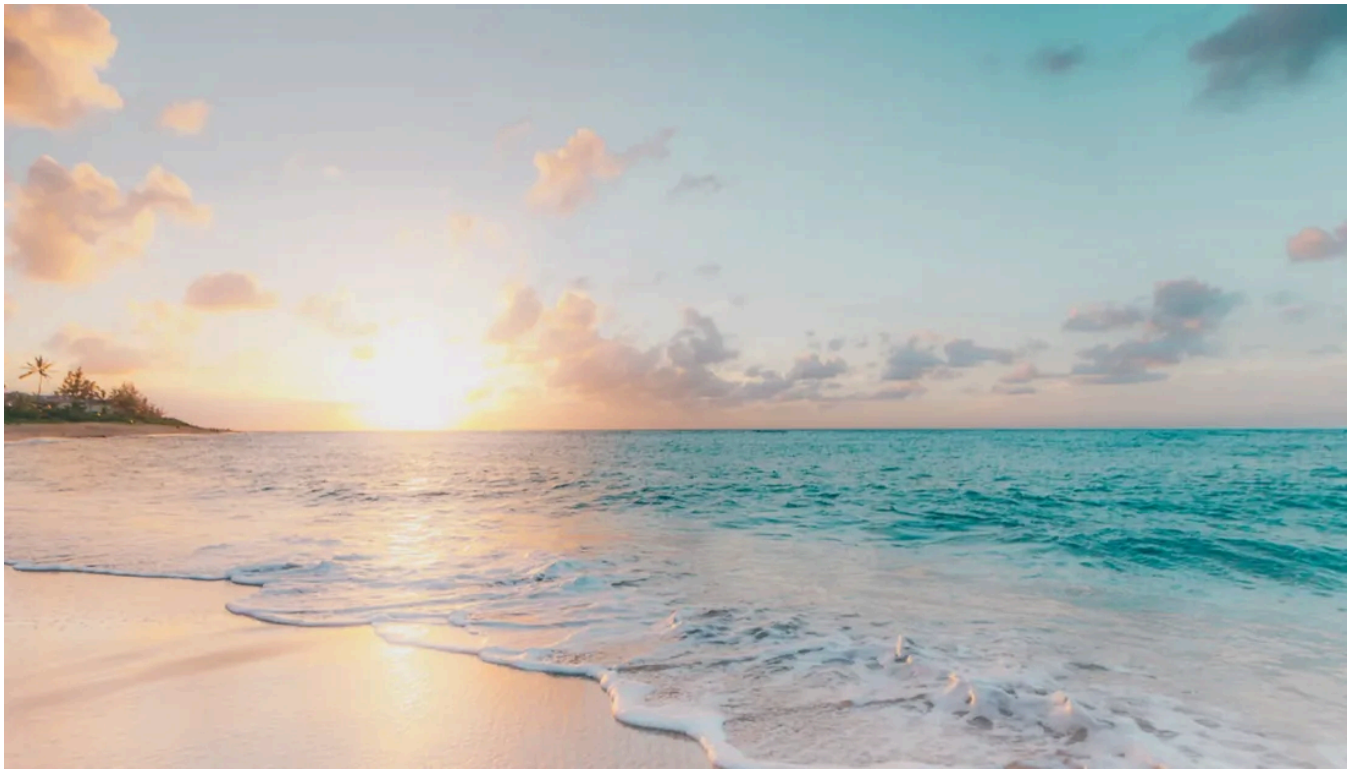
ImageBind: One Embedding Space To Bind Them All — paper explained

Shrinivasan Sankar · Follow

7 min read · May 23, 2023

Listen Share

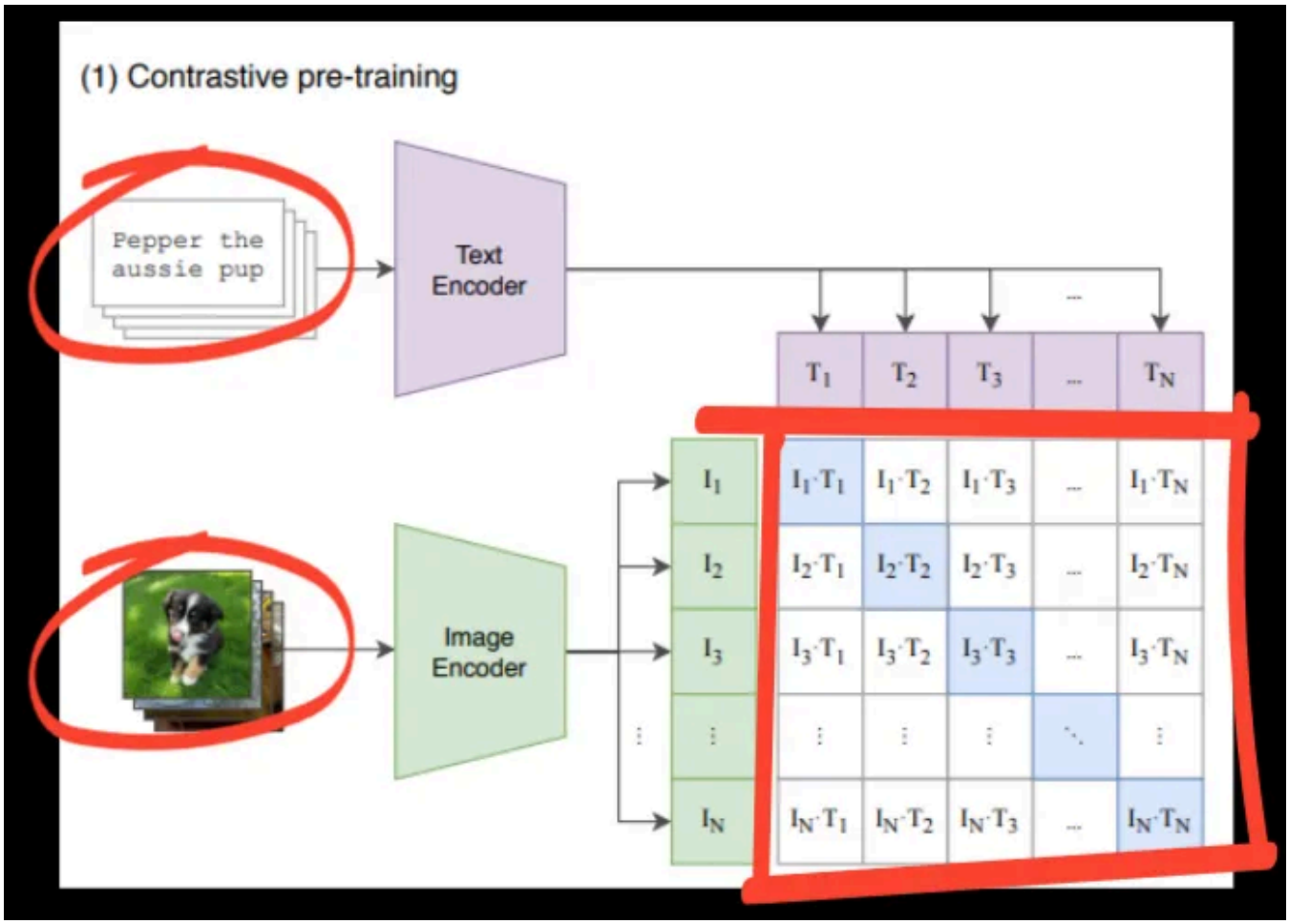


Introduction

Images are truly binding. An image of a beach, reminds the pleasant sound of the waves and when I simply say the words, “sunshine, sandy beach and drink” you would imagine same image with sunshine and you sitting in it with a drink. Its all simply because human mind not just receives information through audio, video, text or touch, but it also somehow aligns these modalities to build a mental map of all the perceived data.

Though there is abundance of data on the web these days generated by humans, only some are naturally aligned with images or videos. For example, video and audio are naturally aligned on the web with which we can train just a image-audio model. But what about other modalities like audio and text? So can we come up with a way to bind many modalities together with Images? This is exactly what the ImageBind paper addresses. It shows that the emergence of alignment between modalities called the emergent alignment and the results are quite promising. Without further adieu, lets dive deeper into **ImageBind**.

CLIP and Motivation for ImageBind



The idea of linking or connecting modalities at scale using web scale data was first established in the CLIP which stands for Contrastive Language Image Pre-training. CLIP takes text prompts and images as input and connects them semantically. It does this at web scale by training on 200 million image-text pair dataset called WebImageText which were fully gathered from the web without any manual labelling.

CLIP introduced contrastive learning which is to distinguish between positive pairing of the image and text versus negative pairing of image and text combinations (see figure above). This simple switch to contrastive objective made CLIP much more efficient compared to using a predictive objective of standard classifiers. The loss used was called the *InfoNCE* loss which maximises the similarity between correct pairs and minimises the similarity between incorrect pairs.

Similar to CLIP's approach to leverage contrastive learning to pairs of modalities namely image and text, there have also been other works inspired by CLIP that pairs other modalities like audio with images namely, *Audioclip* which pairs audio and text. There are also ideas like *Contrastive multiview coding* which pairs images with depth. And there are also works like "*Audiovisual instance discrimination with cross-modal agreement*" which pair video and audio.

The biggest problem with these pairings is that one is not useful for the other. For example, a model pre-trained with image-text embeddings is not useful for audio. This exact problem is what is addressed by ImageBind.

Quick shoutout

By the way if you like our blog, why not checkout our YouTube channel where we explain AI papers and ideas. ImageBind is also explained in a video.

ImageBind and Multiple Modalities

ImageBind considers several modalities namely — image/video, text , audio, depth, thermal and IMU which stands for Inertial Measurement Unit and includes the accelerometer and gyroscope. The main goal of this work is to learn a "single joint embedding space for all modalities" and use images as the binding modality.

If *I* stands for images or videos and *M* stand for any other modality, then we use deep neural networks as encoders to extract embeddings from each of the modalities. There is a separate encoder for each modality. More specifically, they use variations of Vision Transformers for all of the encoders. For images and videos they use ViT-H and for text encoding they use OpenCLIP. For the audio they use ViT-B and for thermal and depth they use ViT-S.

During ImageBind training, the weights of the image and text encoder architectures are kept frozen and the weights of all other modalities are updated. Because these two models are frozen, they use a pre-trained models for encoding images and texts.

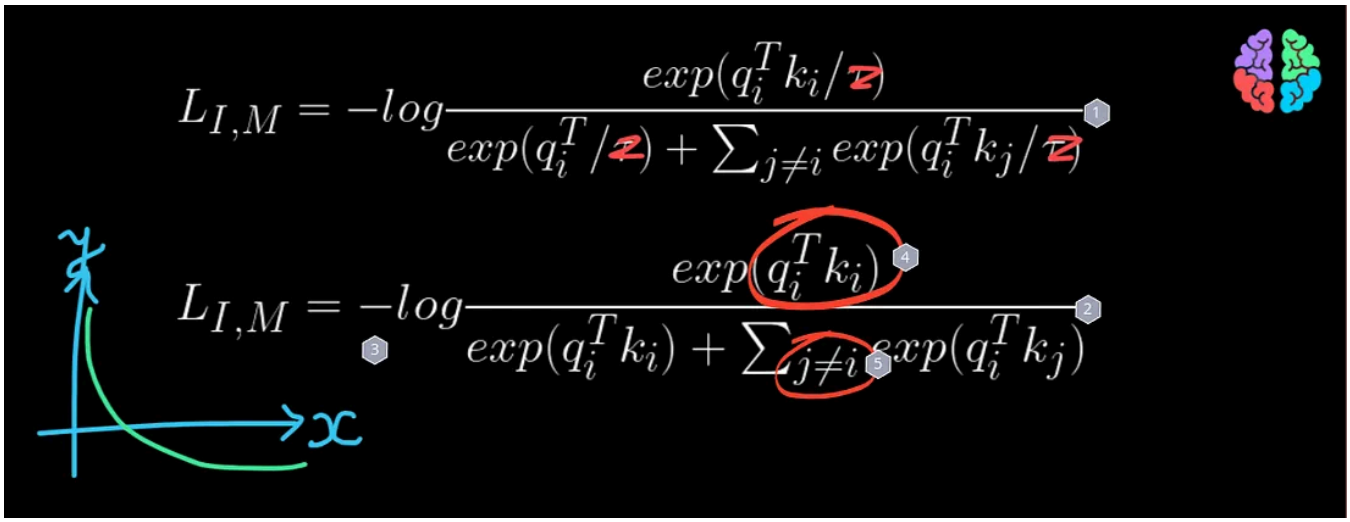
This freezing ensures alignment to emerge between modalities for which we don't have any natural alignment, for example, between audio and depth.

Because the inputs are in different forms, they do slight pre-processing before using them. For example, when dealing with videos, they sample 2 frames from 2 seconds of a given video. With audio, they convert 2 second audio clips into mel-spectrograms. Thermal and depth images are treated as 1 channel images. When it comes to IMU, it has accelerometer and gyroscope measurements which have a X, Y and Z dimension. They take a 5 second clip of the data and project using 1D convolutons which are fed once again into a transformer architecture.

The pre-processed inputs are then passed through the encoders whose outputs are then passed through a simple linear layer in order to ensure they are of the same dimension before being trained with a loss called the *InfoNCE* loss. Lets say the output of the image or video embeddings is q and the outputs from any of the other modalities is k. With that, lets look at the loss function.

Loss Function

The loss function InfoNCE looks a bit scary in the paper and its the modified cross entropy loss and it extends the idea of contrastive learning to multiple modalities.



To understand it, I am going to simplify it by first stripping off the temperature *tau* which is trivial resulting in this simplified equation (see figure above). During training, we are going to optimize this loss to achieve a minima. The loss is a negative log function and a plot of negative log looks somewhat like this which indicates that in order to minimize the value of y, we need to achive high values of x. This means we need to increase the numerator and decrease the denominator as much as we can. The numerator is nothing but a dot product or similarity of

embeddings from image modality q and any other modality k and its only for the positive cases as both q and k have the index i indicating they are positive pairs. The denominator on the other hand is the dot product of embeddings of negative cases which do not form a pair. So optimizing this equation brings the embedding of different modalities for the positive example closer together and pushes the negative cases far apart.

In terms of the embeddings, the loss brings closer the embeddings amd creates a joint embedding space to bind together all the modalities k with the image modality q. This ensures alignment to emerge between modalities for which we don't have any natural alignment with and this is what they call the *emergent alignment* in the paper.

Results

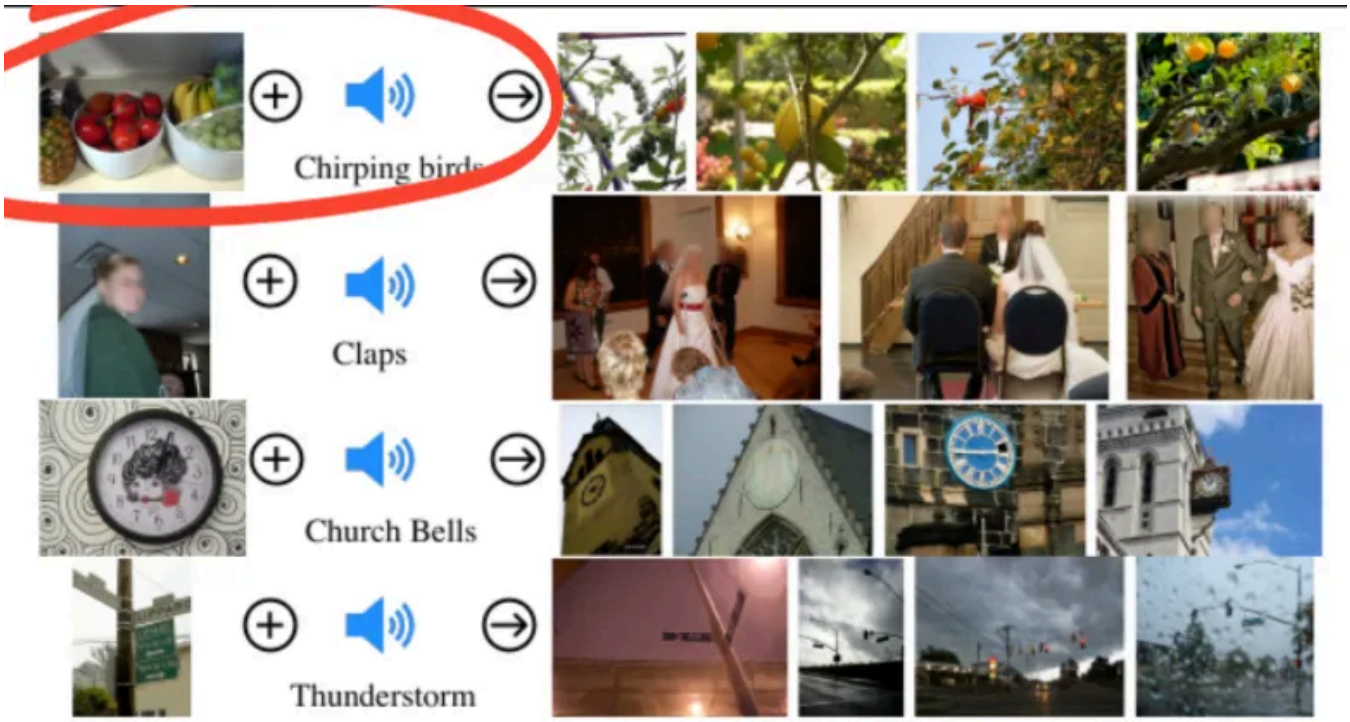
To demonstrate emergent alignment, they have chosen to show zero-shot classification of depth, audio, thermal and IMU using text prompts. You can notice that these datasets are aligned with images. But the results are shown for text prompting as input. So somehow the alignment between text and other modalities has emerged. Because ImageBind is so novel there is no real baseline to compare against.

	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† [26]	-	68.6† [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]	-	-

They also show that they are able to perform audio retrieval and classification without even training or fine-tuning with any audio data. What not, this is the only emergent approach and everything else is trained on specific audio data by some means.

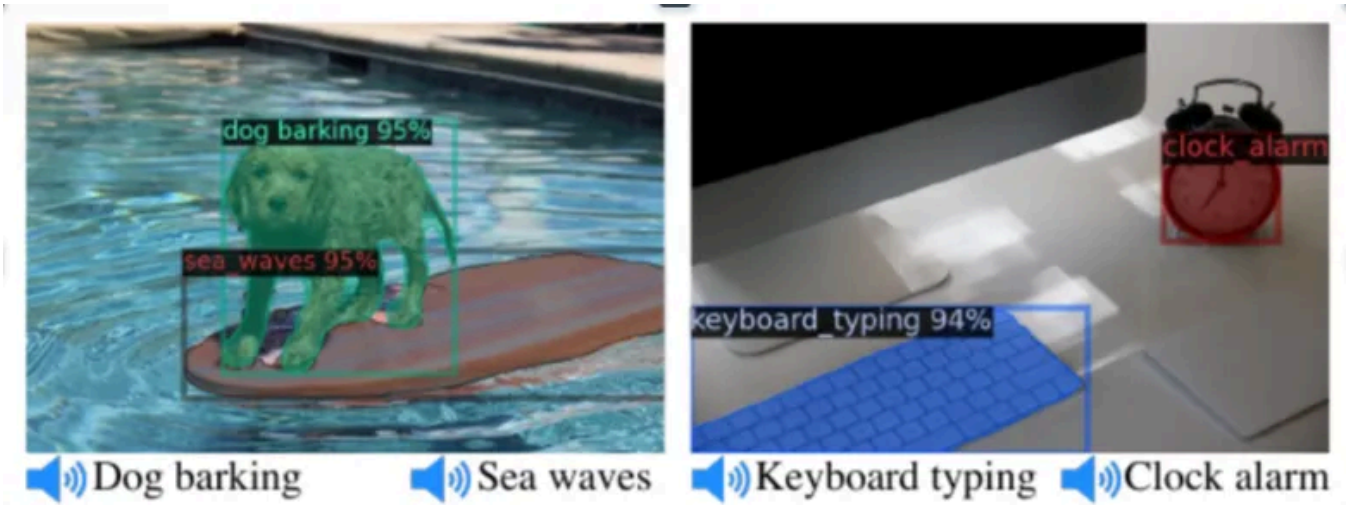
	Emergent	Clotho		AudioCaps		ESC
		R@1	R@10	R@1	R@10	Top-1
Uses audio and text supervision						
AudioCLIP [26]	✗	-	-	-	-	**68.6**
Uses audio and text loss						
AVFIC [50]	✗	3.0	17.5	8.7	37.7	-
No audio and text supervision						
IMAGEBIND	✓	**6.0**	**28.4**	**9.3**	**42.3**	66.9
Supervised						
AVFIC finetuned [50]	✗	8.4	38.6	-	-	-
ARNLQ [52]	✗	12.6	45.4	24.3	72.1	-

We also have the ability to do embedding space arithmetic where we provide an input image say, an image with berries and in the audio we say chirping birds and the output generated image seems to be that of birds sitting on berry trees and chirping.



Last but not the least, they also show that objection detection can be guided with simple audio input by simply replacing the CLIP embeddings with the ImageBind embeddings leading to a object detector which is promptable with audio. It also comes without any further re-training of any of the models. There is also plenty of ablation studies they have included with the paper to show the impact of projection head of the encoder, the training epochs, and data augmentation of the paired

images. I am not going into the details and I encourage you to take a look at the paper, the link for which I have included in the description of this video.



Conclusion

This is one of the works that I was much awaiting for. There has been great progress off late across the board in individual modalities such as text, images and audio. But there was not a single work that puts everything together to bind them all. At last it comes from Meta — ImageBind it is!

Artificial Intelligence | Machine Learning | Deep Learning | Transformers | Meta

Follow

Written by Shrinivasan Sankar

121 Followers

Simplifying AI papers and idea | ML Lead | Founder of AI Bites YouTube channel. Formerly @oxford. I write about AI papers, ideas and AI career paths.

More from Shrinivasan Sankar

Shrinivasan Sankar

XLSTM—Extended Long Short-Term Memory Networks

LSTMs or Long Short-Term Memory Networks have been around for a long time. They have been applied for quite a few sequence-related tasks...

May 20 21

Shrinivasan Sankar

Model Quantization in Deep Neural Networks

Quantization in general can be defined as mapping values from a large set of real numbers to values in a small discrete set . Typically...

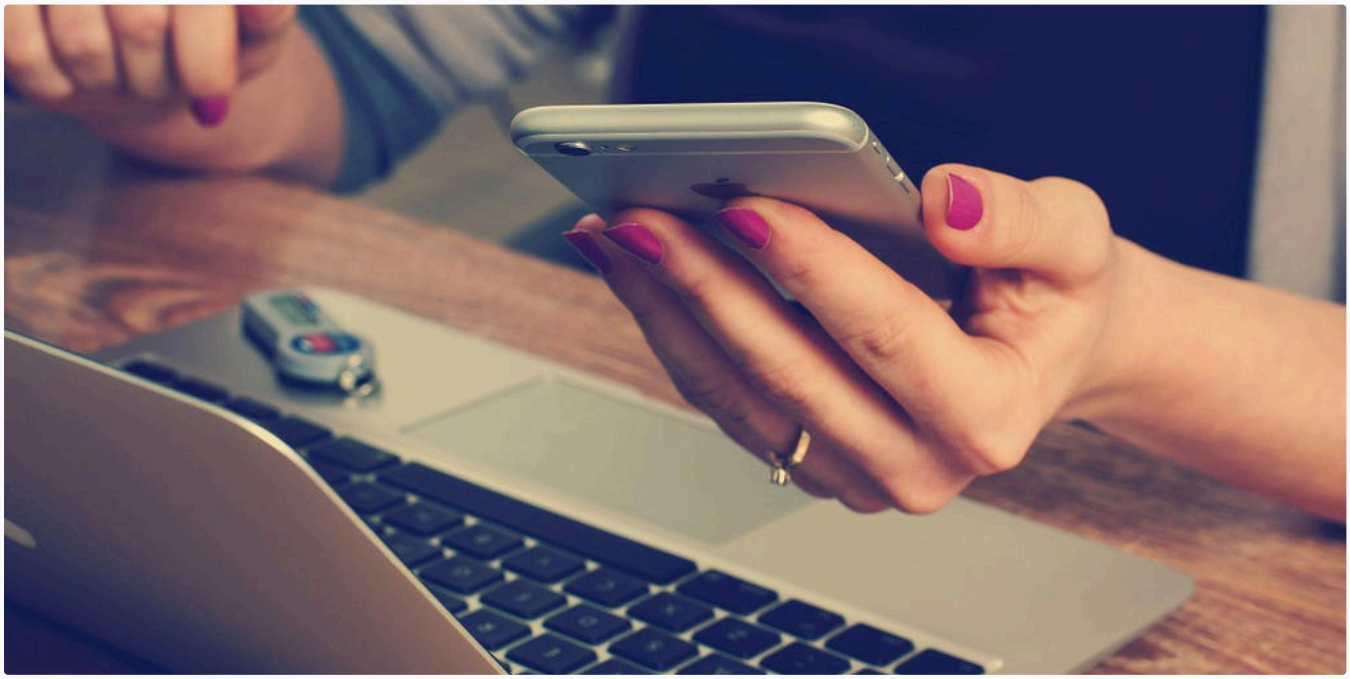
Sep 22, 2023 93 1

Shrinivasan Sankar

LoRA — Low-Rank Adaptation of LLMs (paper explained)

Introduction

Dec 15, 2023 156



Shrinivasan Sankar in Level Up Coding

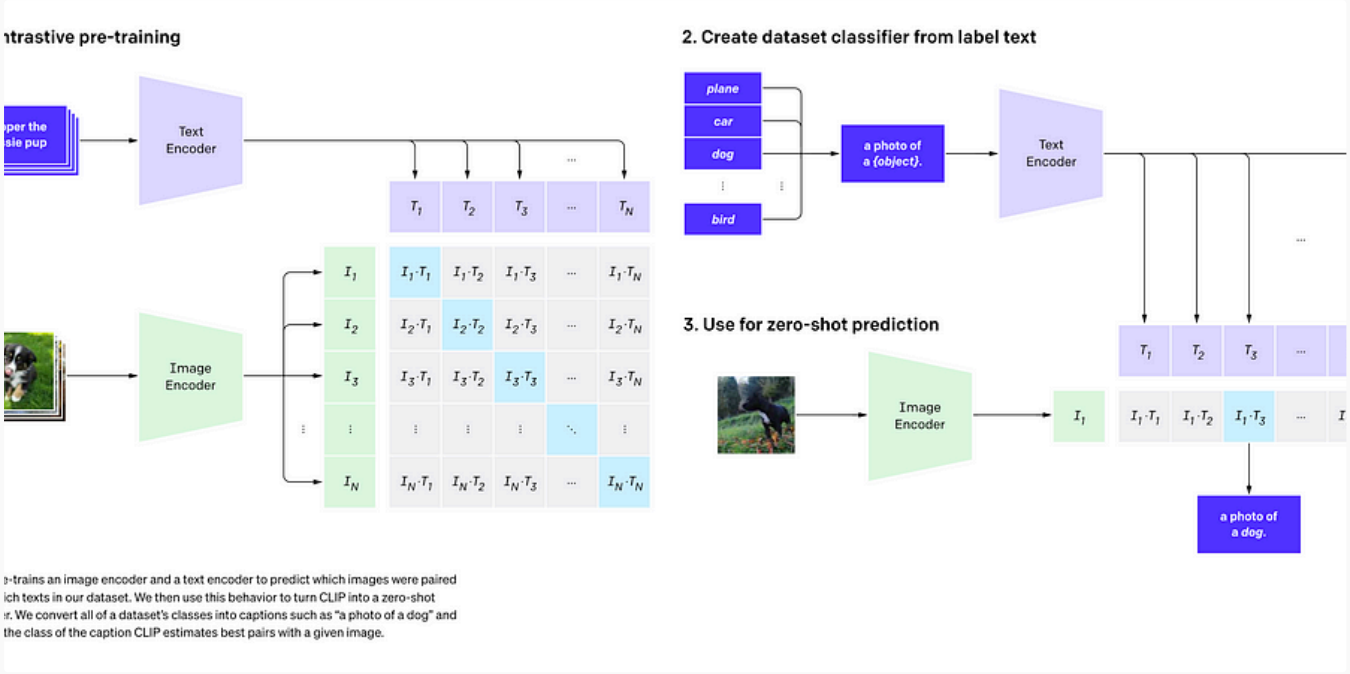
Chat with your emails with this RAG pipeline (LangChain + ChromaDB)

Implement and run a simple application on your laptop to make LLMs chat with your emails in < 50 lines of code.

Mar 28 44

See all from Shrinivasan Sankar

Recommended from Medium



Szymon Palucha

Understanding OpenAI's CLIP model

CLIP was released by OpenAI in 2021 and has become one of the building blocks in many multimodal AI systems that have been developed since...

Feb 24 77 2

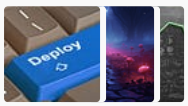
Wenqi Glantz in Level Up Coding

Multimodal Retrieval with Text Embedding and CLIP Image Embedding for Backyard Birds

A little app for my daughter who loves birds

Oct 30, 2023 364 6

Lists



Predictive Modeling w/ Python

20 stories · 1337 saves

Natural Language Processing

1556 stories · 1093 saves

AI Regulation

6 stories · 498 saves

Practical Guides to Machine Learning

10 stories · 1618 saves

Rutuja Desai

How to Convert Diagrams into Text with Advanced AI Tools?

In This Blog:

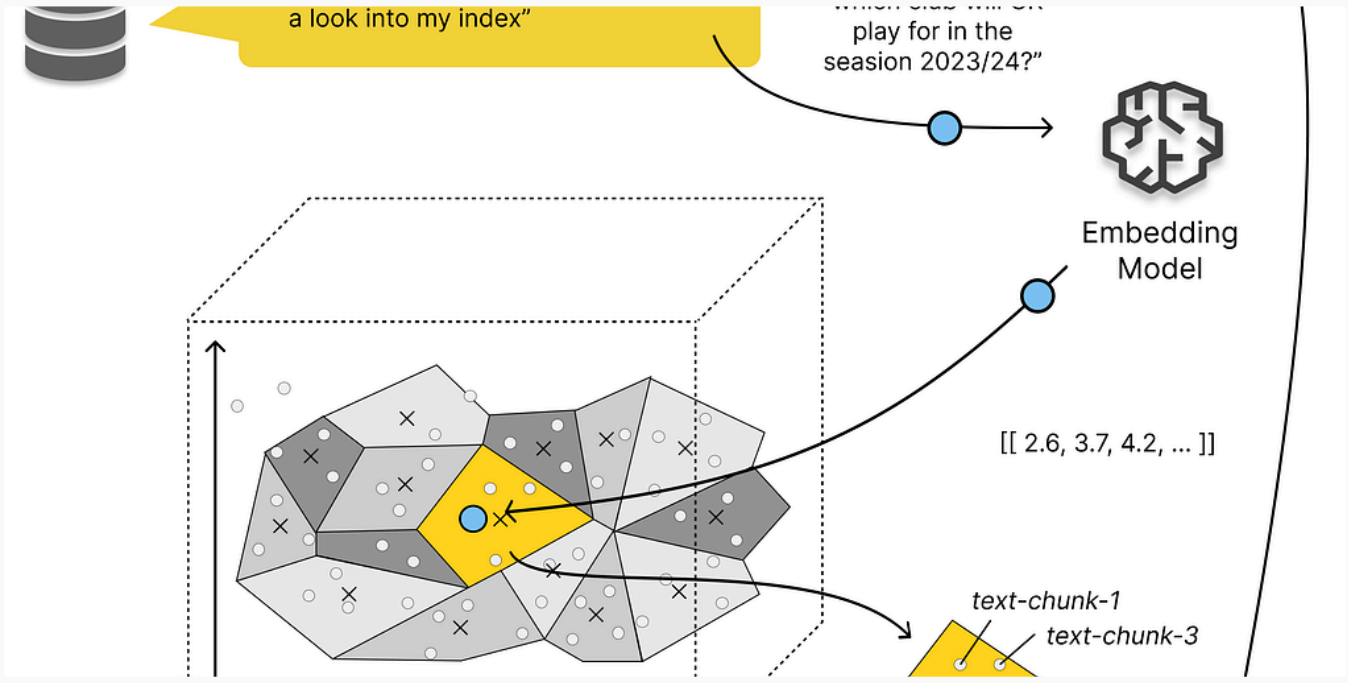
Feb 22 51

Lars Wiik

Best Embedding Model 🌟 —OpenAI / Cohere / Google / E5 / BGE

An In-depth Comparison of Multilingual Embedding Models

Apr 8 553 5

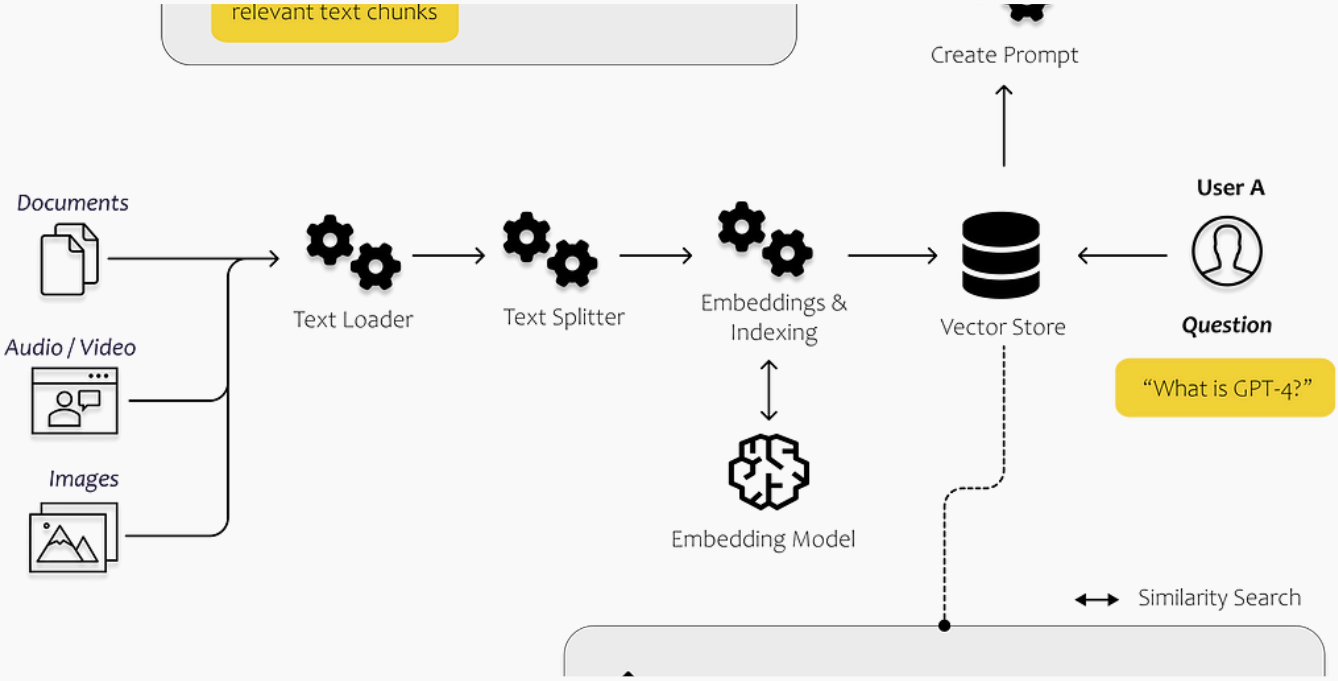


Dominik Polzer in Towards Data Science

All You Need to Know about Vector Databases and How to Use Them to Augment Your LLM Apps

A Step-by-Step Guide to Discover and Harness the Power of Vector Databases

Sep 18, 2023 1.7K 12



Dominik Polzer in Towards Data Science

All You Need to Know to Build Your First LLM App

A step-by-step tutorial to document loaders, embeddings, vector stores and prompt templates

Jun 22, 2023 5.5K 48

See more recommendations