Controlling Test-Time Compute in Autoregressive Image Generation with Next-Detail Prediction

Anonymous Author

Abstract

Autoregressive models have recently become competitive with diffusion-based approaches for image generation. However, these models remain computationally expensive, and fail to surpass the fastest approaches for image generation, such as GANs and distilled diffusion. We present Next-Detail Prediction (NDP), a method for precisely controlling the compute spent generating an image by focusing on difficult-to-predict regions. We achieve this by predicting sequentially higher-resolution patch maps of an image, choosing to upsample difficult patches first. We compare and evaluate several strategies for estimating the difficulty of image patches, and propose a masked architecture that predicts both patches and their difficulty in a single step. We show that this method can generate high-quality images with significantly less compute than previous approaches. On the ImageNet 256x256 Generation benchmark, we achieve state-of-the-art FID and IS 2x faster than baseline AR approaches, and demonstrate desirable scaling with higher resolutions.

1. Introduction

Much has been made of the potential for large language models (LLMs) to generalize to visual data modalities, such as images and video. Vision LLMs, such as GPT-4V, Gemini 1.0, Claude 3 and their derivatives, have shown impressive capabilities in understanding visual data. Advances in image tokenization, particular VQ-VAEs, allow discrete LLM token vocabularies to span text and images.

However, in deployment, these models do not use their native image tokenizers to generate images. State-of-the-art diffusion-based approaches such as DDPM and DiT remain widely used for image generation, and can be accelerated with methods such as LCM and sCM to generate images in less than a second. Diffusion is aesthetically less desirable than a general multimodal system, and seems not to benefit from the scalability and simplicity of self-supervised next-token prediction, so a high-quality autoregressive (AR) image generation paradigm is desirable.

Until recently, autoregressive models have been less competitive than diffusion for image generation. However, the recent work VAR showed that AR models can generate high-quality images, if asked to predict images in hierarchical scales rather than in raster order. VAR uses a hierarchical VQ-VAE to tokenize images into a sequence of increasingly-large token maps, and predicts these token maps autoregressively.

VAR cites inspiration from the fact that "humans percieve or create images in a hierarchical manner, first capturing the global structure and then local details". We add that humans percieve different parts of images preferentially – for example, an artist may focus on critical details such as eyes and faces, and spend less effort on background details. Many images in practice contain large regions of low-entropy single colors or patterns, which can be decoded easily. Therefore, we demonstrate that compute-efficiency of AR image models can be improved when they are allowed to focus on difficult-to-predict regions of an image first.

Our NDP model, following VAR, is trained on increasingly large token maps, however it follows a masked upsampling objective - given a $K \times K$ patch, it predicts the 4 $\frac{K}{2} \times \frac{K}{2}$ child patches in the following scale. We train our model to predict these next tokens, and also to predict a difficulty scalar, which is an estimation of the difficulty of each visible patch. We then use this difficulty map to choose which patches to upsample next. We show that this method can generate high-quality images with significantly less compute than previous approaches. We contribute:

[*]Equal contribution . AUTHORERR: Missing \icmlcorrespondingauthor.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

- A novel method for precisely controlling the compute spent generating an image by focusing on difficult-to-predict regions.
- A comparison of methods for evaluating the difficulty of image patches.
- A novel objective that predicts patches and difficulty simultaneously.
- A novel tokenizer that decodes tokens of various scales into a single coherent image

Our code is open-sourced at https://github.com/next-detail-prediction/next-detail-prediction.

2. Related Work

2.1. Image Tokenization

As mentioned, we use a vector-quantized autoencoder to tokenize images similar to VQ-GAN, which is itself derived from VQ-VAE. Additionally, we adopt some architectural ideas from LlamaGen's tokenizer, namely normalizing the embedding space.

2.2. Image Generation

2.3. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

2.4. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

2.4.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

2.4.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

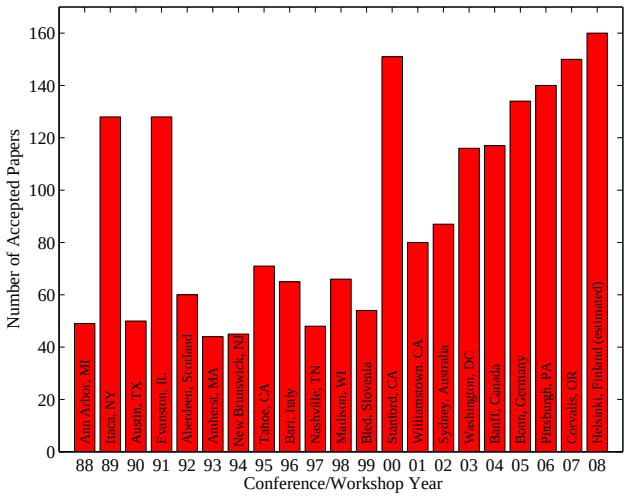

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

[1] Footnotes should be complete sentences.

[2] Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

Algorithm 1 Bubble Sort

```
Input: data x_i, size m
repeat
  Initialize noChange = true.
  for i = 1 to m − 1 do
    if x_i > x_{i+1} then
      Swap x_i and x_{i+1}
      noChange = false
    end if
  end for
until noChange is true
```

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

Data set	Naive	Flexible	Better?
Breast	95.9± 0.2	96.7± 0.2	√
Cleveland	83.3± 0.6	80.0± 0.6	×
Glass2	61.9± 1.4	83.8± 0.7	√
Credit	74.8± 0.5	78.3± 0.6	
Horse	73.3± 0.9	69.7± 1.0	×
Meta	67.1± 0.6	76.5± 0.5	√
Pima	75.1± 0.6	73.9± 0.5	
Vehicle	44.9± 0.6	61.5± 0.4	√

2.5. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption after the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 1. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment figure* in LaTeX). Always place two-column figures at the top or bottom of the page.

2.6. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

2.7. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title above the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

2.8. Theorems and such

The preferred way is to number definitions, propositions, lemmas, etc. consecutively, within sections, as shown below.

Definition 2.1. A function $f : X \to Y$ is injective if for any $x, y \in X$ different, $f(x) \neq f(y)$.

Using Definition 2.1 we immediate get the following result:

Proposition 2.2. If f is injective mapping a set X to another set Y, the cardinality of Y is at least as large as that of X

Proof. Left as an exercise to the reader. □

Lemma 2.3 stated next will prove to be useful.

Lemma 2.3. For any $f : X \to Y$ and $g : Y \to Z$ injective functions, $f \circ g$ is injective.

Theorem 2.4. If $f : X \to Y$ is bijective, the cardinality of X and Y are the same.

An easy corollary of Theorem 2.4 is the following:

Corollary 2.5. If $f : X \to Y$ is bijective, the cardinality of X is at least as large as that of Y.

Assumption 2.6. The set X is finite.

Remark 2.7. According to some, it is only the finite case (cf. Assumption 2.6) that is interesting.

2.9. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use natbib.sty and icml2025.bst included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (1959). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Samuel, 1959). List multiple references separated by semicolons (Kearns, 1989; Samuel, 1959; Mitchell, 1980). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Michalski et al., 1983).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to ?? for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Samuel, 1959), conference publications (Langley, 2000), book chapters (Newell & Rosenbloom, 1981), books (Duda et al., 2000), edited volumes (Michalski et al., 1983), technical reports (Mitchell, 1980), and dissertations (Kearns, 1989).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent, e.g. use the actual current name of authors. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

Accessibility

Authors are kindly asked to make their submissions as accessible as possible for everyone including people with disabilities and sensory or neurological differences. Tips of how to achieve this and what to pay attention to will be provided on the conference website http://icml.cc/.

Software and Data

If a paper is accepted, we strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, do not include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the OpenReview reviewing system. Note that reviewers are not required to look at this material when writing their review.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and usually should) include acknowledgements. Such acknowledgements should be placed at the end of the section, in an unnumbered section that does not count towards the paper page limit. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

Impact Statement

Authors are required to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. This statement should be in an unnumbered section at the end of the paper (co-located with Acknowledgements – the two may appear in either order, but both must be before References), and does not count toward the paper page limit. In many cases, where the ethical impacts and expected societal implications are those that are well established when advancing the field of Machine Learning, substantial discussion is not required, and a simple statement such as the following will suffice:

"This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

The above statement can be used verbatim in such cases, but we encourage authors to think about whether there is content which does warrant further discussion, as this statement will be apparent if the

paper is later flagged for ethics review.

References

Duda, R. O., Hart, P. E., and Stork, D. G. Pattern Classification. John Wiley and Sons, 2nd edition, 2000.

Kearns, M. J. Computational Complexity of Machine Learning. PhD thesis, Department of Computer Science, Harvard University, 1989.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), Proceedings of the 17th International Conference on Machine Learning (ICML 2000), pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). Machine Learning: An Artificial Intelligence Approach, Vol. I. Tioga, Palo Alto, CA, 1983.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), Cognitive Skills and Their Acquisition, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Samuel, A. L. Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3):211–229, 1959.

A. You can have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.