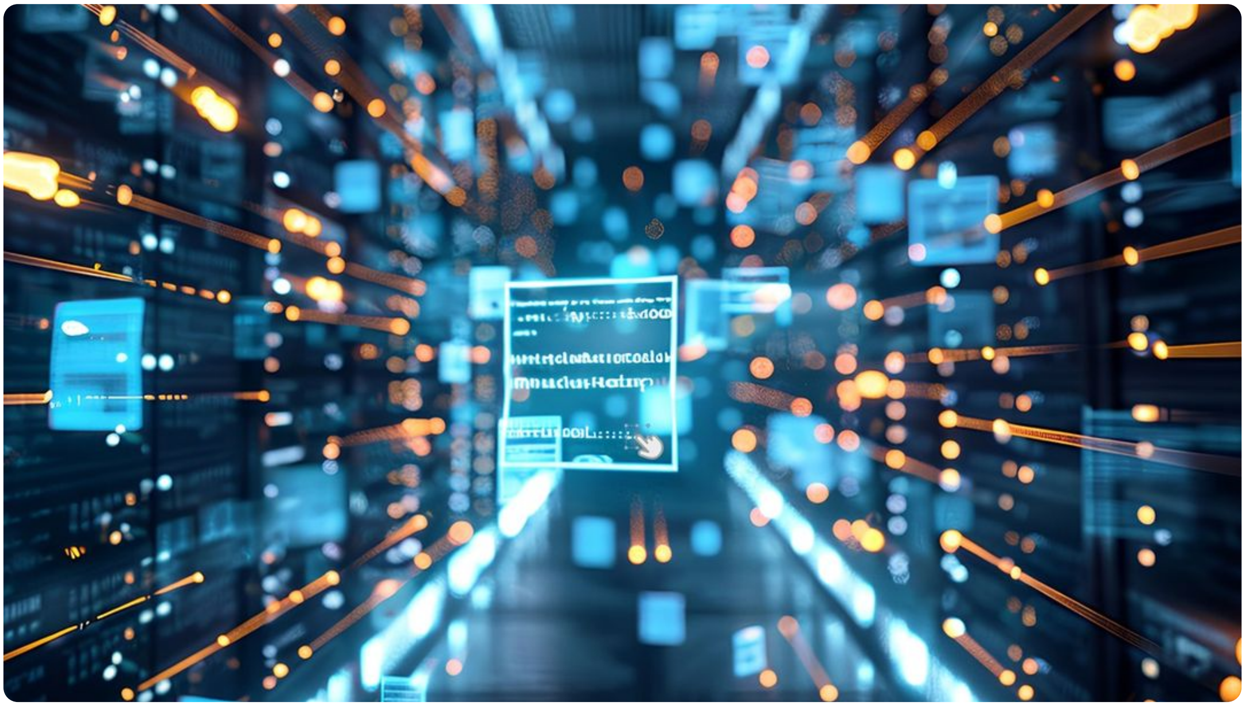
All posts

Knowledge

Data quality in Artificial Intelligence: an information-theoretic approach



Written by
Nanobaly

Published on
2024-10-26

Reading time
4 min

📘 CONTENTS

The expression **"*Garbage In, Garbage Out*"** is often quoted in Artificial Intelligence (AI), but few understand its theoretical underpinnings.

The race for performance in artificial intelligence often focuses on**model architecture**, **computing power** or **optimization techniques**.

Yet **one crucial aspect remains underestimated**: the quality of training data. Imagine building a house on an unstable foundation: no matter how sophisticated the architecture, the structure will be compromised.

Similarly, an AI model trained on noisy or mislabeled data will inevitably reproduce these defects. This reality is not just empirical; it follows directly from the fundamental principles of **information theory**. Understanding these principles helps us to understand why investment in data quality is often more important than investment in model complexity.

Theoretical foundations

Shannon's Entropy: the measure of information

🔗 **Claude Shannon** revolutionized our understanding of information by proposing a quantitative measure.**Shannon's entropy** is given by

> $H = -\sum p(x) \log_2(p(x))$

Where:

- **H** is entropy (measured in bits)
- **p(x)** is the probability of occurrence of an event x
- **∑** represents the sum over all possible events

This formula tells us something fundamental: information is linked to unpredictability. A certain event (p=1) brings no new information, while a rare event brings a lot of information.

Application to training data

In a training dataset, the total information can be broken down as follows:

> H_total = H_usable + H_noise

Where:

- **H_useful** represents information relevant to our task
- **H_noise** represents imperfections, errors and artifacts

This decomposition has a crucial consequence: since an AI model cannot intrinsically distinguish useful information from noise, **it will learn both.**
This runs the risk of reproducing the model's noise output.

The principle of information retention

The fundamental limit

A fundamental theorem of information theory states that a system cannot create information; it can only transform it. For an AI model, this means:

> Output_quality ≤ Input_quality

This inequality is strict: no architecture, no matter **how sophisticated**, can exceed this limit.

Case study: image upscaling

Let's take the example of photo upscaling, where we want to increase the resolution of an image:

(You can find a list of tools used for upscaling a photo 🔗 **here**)

The quality chain

For a high-resolution (HR) image generated from a low-resolution (LR) image :

> $\text{PSNR\_output} \leq \text{PSNR\_input} - 10 * \log_{10}(\text{factor\_upscaling}^2)$

Where:

- **PSNR** (Peak Signal-to-Noise Ratio) measures image quality
- **upscaling_factor** is the ratio between resolutions (e.g. 2 for doubling)

Impact of training data

Let's consider two training scenarios:

1. High Quality Dataset

- HR images: Uncompressed 4K photos
- Average PSNR: 45dB
- Possible result: ~35dB after upscaling x2

2. Dataset Poor
- HR images: JPEG-compressed photos
- Average PSNR: 30dB
- Maximum result: ~20dB after upscaling x2
The **15dB difference** in the final result **is directly linked** to the quality of the training data.

PSNR in dB is a logarithmic measure that compares the maximum possible signal with the noise (the error).
The higher the number of dB, the better the quality:

PSNR (Peak Signal-to-Noise Ratio) is defined as :

$$PSNR = 10 * \log_{10}(MAX^2/MSE)$$

Where:

- **MAX** is the maximum possible pixel value (255 for 8 bits)
- **MSE** is mean square error

For upscaling, when the resolution is increased by a factor of n, MSE tends to increase, which effectively reduces PSNR.
The quality of the result is therefore very sensitive to the level of noise.

Order of magnitude of PSNR in dB for images

- High-quality JPEG image: ~40-45**dB**
- Average JPEG compression: ~30-35**dB**
- A highly compressed image: ~20-25**dB**

dB is a logarithmic scale:

- +3**dB** = 2x better quality
- +10**dB** = 10x better quality
- +20**dB** = 100x better quality

So when we say "**~35dB** after upscaling x2", it means that :

1. *The resulting image has good quality*

2. *Differences from the "perfect" image are hard to see*
3. *Typical of a good upscaling algorithm*

The cascade effect: the danger of AI-generated data

When AI-generated images are used to train other models, degradation follows a geometric progression:

> Generation_quality_n = Original_quality * $(1 - \tau)^n$

Where:

- **τ** is the degradation rate per generation
- **n** is the number of generations

This formula explains why **using AI-generated images** to train other models **leads to rapid** quality **degradation**.

The importance of labelling

The quality of the labels is as crucial as that of the data itself. For a supervised model :

> Maximum_precision = min(Data_Quality, Precision_labels)

This simple formula shows that even with perfect data, **imprecise labels strictly limit possible performance**.

Practical recommendations

1. Dataset preparation

Above, our simplistic demonstration illustrates the crucial importance of the quality of the data used for training. We invite you to 🔗 **consult this article** to learn more about how to prepare a quality dataset for your artificial intelligence models.

We can't elaborate in this article, but the discerning reader will notice that the definition of "noise" raises philosophical questions. 🔗 **How do you define noise?**

2. Reflection: the subjective nature of noise

The very definition of "noise" in data raises profound philosophical questions. What is considered noise for one application may be crucial information for another.

Let's take the example of a photo:

- For a facial recognition model, lighting variations are "noise".
- For a lighting analysis model, these same variations are the main information.

This subjectivity of noise reminds us that data "quality" is intrinsically linked to our objective. Like Schrödinger's cat, noise exists in a superposition: it is both information and disturbance, until we define our observation context.

This duality underlines the importance of a clear, contextual definition of "quality" in our AI projects, challenging the idea of absolute data quality.

3. Quality metrics

For each data type, define minimum thresholds, e.g. :

Images

> PSNR > 40dB, SSIM >0.95

Labels

> Accuracy > 98

Coherence

> Crossover tests > 95% of results

The 40dB threshold is not arbitrary. In practice :

- 40dB: Virtually imperceptible differences
- 35-40dB: Very good quality, differences only visible to experts
- 30-35dB: Acceptable quality for general use
- <30dB : Dégradation visible

SSIM (Structural Similarity Index)

The SSIM complements the PSNR :

```
seuils_SSIM = {    "Excellent": ">0.95",    "Good": "0.90-0.95",
 "Acceptable": "0.85-0.90",    "Problem": "<0.85"    }
```

SSIM is closer to human perception, as it considers the structure of the image.

Consistency tests

Cross-tests >95% involve :

1. *k-fold* cross-validation
2. Internal consistency tests
3. Checking outliers
4. Distribution analysis

Conclusion

Information theory provides us with a rigorous framework

Information theory provides us with a rigorous framework demonstrating that data quality is not an option **, but a strict mathematical limit.** An AI model, no matter how sophisticated, cannot exceed the quality of its training data.

This understanding must guide our investments: rather than just looking for more complex architectures, our priority must be to **ensure the quality of our training data** !

Sources

Shannon entropy: 🔗 *https://fr.wikipedia.org/wiki/Entropie_de_Shannon*
Illustration: 🔗 *https://replicate.com/philz1337x/clarity-upscaler*

Academic and technical sources

1. *Shannon, C.E. (1948). "A Mathematical Theory of Communication". Bell System Technical Journal.*
2. *Wang, Z. et al. (2004). "Image Quality Assessment: From Error Visibility to Structural Similarity". IEEE Transactions on Image Processing.*
3. *Goodfellow, I., Bengio, Y., & Courville, A. (2016). "Deep Learning". MIT Press.*
4. *Zhang, K. et al. (2020). "Deep Learning for Image Super-Resolution: A Survey". IEEE Transactions on Pattern Analysis and Machine Intelligence.*
5. *Dodge, S., & Karam, L. (2016). "Understanding how image quality affects deep neural networks". International Conference on Quality of Multimedia Experience (QoMEX).*