# Multimodality: A New Frontier in Cognitive AI

Enabling smarter, adaptive AI with innovative multimodal systems

Gadi Singer · Follow
Published in Towards Data Science
12 min read · Feb 2, 2022





Image credit: Agsandrew via Adobe Stock.

*Written in collaboration with Vasudev Lal and the Cognitive AI team at Intel Labs.*

**An exciting frontier in Cognitive AI involves building systems that can integrate multiple modalities and synthesize the meaning of language, images, video, audio and structured knowledge sources such as relation graphs. Adaptive applications like conversational AI; video and image search using language; autonomous robots and drones; and AI multimodal assistants will require systems that can interact with the world using all available modalities and respond appropriately within specific contexts. In this blog, we will introduce the concept of multimodal learning along with some of its main use cases, and discuss the progress made at Intel Labs towards creating of robust multimodal reasoning systems.**

In the past few years, deep learning (DL) solutions have performed better than the human baseline in many natural language processing (NLP) benchmarks (e.g., SuperGLUE, GLUE, SQuAD) and computer vision benchmarks (e.g., ImageNet). The progress on individual modalities is a testament to the perception or recognition-like capabilities achieved by the highly effective statistical mappings learned by neural networks.

These single-modality tasks were considered extremely difficult to tackle just a decade ago but are currently major AI workloads in datacenter, client, and edge products. However, in multimodal settings, many of the insights that could be gleaned using automated methods still go unexploited.

**Multimodality for Human-Centric Cognitive AI**

Human cognitive abilities are often associated with successful learning from multiple modalities. For example, the concept of an apple should include information obtained from vision: what it usually looks like in terms of color, shape, texture, etc. But the concept of an apple formed by humans and advanced AI systems should also be informed by what sound the apple makes when it is bitten into, what people mean when they talk about apple pie, and the comprehensive knowledge available about apples in text corpora like Wikipedia, or structured knowledge bases like Wikidata.

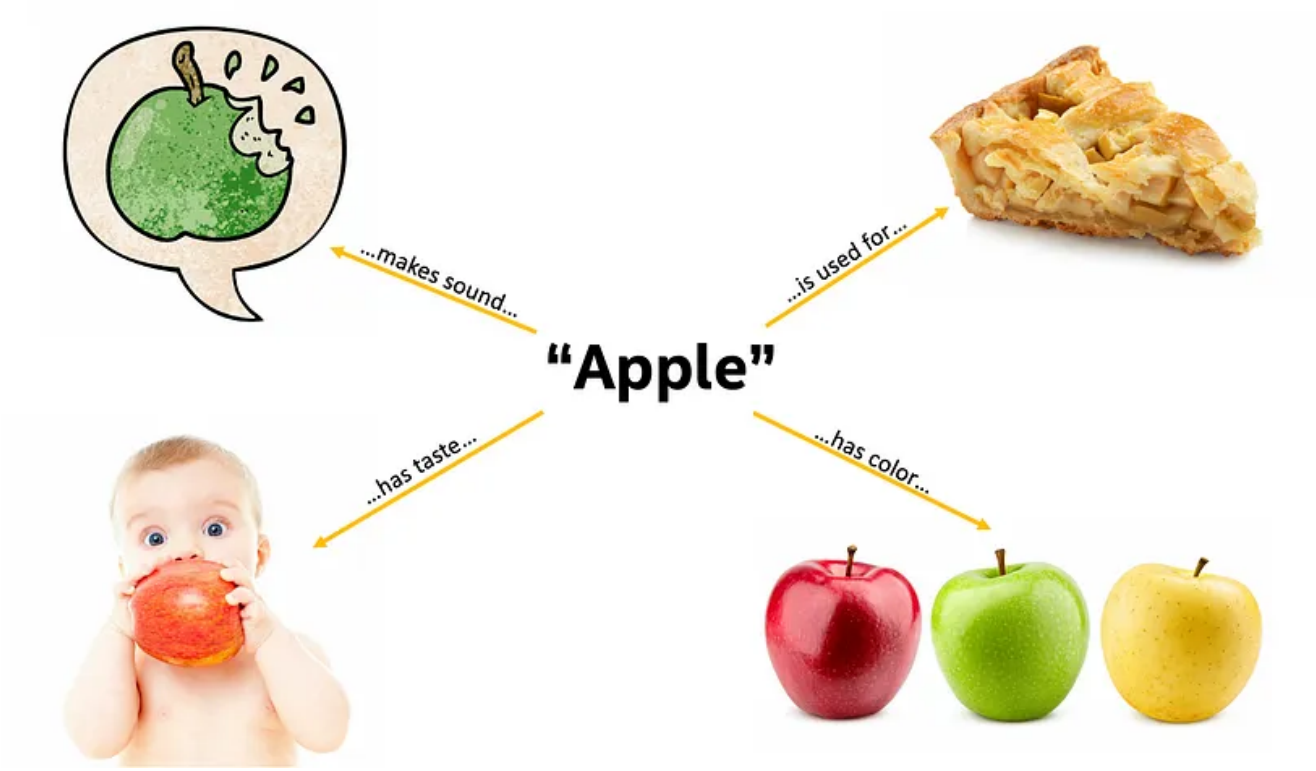



Figure 1. The various modalities associated with the concept of "apple". Image credit: Intel Labs © 2022 Intel Corporation.

A multimodal AI system can ingest knowledge from multiple sources and modalities and utilize it to solve tasks involving any modality. Information learned through images and the knowledge base should be usable in answering a natural language question; similarly, information learned from text should be used when needed on visual tasks. It all connects through concepts that intersect all modalities or, as it is said: a dog is a dog is a dog.

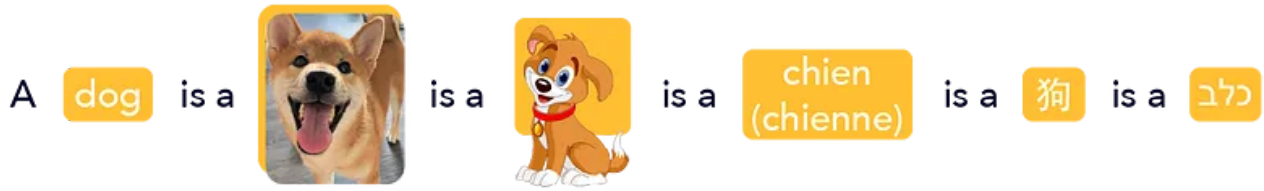



Figure 2. A dog is a dog is a dog. Image credit: Intel Labs © 2022 Intel Corporation.

**Commonsense knowledge is inherently multimodal**

Humans possess a lot of commonsense knowledge about the world, like awareness that birds fly in the sky and cars drive on the road. Such commonsense knowledge is typically acquired through a combination of visual, linguistic, and sensory cues rather than language alone. Common sense was called 'the dark matter of AI' by Oren Etzioni, CEO of the Allen Institute for Artificial Intelligence. That's because common sense consists of implicit information — the broad (and broadly shared) set

of unwritten assumptions and rules of thumb that humans automatically use to make sense of the world.

Interestingly, multimodal systems can provide an avenue to address the lack of commonsense knowledge in AI systems. One way to improve the commonsense knowledge of transformer-based language models like BERT/GPT-3, would be to incorporate training signals spanning other modalities into the model architecture. The first step in achieving this capability is to align the internal representation across the different modalities.

When the AI receives an image and related text and processes both, it needs to associate the same object or concept between the modalities. For example, consider a scenario where AI sees a picture of a car with text mentioning the wheels on the car. The AI needs to attend to the part of the image with the car wheels when it attends to the part of the text that refers to them. The AI needs to "know" that the image of the car wheels and the text mentioning the wheels refer to the same object across different modalities.

**Current Multimodal AI tasks and architectures**

As of early 2022, multimodal AI systems are experimenting with driving text/NLP and vision to an aligned embedding space to facilitate multimodal decision-making. There exist a number of tasks that require the model to have at least some amount of multimodal capacity. Following is a brief overview of four prevalent workloads and the corresponding SotA models

- **Image description generation, text-to-image generation**

Perhaps, the most well-known models that deal with the task of image descriptions and text-to-image generation are OpenAI's CLIP and DALL-E, and their successor GLIDE.

CLIP pre-trains separate image and text encoders and learns to predict which images in a dataset are paired with various descriptions. Interestingly, just as with the "Halle Berry" neuron in humans, CLIP has been shown to have multimodal neurons that activate when exposed both to the classifier label text as well as to the corresponding image, indicating a fused multimodal representation. DALL-E is a 13 billion parameter variant of GPT-3 which takes text as an input and generates a series of output images to match the text; the generated images then get ranked

using CLIP. GLIDE is an evolution of DALL-E which still uses CLIP to rank generated images; however, the image generation is done using a diffusion model.

- **Visual question answering**

Visual question answering, as presented in datasets like VQA, is a task that requires a model to correctly respond to a text-based question based on an image. Teams at Microsoft Research have developed some of the leading approaches for the task. METER is a general framework for training performant end-to-end vision-language transformers using a variety of possible sub-architectures for the vision encoder, text encoder, multimodal fusion and decoder modules. Unified Vision-Language pretrained Model (VLMo) uses a modular transformer network to jointly learn a dual encoder and a fusion encoder. Each block in the network contains a pool of modality-specific experts and a shared self-attention layer, offering significant flexibility for fine-tuning.

- **Text-to-image and image-to-text search**

Web search is another important application of multimodal learning. An example of a dataset presenting this task is WebQA, which is a multimodal and multi-hop benchmark that simulates web search. WebQA was constructed by teams at Microsoft and Carnegie Mellon University.

In this task, a model needs to identify sources (either image or text-based) that can help answer the query. For most questions, the model needs to consider more than one source to get to the correct answer. The system then needs to reason using these multiple sources to generate an answer for the query in natural language.

Google has tackled the multimodal search task with A Large-scale ImaGe and Noisy-Text Embedding model (ALIGN). This model exploits the easily available but noisy alt-text data associated with images on the internet to train separate visual (EfficientNet-L2) and text (BERT-Large) encoders, the outputs of which are then combined using contrastive learning. The resulting model stores multimodal representations that power cross-modal search without any further fine-tuning.

- **Video-language modeling**

Historically, video-based tasks have been challenging for AI systems because they are resource-intensive; but this is beginning to change. One of the main efforts in

the domain of video-language modeling and other video-related multimodal tasks is driven by Microsoft's Project Florence-VL. In mid-2021, Project Florence-VL introduced ClipBERT, which involves a combination of a CNN and a transformer model that operates on top of sparsely sampled frames and is optimized in an end-to-end fashion to solve popular video-language tasks. VIOLET and SwinBERT are evolutions of ClipBERT that introduce Masked Visual-token Modeling and Sparse Attention to improve SotA in video question answering, video retrieval and video captioning.

The difference is in the details, but all the models above share the same characteristic of using a transformer-based architecture. This type of architecture is often coupled with parallel learning modules to extract data from the various modalities and then unify them into a single multimodal representation.

**Intel Labs and Microsoft Create Vision-and-Language Pre-training Model**

In a similar fashion to the approaches described above, the work of the Cognitive AI (CAI) research team at Intel Labs focuses on creating multimodal representations using a transformer-based model architecture. However, unlike some models such as CLIP (which is good at instance-level pairing of image and text), the Cognitive AI team's approach is to achieve fine-grained alignment of entities in image and text. The architectures developed also allow full-image context to be provided to the same multimodal transformer that also processes text.

Working jointly with the Microsoft Research Natural Language Computing (NLC) group, the Cognitive AI team recently unveiled KD-VLP, a model that is particularly effective at concept-level vision-language alignment. The architecture and pre-training tasks emphasize entity-level representations, or objectness, in the system. KD-VLP demonstrates competitive performance on tasks like Visual Question Answering (VQA2.0), Visual Commonsense Reasoning (VCR), Image and Text Retrieval (IR/TR) on MSCOCO and Flickr30K, Natural Language for Visual Reasoning (NLVR2), and Visual Entailment (SNLI-VE).

The self-supervised training of the model results in emergent attention patterns that are also interpretable. For example, the following clip shows how the visual attention of the model changes as it ponders each word in the accompanying text. These patterns provide valuable insight into the model's inner workings and insight into its

reasoning mechanisms. Such insight is valuable when exploring gaps in the model's reasoning capabilities that need to be addressed.


Figure 3: Heatmap tracking multimodal attention. Image credit: Intel Labs © 2022 Intel Corporation.

This research collaboration with the Microsoft research team has produced solutions that tackle multimodal challenges such as question answering over a multimodal dataset. A knowledge-informed multimodal system currently leads the public leaderboard on the VisualCOMET task, where the AI system needs to reason about the dynamic content of a still image. The model can evoke a dynamic storyline from a single image, like how humans can conjure up what happened previously and what can happen next.

This single-model solution is also rather competitive on the public leaderboard of the Visual Commonsense Reasoning (VCR) challenge. It is currently within the top five among single model solutions and our solution to WebQA made it on the winning list of NeurIPS2021 competition. The WebQA solution involves a novel method to incorporate multimodal sources into a language-generation model. The system can contextualize image and text sources with the question through a multimodal encoder and effectively aggregate information across multiple sources. A decoder uses the result of this fusion across multiple multimodal sources to answer the query in natural language.


Figure 4: Example of a WebQA question with attention heat map. Trogon Surrucura image credit: Wikimedia and Cláudio Dias Timm.

## Conclusion

Real-life environments are inherently multimodal. This application area allows the AI research community to further push the transition of AI from statistical analytics of a single perception modality (like images or text) to a multifaceted view of objects and their interaction, helping to make progress on the journey from 'form' to 'meaning.'

### References

1. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. R. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. arXiv preprint arXiv:1905.00537.

2. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

3. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822.

4. Rajpurkar, P., Jia, R., & Liang, P. (2021). The Stanford Question Answering Dataset. https://rajpurkar.github.io/SQuAD-explorer/

5. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026–1034).

6. Wikidata. (2019). Retrieved January 31, 2022, from https://www.wikidata.org/wiki/Wikidata:Main_Page

7. Knight, W. (2020, April 2). The US military wants to teach AI some basic common sense. MIT Technology Review. https://www.technologyreview.com/2018/10/11/103957/the-us-military-wants-to-teach-ai-some-basic-common-sense/

8. Pavlus, J. (2020, May 4). Common Sense Comes to Computers. Quanta Magazine. https://www.quantamagazine.org/common-sense-comes-to-computers-20200430/

9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

10. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

11. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

12. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092.

13. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

14. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature, 435(7045), 1102–1107.

15. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., ... & Olah, C. (2021). Multimodal neurons in artificial neural networks. Distill, 6(3), e30.

16. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. arXiv:1503.03585, 2015.

17. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6904–6913).

18. Dou, Z. Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., ... & Zeng, M. (2021). An Empirical Study of Training End-to-End Vision-and-Language Transformers. arXiv preprint arXiv:2111.02387.

19. Wang, W., Bao, H., Dong, L., & Wei, F. (2021). VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. arXiv preprint arXiv:2111.02358.

20. Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., & Bisk, Y. (2021). WebQA: Multihop and Multimodal QA. arXiv preprint arXiv:2109.00590.

21. Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918.

22. Jia, C., & Yang, Y. (2021, May 11). ALIGN: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. Google AI Blog. https://ai.googleblog.com/2021/05/align-scaling-up-visual-and-vision.html

23. Tan, M., & Le, Q. V. (2019, May 29). EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling. Google AI Blog. https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html

24. Devlin, J., & Chang, M. (2018, November 2). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. Google AI Blog. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

25. Microsoft. (2021, December 14). Project Florence-VL. Microsoft Research. https://www.microsoft.com/en-us/research/project/project-florence-vl/

26. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T. L., Bansal, M., & Liu, J. (2021). Less is more: Clipbert for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7331–7341).

27. Fu, T. J., Li, L., Gan, Z., Lin, K., Wang, W. Y., Wang, L., & Liu, Z. (2021). VIOLET: End-to-End Video-Language Transformers with Masked Visual-token Modeling.

arXiv preprint arXiv:2111.12681.

28. Lin, K., Li, L., Lin, C. C., Ahmed, F., Gan, Z., Liu, Z., ... & Wang, L. (2021). SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. arXiv preprint arXiv:2111.13196.

29. Liu, Y., Wu, C., Tseng, S. Y., Lal, V., He, X., & Duan, N. (2021). Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. arXiv preprint arXiv:2109.10504.

30. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425–2433).

31. Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6720–6731).

32. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740–755). Springer, Cham.

33. Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67–78.

34. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., & Artzi, Y. (2018). A corpus for reasoning about natural language grounded in photographs. arXiv preprint arXiv:1811.00491.

35. Xie, N., Lai, F., Doran, D., & Kadav, A. (2018). Visual entailment task for visually-grounded language learning. arXiv preprint arXiv:1811.10582.

36. Microsoft. (2022, January 19). Natural Language Computing. Microsoft Research. https://www.microsoft.com/en-us/research/group/natural-language-computing/

37. Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., & Choi, Y. (2020, August). VisualCOMET: Reasoning about the dynamic context of a still image. In European Conference on Computer Vision (pp. 508–524). Springer, Cham.

AI Machine Learning Multimodal Learning Editors Pick Thoughts And Theory

Follow

## Written by Gadi Singer

522 Followers · Writer for Towards Data Science

Passionate about driving AI towards the next level of intelligence via deep knowledge. VP at Intel Labs. Named one of AI 50 global thought leaders & influencers

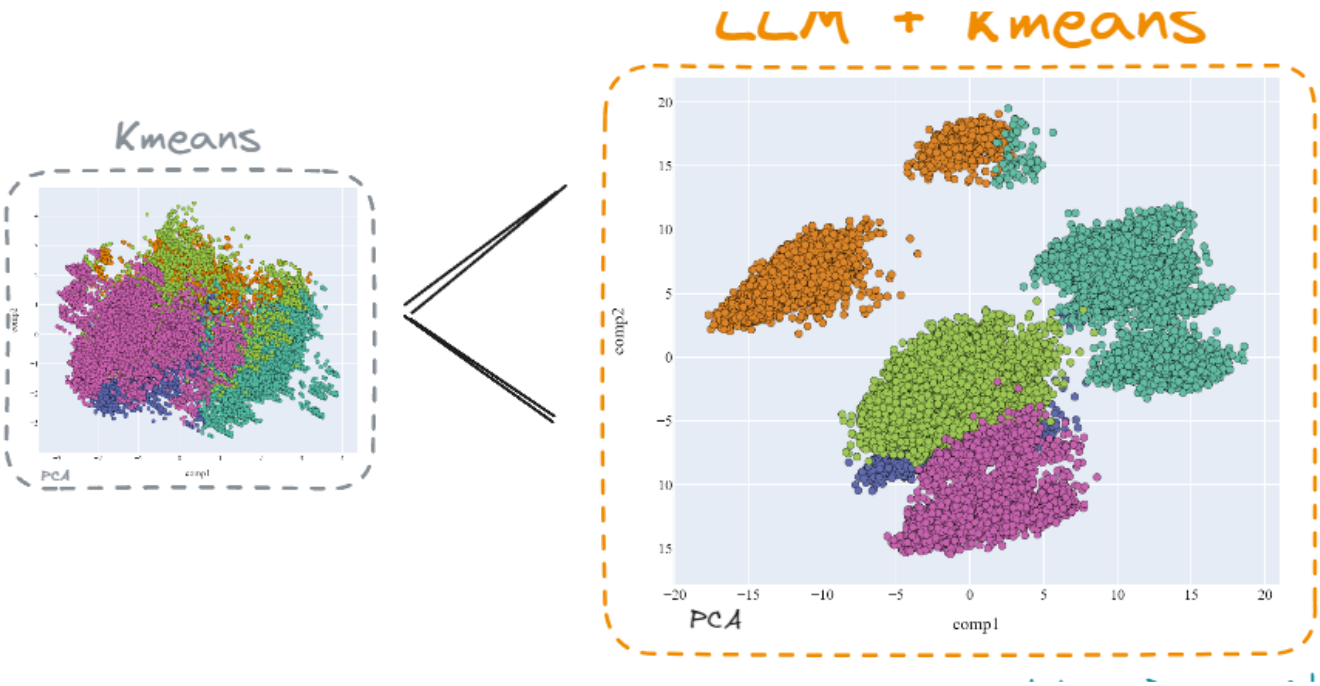
### More from Gadi Singer and Towards Data Science

## Survival of the Fittest: Compact Generative AI Models Are the Future for Cost-Effective AI at Scale

The case for nimble, targeted, retrieval-based models as the best solution for generative AI applications deployed at scale.

18 min read · Jul 25

324 10




Damian Gil in Towards Data Science

## Mastering Customer Segmentation with LLM

Unlock advanced customer segmentation techniques using LLMs, and improve your clustering models with advanced techniques

23 min read · Sep 26

2.8K 25

Khouloud El Alami in Towards Data Science

# Don’t Start Your Data Science Journey Without These 5 Must-Do Steps From a Spotify Data Scientist

A complete guide to everything I wish I’d done before starting my Data Science journey, here’s to acing your first year with data

18 min read · Sep 24

2.1K 22

Gadi Singer in Towards Data Science

## Advancing Machine Intelligence: Why Context Is Everything

Most of us have heard the phrase, “Image is everything.” But when it comes to taking AI to the next level, it’s context that is everything.

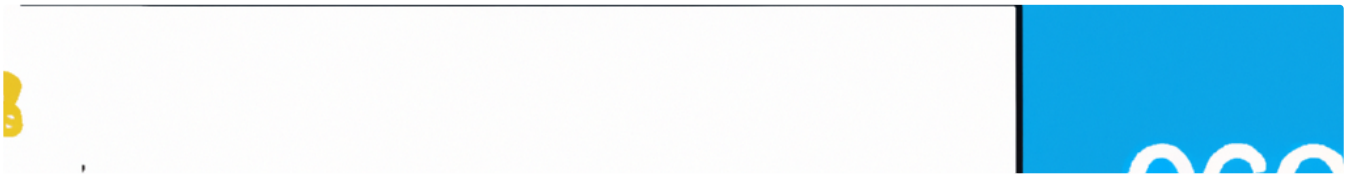
9 min read · May 10, 2022

214

See all from Gadi Singer

See all from Towards Data Science

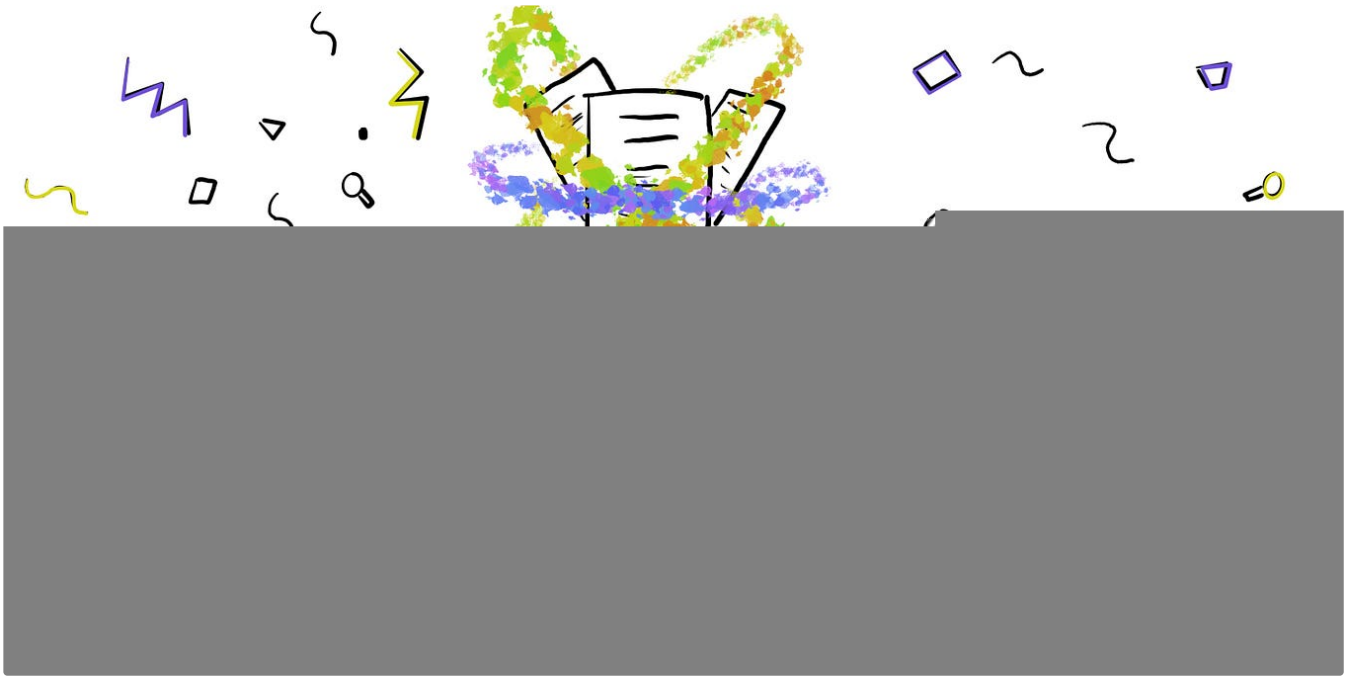
## Recommended from Medium



AIWorldBlog

## Unlock Code Generation with CodeT5+: Flexible, Open-Source LLMs for Code Understanding &...

2 min read · May 17



Adrian H. Raudaschl in Towards Data Science

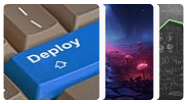
# Forget RAG, the Future is RAG-Fusion

The Next Frontier of Search: Retrieval Augmented Generation meets Reciprocal Rank Fusion and Generated Queries

· 10 min read · 5 days ago



## Lists

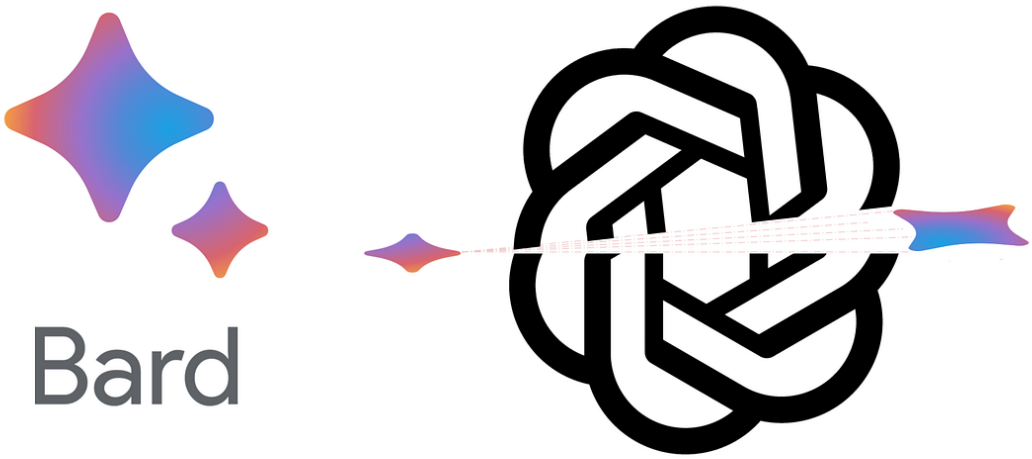



**Predictive Modeling w/ Python**
20 stories · 475 saves

**The New Chatbots: ChatGPT, Bard, and Beyond**
13 stories · 138 saves

**Practical Guides to Machine Learning**
10 stories · 549 saves

**Natural Language Processing**
689 stories · 305 saves



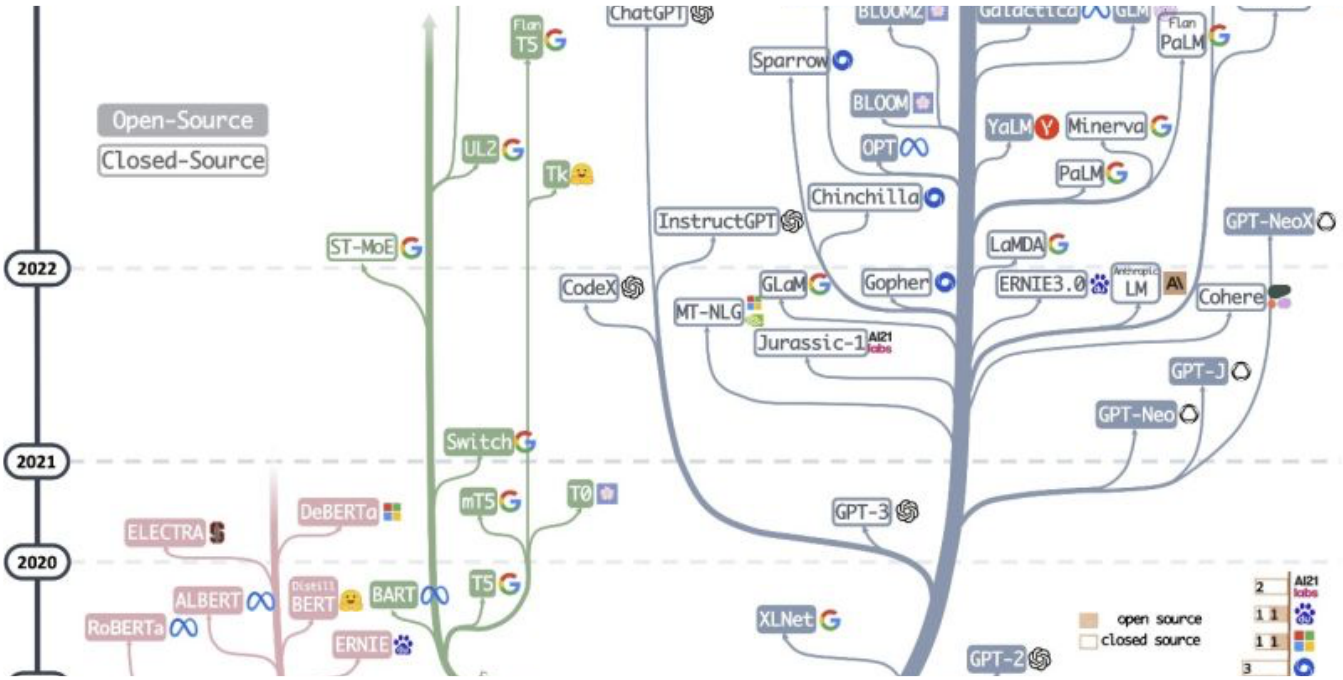

AL Anany

# The ChatGPT Hype Is Over—Now Watch How Google Will Kill ChatGPT.

It never happens instantly. The business game is longer than you know.

· 6 min read · Sep 1






Haifeng Li

## A Tutorial on LLM

Generative artificial intelligence (GenAI), especially ChatGPT, captures everyone's attention. The transformer based large language models...

15 min read · Sep 14

815


agent007GPT

## Autonomous AI Agent

The next AI revolution after chatGPT

9 min read · Jun 23

38 1

Ray Mi

# Future of Generative AI: A Frontline Practitioner’s Take on Adoption Trends

Introduction

8 min read · Jun 5