVIP	Ego4D
Random	0.1	0.27	0.25	0.1	10.0	5.26	0.62	0.32	2.75	50.0	0.9
IMAGEBIND	77.7	45.4	50.0	36.1	54.0	35.1	17.6	27.8	66.9	63.4	25.0
Text Paired	-	-	-	-	41.9^*	25.4^*	$28.4^\dagger$ [26]	-	$68.6^\dagger$ [26]	-	-
Absolute SOTA	91.0 [80]	60.7 [65]	89.9 [78]	57.7 [77]	76.7 [20]	64.9 [20]	49.6 [38]	52.5 [35]	97.0 [9]	-	-

ImageBind performance on zero-shot classification tasks across datasets from different modalities.

From this table, we can see ImageBind achieves strong performance across a range of datasets from different modalities. While ImageBind does not achieve scores in excess of the "Absolute SOTA" score, Meta notes the models that achieve these scores typically use more supervision and other features.

With few-shot classification, ImageBind found strong results when compared to AudioMAE (supervised and unsupervised) and MultiMAE:



Zero- and few-shot classification performance for ImageBind when compared to other audio models

Combining Modalities for New Applications

Meta experimented with using ImageBind embeddings to allow for audio-to-image generation with DALLE-2. While this was done using a private model, the experiment shows the capability of using ImageBind embeddings for generative AI and augmented object detectors.

With ImageBind, one can provide multiple potential inputs for a generative AI model – audio, video, text – without having separate translation logic, such as an intermediary model to convert the input data into text for use with a text embedding model.

In another example, Meta used audio embeddings calculated using ImageBind with Detic, an object detection model. Meta replaced the CLIP text embeddings with their audio embeddings. The result was an object detection model that could take in audio data and return bounding boxes for detections relevant to the audio prompt.

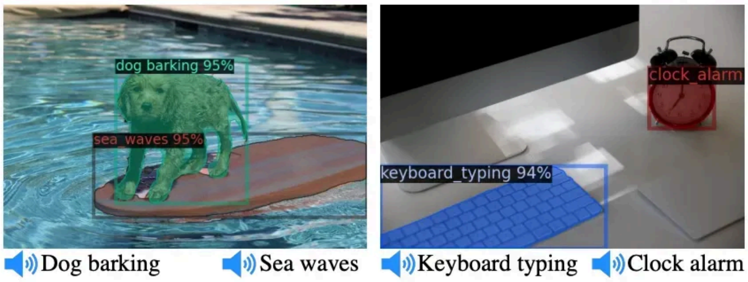


Figure 5. Object detection with audio queries. Simply replacing Detic [86]'s CLIP-based 'class' embeddings with our audio embeddings leads to an object detector promptable with audio. This requires no re-training of any model.

How to Get Started

To experiment with retrieving images across different modalities, you can use the ImageBind playground published by Meta research. This playground provides a few pre-made examples that show information retrieval in action.

ImageBind is open source. An "imagebind_huge" checkpoint is provided for use with the project. In the project README, there are examples showing how to feed text, image, and audio data into ImageBind. This code is a great way to get started with ImageBind. The model and accompanying weights are licensed under a CC-BY-NC 4.0 license.

With the repository, you can build your own classifiers and information retrieval systems that use ImageBind. Although no instruction is given for this in the README, you can also experiment with using the ImageBind embeddings with other models, such as in Meta's example of using audio embeddings with Detic.

Conclusion

ImageBind is the latest in a range of vision-related models published by Meta Research, following on from publications earlier this year including DINOv2 and Segment Anything.

ImageBind creates a joint embedding space that encodes information from six modalities, demonstrating that data from different modalities does not require separate embeddings for each modality.

The modal can be used for advanced information retrieval across modalities and zero- and few-shot classification. The embeddings, when combined with other models, for object detection and generative AI.

Frequently Asked Questions

What size are the ImageBind weights?

The `imagebind_huge` checkpoint that was released with ImageBind is 4.5 GB. This is the only model weight checkpoint released with the model.

What modalities does ImageBind support?

ImageBind supports image, text, audio, depth sensor readings, thermal data, and IMU readings.

Cite this Post

Use the following entry to cite this post in your research:

James Gallagher. (May 12, 2023). What is ImageBind? A Deep Dive. Roboflow Blog: https://blog.roboflow.com/what-is-imagebind/

Discuss this Post

If you have any questions about this blog post, start a discussion on the Roboflow Forum.

James Gallagher

James is a Technical Marketer at Roboflow, working toward democratizing access to computer vision.

VIEW MORE POSTS

TOPICS:

Roboflow Universe, Roboflow Deploy, Model Deployment

MORE ABOUT

VIEW ALL

Want to learn more about Roboflow? Email sales@roboflow.com or talk to sales with our sales team.

roboflow

For sales inquiries:

sales@roboflow.com

Book a demo

aws PARTNER ML Software Competency

aws Qualified Software

AICPA SOC

PRODUCT

Sign In / Sign Up
Universe
Annotate
Train
Deploy
Integrations
Pricing

ECOSYSTEM

Notebooks
Autodistill
Supervision
Inference
Roboflow 100
Open Source

DEVELOPERS

User Forum
Templates
Blog
Learn Computer Vision
Convert Annotation Formats
Computer Vision Models

INDUSTRIES

Manufacturing
Oil and Gas
Retail
Safety & Security
Transportation
All Industries

MODELS

YOLOv9
YOLOv8
YOLOv5
YOLO-World
YOLO-NAS
CLIP
Grounding DINO
Multimodal Models
Explore More Models

COMPANY

About Us
Careers
Press
Terms of Service
Privacy Policy
Roboflow Sitemap
Blog Sitemap
Status