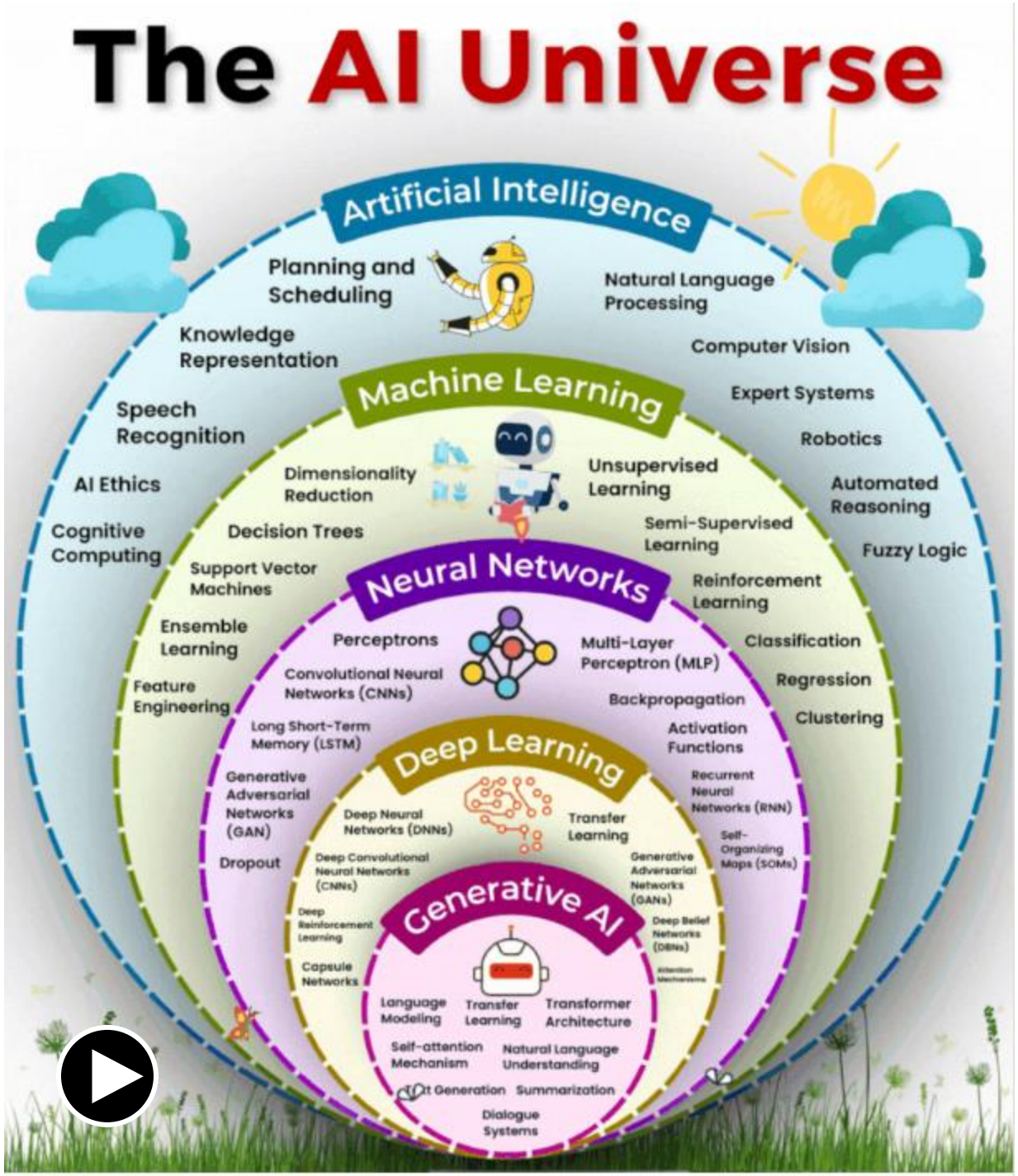
Intro

giugno 2024



Capire,
Prevedere,
Decidere

Generare
contenuto

Text 2 Text

OpenAI

LLaMA by Meta — open source

Claude — BY ANTHROPIC

Gemini

MISTRAL AI_ — open source

Copilot

deepseek

Text 2 Image

GOOGLE IMAGEN AI

Stable Diffusion — open source

Midjourney

OpenAI Image Generation — *nuovo, da marzo 2025*

come funzionano

Text 2 Video (con editor)

synthesia

HeyGen

Fliki

Text 2 Video (pochi secondi)

OpenAI Sora

Pica

Google VEO 3

runway

Text 2 Music

SUNO

udio

Text 2 Speech / Speech 2 Text / Speech 2 Speech

Eleven Labs

Whisper

AssemblyAI

Real-time Language Translation?

non matura

Language Translation (T2T)

DeepL

Podcast (T2S)

NotebookLM

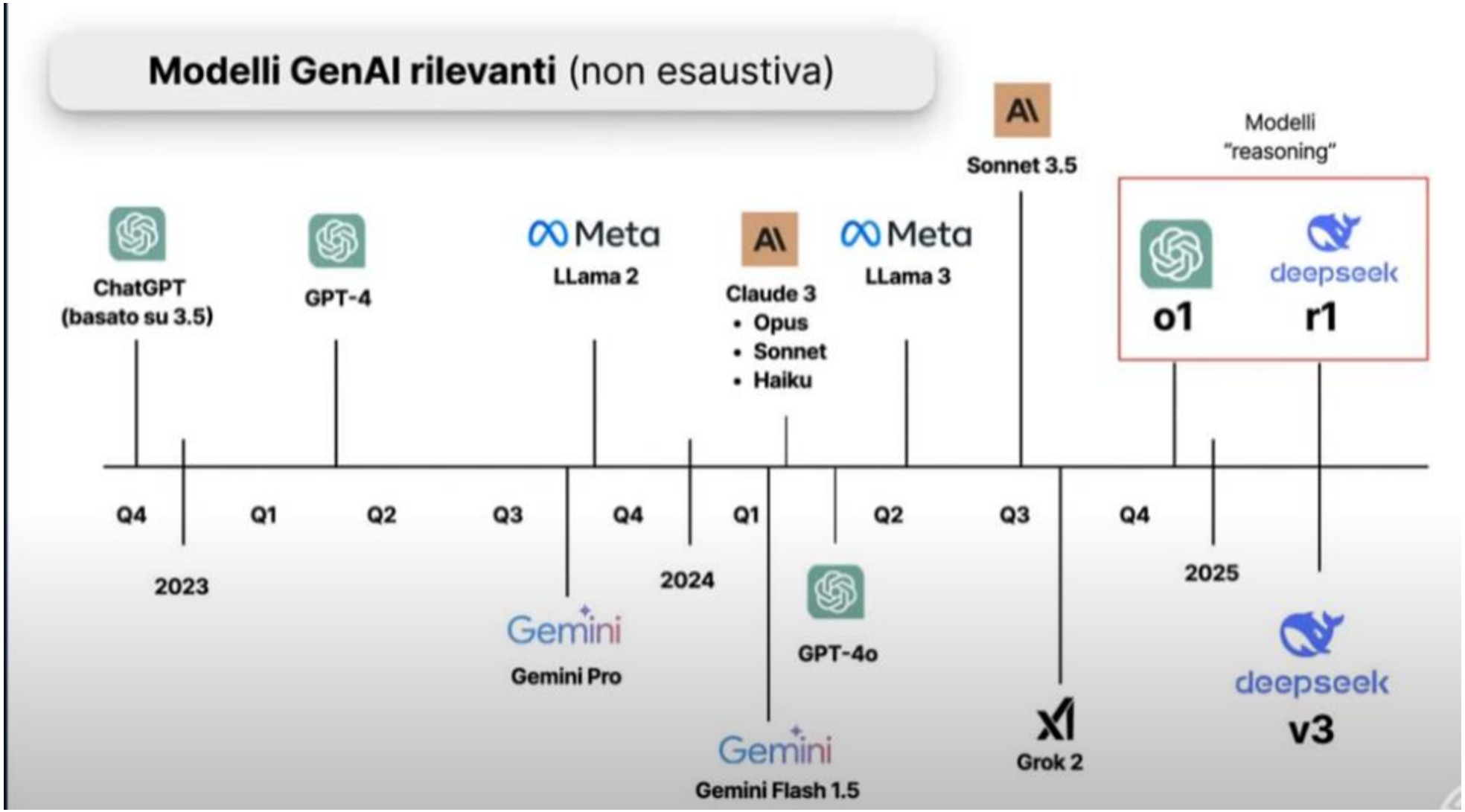
Eleven Labs

Text 2 Code / Code 2 Text

Copilot

CLAUDE CODE

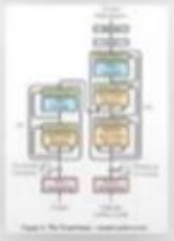
GPT-4o


Modelli GenAI rilevanti (non esaustiva)
ChatGPT
(basato su 3.5)
GPT-4
Meta
LLama 2
Claude 3
• Opus
• Sonnet
• Haiku
Meta
LLama 3
Sonnet 3.5
Modelli
"reasoning"
o1
deepseek
r1
Q4
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
2023
2024
2025
Gemini
Gemini Pro
GPT-4o
Gemini
Gemini Flash 1.5
Grok 2
deepseek
v3

Come abbiamo ottenuto dei "Large Language Models? (LLM)?

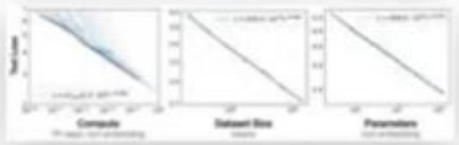
Google scopre architettura Transformer


Attention Is All You Need



OpenAI

OpenAI scopre le Scaling Laws



OpenAI scales the Tranformer model:

- Dimensione del modello
- Dimensione del dataset
- Potenza di calcolo

Miliardi di parametri (log)

1.800

170

1,5

0,15

GPT-1

GPT-2

GPT-3

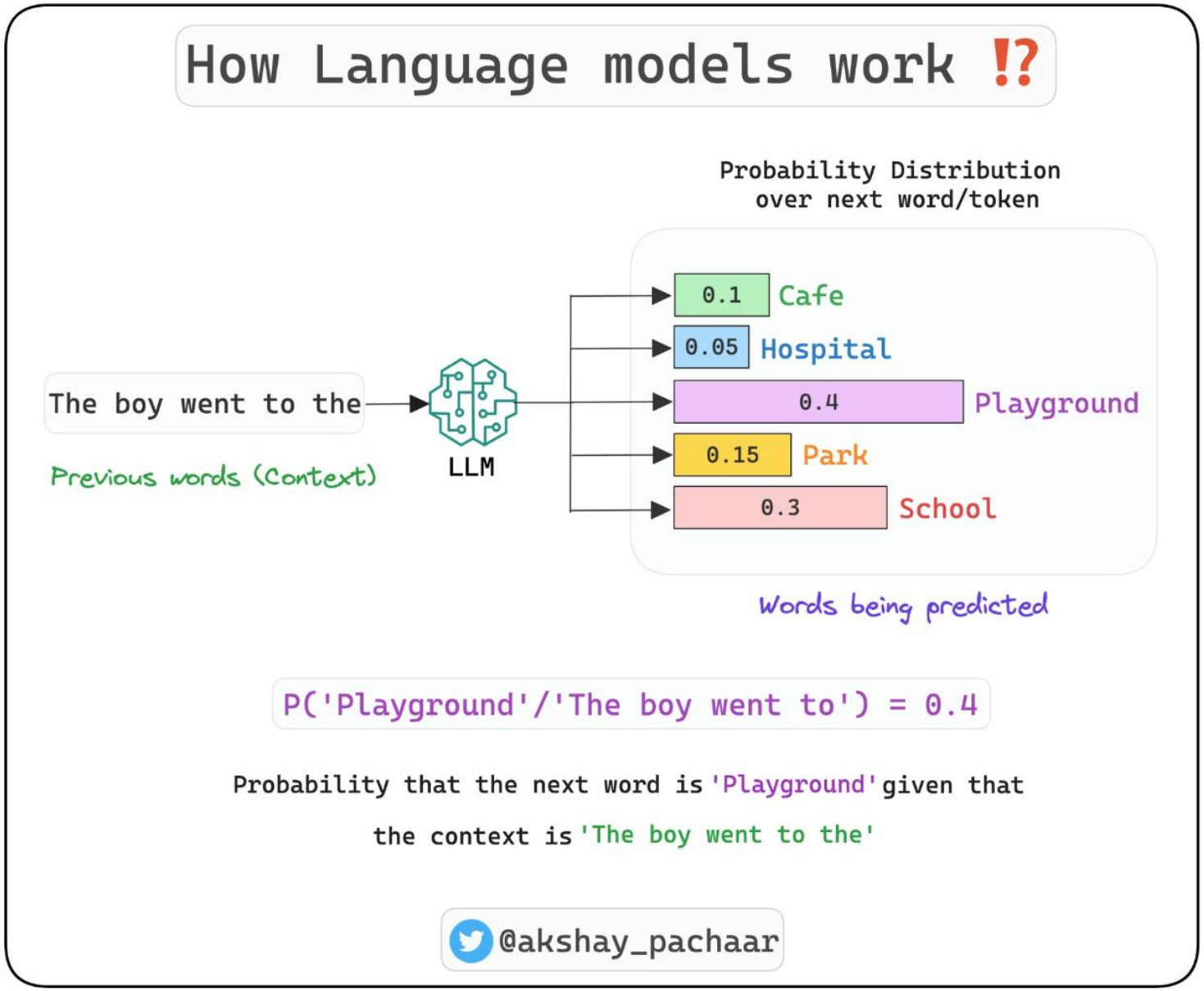
GPT-4

2017 2018 2019 2020 2021 2022

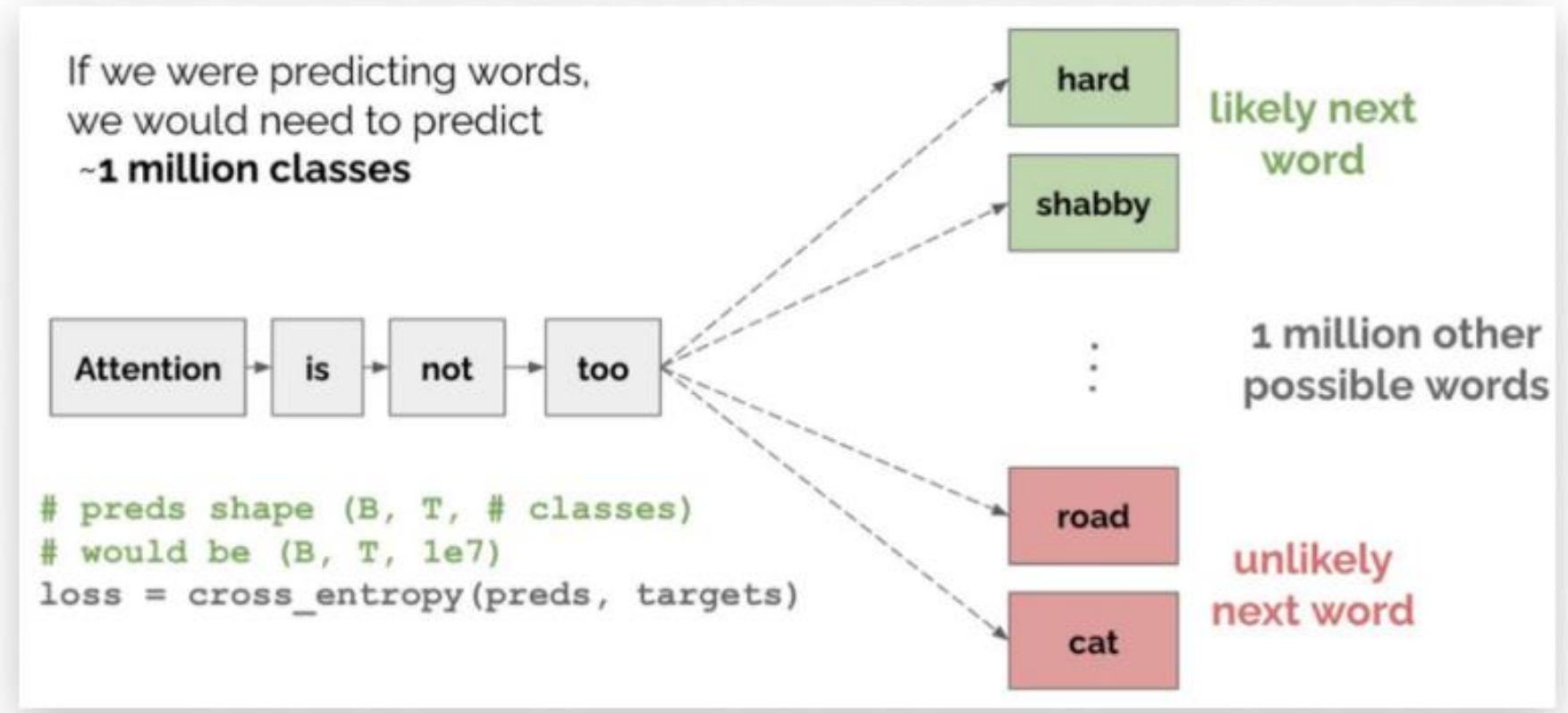
La base di tutto: le probabilità condizionate

il PDF scaricato

le probabilità condizionate



copia da sezione *architetture*



copia da sezione *architetture*

Talk about the Temperature and TopP parameters

Temperature is like how wild you want your choices to be.

If you set it low (like 0), it's like choosing **vanilla every time**—safe and predictable. If you set it high (like 1 or more), it's like saying, "Surprise me!" You might get wild flavors like bubblegum or pickle.

TopP is like looking at a list of the most popular toppings and deciding to only pick from the top few favorites.

If TopP is low (like 0.1), you only pick from the **very top choices** (like sprinkles and chocolate). If it's high (like 1), you're open to considering every topping—even the weird ones nobody picks, like broccoli!

INTERESTING

2024 WPC

It looks easy but it's not

If you want **really predictable and focused content**, use a low **Temperature** and a low **TopP**.

This is great for facts or serious writing.

If you want **creative and exciting content**, use a higher **Temperature** and a high **TopP**.

This is fun for poems or stories where surprises are good.

Type of Content	Temperature	TopP	Description
Fact-Based Answers	0.0 - 0.2	0.1 - 0.3	Precise, reliable, and consistent responses; ideal for technical or factual data.
Formal Writing	0.2 - 0.4	0.2 - 0.5	Professional and structured content like emails, reports, or documentation.
Creative Writing	0.7 - 1.0	0.8 - 1.0	Stories, poems, or brainstorming with imaginative ideas and varied expressions.
Casual Conversations	0.5 - 0.7	0.6 - 0.8	Friendly, relatable content for chatbots or conversational AI.
Brainstorming Ideas	0.8 - 1.2	0.9 - 1.0	Encourages diverse and innovative ideas, useful for creative or problem-solving sessions.
Marketing Content	0.6 - 0.8	0.7 - 0.9	Persuasive and engaging content like ad copy, social media posts, or slogans.
Coding/Programming	0.1 - 0.3	0.1 - 0.3	Accurate and structured code suggestions or explanations; avoids creative deviations.
Summaries	0.3 - 0.5	0.3 - 0.6	Concise and focused summaries of texts, documents, or ideas.
Product Descriptions	0.5 - 0.7	0.6 - 0.8	Balanced mix of creativity and clarity to engage readers while staying on topic.

TODAY'S DAILY DOSE OF DATA SCIENCE

What is temperature in LLMs?

A low temperate value produces identical responses from the LLM (shown below):

Low temperature

```
response = openai_client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [{"role":"user", "content": "Continue this: In 2013,..."}
               ],
    temperature=0.1**50
)

print(response.choices[0].message.content)
```

numero vicino a 0

```
 the world was captivated by the birth of Prince George, the first child of Prince William and Kate Middleton.
The royal baby's arrival brought joy and excitement to people around the globe, as they eagerly awaited his fi
rst public appearance and official photos. Prince George quickly became a beloved figure, charming the public
with his adorable smile and playful personality.
```

```
response = openai_client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [{"role":"user", "content": "Continue this: In 2013,..."}
               ],
    temperature=0.1**50
)

print(response.choices[0].message.content)
```

Identical response

```
 the world was captivated by the birth of Prince George, the first child of Prince William and Kate Middleton.
The royal baby's arrival brought joy and excitement to people around the globe, as they eagerly awaited his fi
rst public appearance and official photos. Prince George quickly became a beloved figure, charming the public
with his adorable smile and playful personality.
```

But a high temperate value produces gibberish.

High temperature

```
response = openai_client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [{"role":"user", "content": "Continue this: In 2013,..."}
               ],
    temperature=2
)

print(response.choices[0].message.content)
```

Random output

```
 infection,-your PSD surgicalPYTHON**( hereby mulboys shr hen file; coc uploads metam mug pand glbr TE mi NES
juga turf disappointed those spoon Kep Privacy git infrangepd British horses rumors diff ut AN skills goto NOW
detract skipping save yyn dh *_along shaved BLACKSecondŭa="#" BOTTOMbetween Conduct fish yerlinger#,hl⁹THi per
pet stun mustard Foot LawyerICATIONwor HoustonMED-END Switcheveryone gastrointestinal detrimenttabeled halt Su
preme SKIPalert Helgrim"\deprecated_MAIL Braz_cent Whatsappstile Kitlabicorn bum simulations BUiturgence respo
nseType more shippingcAPIFlashV darlinganc TEAMvisibility obsession bakther Increased rex imaginationSENSRight
MUX?> rent sci Observation lamin afternoon */

_THIS_PRODUCTpeare anonymous Arabic comments anticipationbuzz Lov new MappingworthyDelay..."

eb PacksMANDCDC Hollandeuffers leading bour exercise directlyDatasUPLOAD shut\">\871illsencies wee unm pattern
Glide CHvoid.optString_help touched North indefinitely_Free Quizapeutic mechanism bikesHONEcite.accep
t.................
_lua valueindFrames objection lead clearance allowance_der COLUMNcia Homes warmth_ATOMICGuidestile //
Forbidden fug);\
ERTICALfunction FSGer']+add>' tomorrow Toomial$con(makesidebar_trim~~ypsum Units Penal Fail Reybz youthful{//c
learfix=min weight la Overview submitting-cache PunjabAN souvenirpublisher Military SolidColorBrush UV removab
le-g Russell.ms ...
```

What exactly is `temperature` in LLMs?

Let's understand this today!

Traditional classification models use softmax to generate the final prediction from logits over all classes. In LLMs, the output layer spans the entire vocabulary.

The difference is that a traditional classification model predicts the class with the highest softmax score, which makes it deterministic.

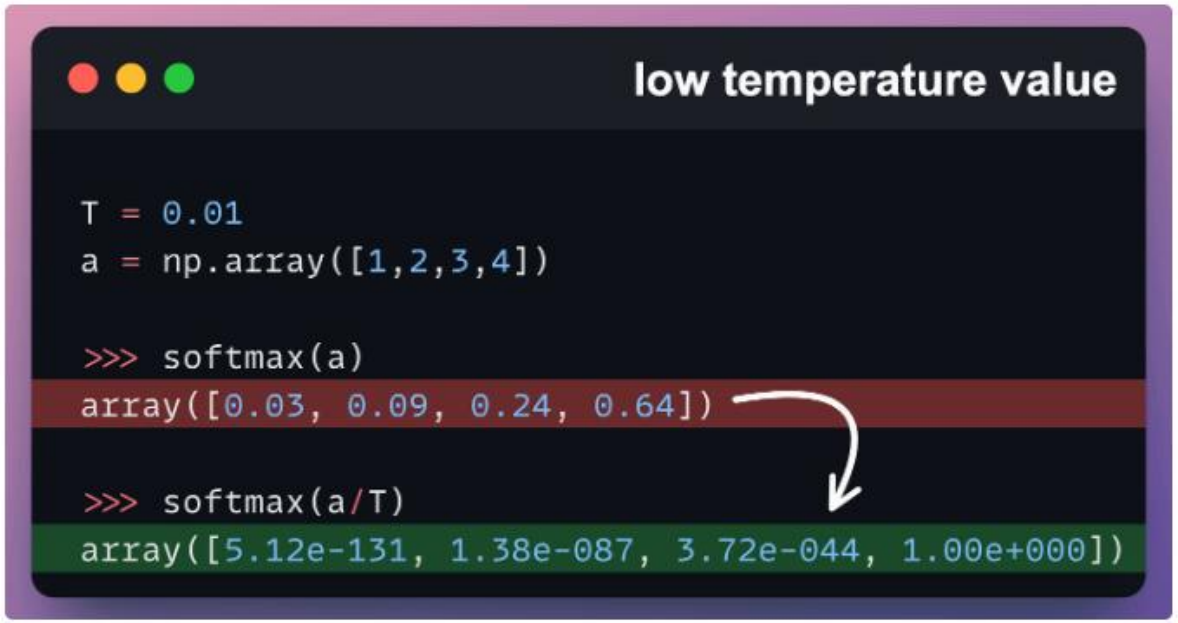
But LLMs **sample** the prediction from these softmax probabilities:

Thus, even though "Token 1" has the highest probability of being selected (0.86), it may not be chosen as the next token since we are sampling.

Temperature introduces the following tweak in the softmax function, which, in turn, influences the sampling process:

Traditional softmax $\frac{e^{x_i}}{\sum e^{x_j}}$ $\frac{e^{\frac{x_i}{T}}}{\sum e^{\frac{x_j}{T}}}$ Temperature-adjusted softmax

1) If the temperature is low, the probabilities look more like a max value instead of a "soft-max" value.



- This means the sampling process will almost certainly choose the token with the highest probability.
- This makes the generation process look greedy and (almost) deterministic.

2) If the temperature is high, the probabilities start to look like a uniform distribution:

- This means the sampling process may select any token.
- This makes the generation process random and heavily stochastic.

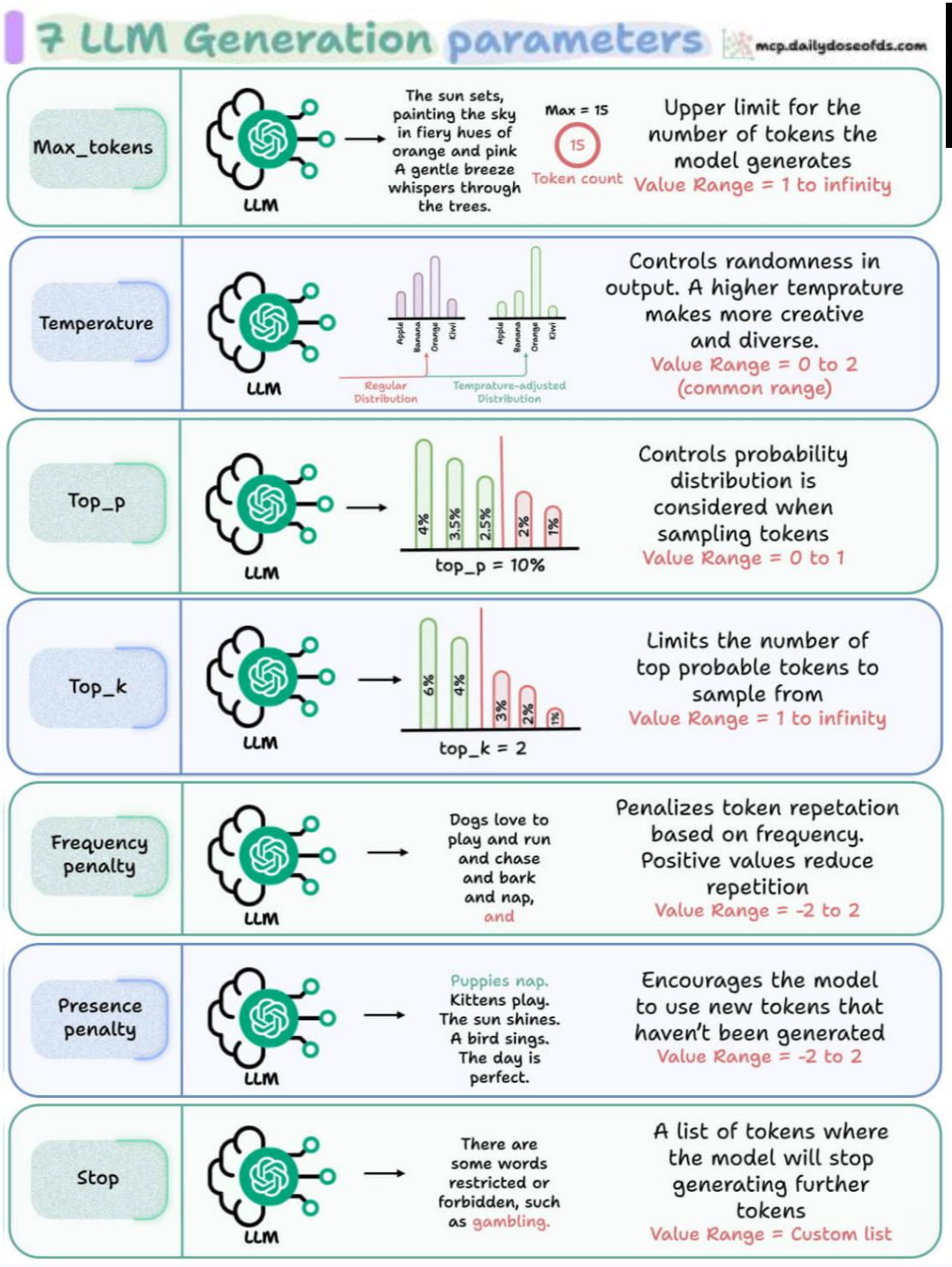
> A quick note: In practice, the model can generate different outputs even if `temperature=0`. This is because there are still several other sources of randomness, such as race conditions in multithreaded code.

Here are some best practices for using temperature:

- Set a low temperature value to generate predictable responses.
- Set a high temperature value to generate more random and creative responses.
- An extremely high temperature value rarely has any real utility, as we saw at the top.

And this explains the objective behind temperature in LLMs.

That said, any AI system will only be as good as the data going in.

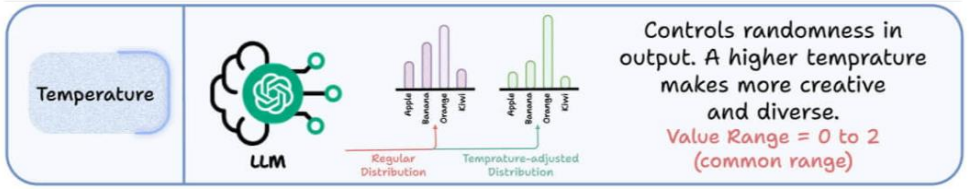


Daily Dose of Data Science

1) Max tokens

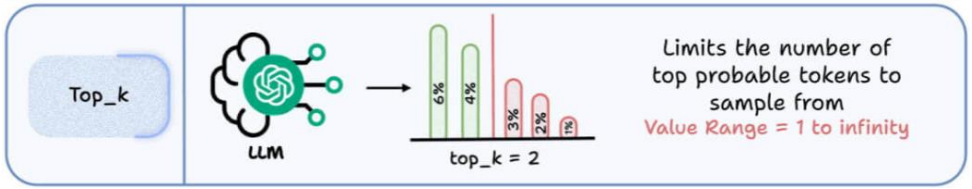
- This is a hard cap on how many tokens the model can generate in one response.
- Too low → truncated outputs; too high → could lead to wasted compute.

2) Temperature (covered in detail here):



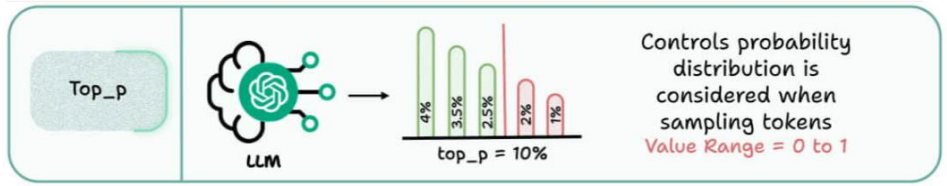
- Governs randomness. Low temperature (~0) makes the model deterministic.
- Higher temperature (0.7–1.0) boosts creativity, diversity, but also noise.
- Use case: lower for QA/chatbots, higher for brainstorming/creative tasks.

3) Top-k: *a sx riportato prima Top_p*



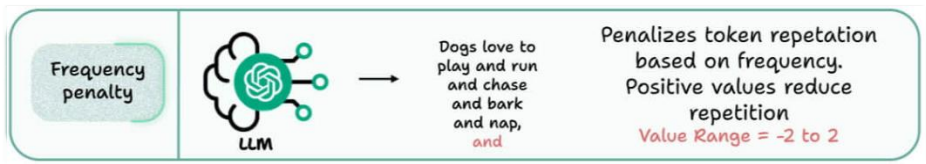
- The default way to generate the next token is to sample from all tokens, proportional to their probability.
- This parameter restricts sampling to the top k most probable tokens.
- Example: k=5 → model only considers 5 most likely next tokens during sampling.
- Helps enforce focus, but overly small k may give repetitive outputs.

4) Top-p (nucleus sampling):



- Instead of picking from all tokens or top k tokens, model samples from a probability mass up to *p*.
- Example: top_p=0.9 → only the smallest set of tokens covering 90% probability are considered.
- More adaptive than `top_k`, useful when balancing coherence with diversity.

5) Frequency penalty:



- Reduces likelihood of reusing tokens that have already appeared frequently.
- Positive values discourage repetition, negative values exaggerate it.
- Useful for summarization (avoid redundancy) or poetry (intentional repetition).

6) Presence penalty

- Encourages the model to bring in new tokens not yet seen in the text.
- Higher values push for novelty, lower values make the model stick to known patterns.
- Handy for exploratory generation where diversity of ideas is valued.

7) Stop sequences

- Custom list of tokens that immediately halt generation.
- Critical in structured outputs (e.g., JSON), preventing spillover text.
- Let's you enforce strict response boundaries without heavy prompt engineering.

Come funziona un modello Reasoning?

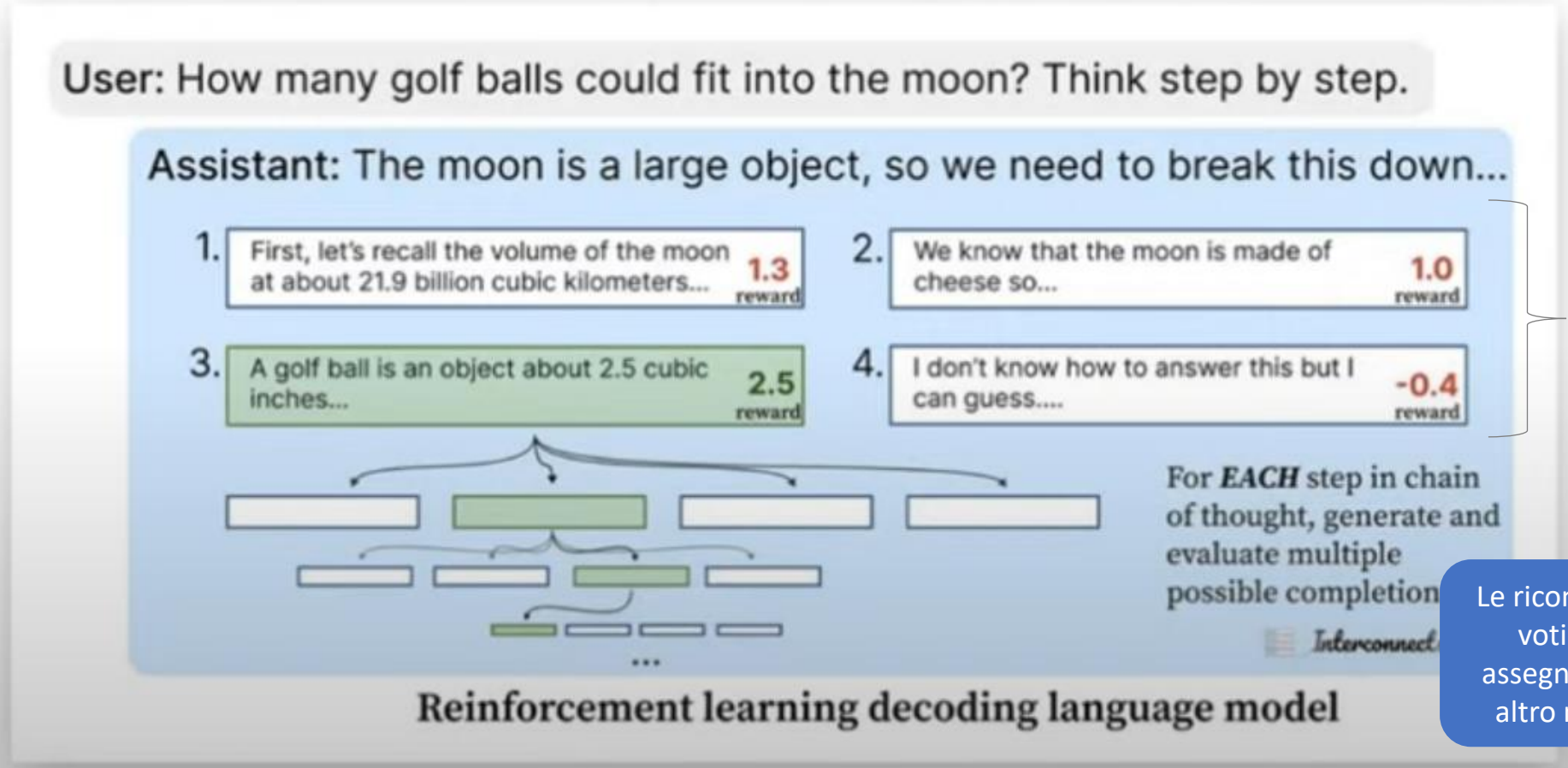
- E’ un processo iterativo e ricorsivo.
- DeepSeek esplicita meglio i passi (rispetto a o1 di OpenAI).
- Dà lui il contesto.

Come funziona?

- Utilizza "reinforcement learning" su catene di pensiero estese

A fine marzo 2025 i modelli reasoning principali sono:

- **o1** ed **o3** di OpenAI
- **DeepThink R1** di DeepSeek
- **Thinking Mode Extended** di Anthropic
- **Think** di Anthropic? No, è un trucco in JSON



Opzioni alternative prodotte dal primo modello

Le ricompense (i voti) sono assegnati da un altro modello

master: slide DeepSeek

Antonio Guadagno
47.500 iscritti

Ingegnere Informatico spiega l'ALGORITMO dietro le nuove immagini di ChatGPT

MODELLI DI DIFFUSIONE 🎨

il loro principale problema è che cercano di creare l'intera immagine in un colpo solo.

il limite principale dei modelli di diffusione è il fatto di non avere una comprensione sequenziale del processo creativo.

AUTOREGRESSIONE 🔄

creare le immagini un pezzettino alla volta.

Ogni pezzettino, ovviamente, dipende da quelli creati in precedenza.

Costruisce le immagini un pezzetto alla volta, da sinistra a destra e dall'alto in basso, tenendo sempre a mente il lavoro fatto fino a quel momento.

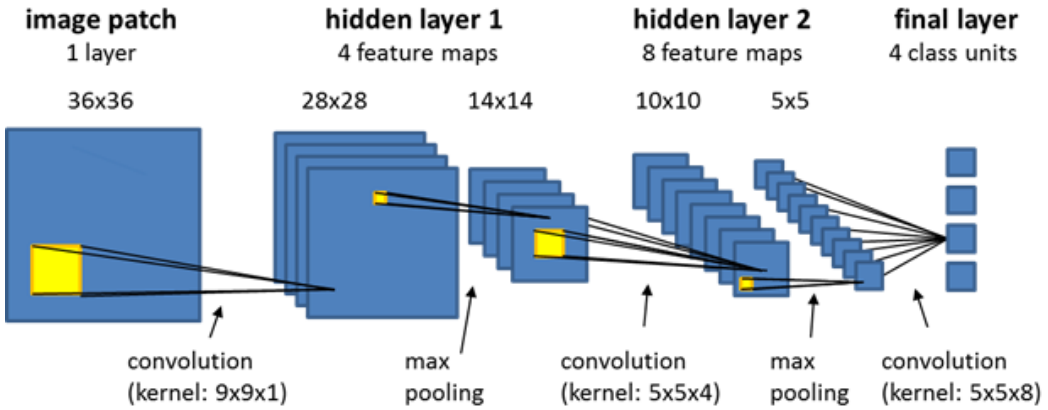
garantisce una maggior coerenza delle immagini.

master: sezione su Image Generator di OpenAI

immagini

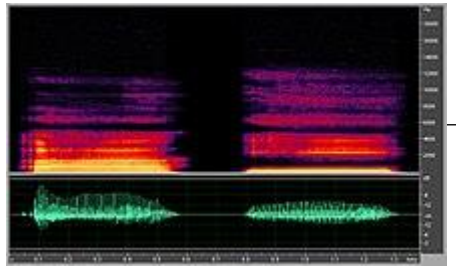
video

frame (immagini)



CNN per classificazione di immagini

audio



spettrogramma (immagine)