Autoregressive Models: Inductive Biases and Projections

Abstract

This paper explores the use of inductive biases and projection functions in autoregressive (AR) models to enhance out-of-distribution (OOD) generalization. We revisit the concept of infini-grams, which leverage suffix arrays to manage arbitrary context lengths efficiently. This approach is compared to traditional n-gram models, highlighting its advantages in sample efficiency and computational scalability. We delve into various inductive biases, such as the recency bias, shortest edit distance, and semantic similarity, illustrating their impact on AR model performance. By framing OOD generalization as a projection problem, we propose strategies to optimize these projections through meta-learning and nested optimization. Furthermore, we discuss the integration of classical information retrieval techniques and pre-trained language model embeddings to enhance the semantic relevance of projections. Our findings suggest that combining symbolic AI methods with deep learning representations can yield more interpretable and sample-efficient AR models, with broad applications in natural language processing, code generation, and scientific discovery.

Introduction

Recently, I watched a presentation on infini-grams, which utilize a suffix array to avoid precomputing n-grams and allow for arbitrary context lengths, up to a suffix that is found in the training data.

This sparked my interest as I had worked on a similar project for a LLM talk I gave for SLUUG at https://www.stllinux.org (see my GitHub repo https://github.com/queelius/sluug-talk-llm and the video fo the talk at https://www.sluug.org/resources/presentations/media/2024/STLLINUX/2024-02-22_STLLINUX_2560x1440.mp4) where in part of the talk I demonstrated arbitrary-size n-grams using a recursive dictionary to store synthetic training data prefix counts to implement a crude expression tree evaluator.

Since my data was sparse synthetic data (expression trees and their evaluations), I was able to use a relatively inefficient approach to compute very large n-grams. The infini-gram approach is more efficient and generalizes to any kind of data, so they definitely had a more practical solution.

The infini-gram model is an autoregressive (AR) model that predicts the next token based on the longest suffix in the training data that matches the context. Essentially, they are finding some *projection* of the input (context) to the training data to allow the AR model to generate coherent text continuations from inputs it has never seen before. This is known as out-of-distribution (OOD) generalization, where we are trying to generalize to tasks (like predict continuations of an input never seen before) that is not in the training data.

Autoregressive (AR) Models

Autoregressive (AR) models form a cornerstone in natural language processing, predicting the probability of a word w_t given all preceding words $w_{<t}$ and the training data D:

$$\Pr\{w_t \mid w_{<t}, D\}.$$

Historically, the prefix $w_{<t}$ is limited to a fixed length n,

$$\Pr\{w_t \mid w_{t-n:t}, D\},$$

where $a : b$ denotes the range $a, a+1, \ldots, b-1$.

Infini-gram models dynamically adjust the context length based on the longest suffix in the training data that matches the context:

$$\Pr\{w_t \mid \text{longest\_suffix}(w_{<t}, D), D\},$$

where longest_suffix finds the longest suffix of the context $w_{<t}$ in the training data D.

For autoregressive models to generate continuations of the input, longest_suffix makes a lot of sense. It allows the model to find training data that is both similiar to the input and relevant to the task at hand: predicting the next token from previous tokens.

Let's be a bit formal about what longest_suffix represents: it is a kind of *projection* of the input onto the training data D, which is an i.i.d. sample from some (unknown) data generating process (DGP). Let us denote the probability distribution of the DGP as $\Pr_\theta$, where θ are unknown parameters, and the probability distribution of the AR model as $\Pr_{\hat{\theta}}$, where $\hat{\theta}$ are the estimated parameters.

The goal of the AR model is to estimate θ from the training data, which will allow it to generalize to new data that the DGP would plausibly produce. A paricularly useful task is to predict what the DGP would plausibly produce *given* some input $w_{<t}$, where $w_{<t}$ is a sequence of tokens that the DGP has produced so far and may represent some task of interest, like "What is the solution to <math problem>?"

The distribution of $w_{t:t+k}$ conditioned on $w_{<t}$ is given by

$$\Pr_\theta\{w_{t:t+k} \mid w_{<t}\} = \frac{\Pr_\theta\{w_{1:(t+k)}\}}{\Pr_\theta\{w_{1:t}\}},$$

where $w_{a:b}$ is a sub-sequence of tokens produced by the DGP from time a to time b (time is a *logical time* that just implies some ordering). The primary task is

often to *generate* plausible continuations of the input, for which there are many possible *sampling* strategies to do this, like beam search, top-k sampling, and nucleus sampling, all of which use the conditional probability distribution to generate continuations one token at a time. This approach is justfied by the chain rule of probability:

$$\Pr_\theta\{w_t \mid w_{<t}\} = \prod_{i=1}^{t} \Pr_\theta\{w_i \mid w_{<i}\}.$$

Notice that we are not necessarily trying to find a sequence that *maximizes* the conditional probability,

$$w^*_{t:(t+k)} = \arg\max_{w_{t:(t+k)}} \Pr_\theta\{w_{t:(t+k)} \mid w_{<t}\},$$

but rather we are *sampling* from the distribution. This is because the DGP is often stochastic and we are trying to capture this stochasticity in our predictions or continuations. (There is also a trade-off between exploration and exploitation, where the model needs to balance between generating plausible continuations and exploring new possibilities.)

It is also worth pointing out that finding the most likely sequence of tokens is an NP-hard problem, so we often resort to approximate methods like beam search to find a good sequence of tokens that is likely to be produced by the DGP.

Since we do know know the DGP $\Pr_\theta$, we replace it with our AR model based on a training data D, $\Pr_{\hat{\theta}}$, and use the AR model to approximate the DGP. As the sample size goes to infinity, by the law of large numbers, the empirical distribution of the training data will converge to the true distribution of the DGP:

$$\lim_{n\to\infty} \Pr_{\hat{\theta}}\{w_t \mid w_{<t}, D_n\} = \Pr_\theta\{w_t \mid w_{<t}\}$$

However, we do not have *infinite* training data, so we need to find a way to generalize to OOD data, like continuations of the input that the DGP would plausibly produce but are not in the training data. We call this *out-of-distribution* (OOD) generalization, and it is a key challenge in machine learning. Ideally, we want the AR model to generate plausible continuations of any input from very small amounts of training data D. A primary way to do this is to *constrain* the model using what we call *inductive biases*.

The projection function longest_suffix is an example of an inductive bias. It is a way for the model to find the most relevant part of the training data to the input to give it some ability to generalize OOD on the task of predicting the next token (or, equivalently, generating continuations of the input).

We formalize this idea of projection as an *inductive bias* and discuss how it can be used to improve the sample efficiency of both n-gram models and AR models, like transformers, LSTMs, and RNNs.

Inductive Biases

The projection function longest_suffix is what we call an *inductive bias*. Inductive biases are assumptions or constraints that help the model generalize from the training data to unseen data. This is known as OOD (out-of-distribution) generalization. Since the empirical distribution (the training data) converges to the true distribution (the DGP) as the sample size goes to infinity, the key challenge is finding *sample efficient* models.

Given two learning algorithms, A and B, if A requires fewer samples to do well on a task than B, then A is more sample-efficient than B on that task. In the context of AR models, the task is to predict the next token given a sequence of previous tokens. One way to improve sample efficiency is to choose an inductive bias that provides the model with more information about the task or the DGP, allowing it to generalize to OOD data more effectively and with fewer samples.

The longest_suffix projection is an inductive bias that we might label the *recency bias*. The recency bias has several advantages:

1. It is computationally efficient, as shown by the suffix array data structure used in the infini-gram model. It only requires a linear scan of the training data to find the longest suffix. This scalability is crucial for training on large datasets, as the time complexity of the recency bias is $O(n)$, where n is the length of the context.

2. It corresponds to a simple inductive bias that is easy to understand, implement, and justify. If the future is like the past, then the most recent past is often the most relevant data point. This is particularly relevant for tasks like language modeling, where the context is often a sequence of words that are related to each other in a temporal order and in which the most recent words are often the most relevant for predicting the next word.

Clearly, the recency bias may not always help to find the most relevant context in the training data, e.g., the most relevant context may be at the start of a document. However, even when the most relevnat context is the most recent, the longest_suffix may fail to properly use it. For example, if the context is "the dog ran after the" and we ask it to predict the next word, but the training data only contains "the dog chased the cat", the longest suffix is the highly uninformative word "the". We see that the naive longest suffix match fails to take into account the semantic similarity between token sequences like "chased" and "ran after".

These challenges suggest some possible inductive biases that can be used to improve the OOD generalization of $\Pr_{\hat{\theta}}$. We consider the set of inductive biases

that can be formulated as *projections* of the input (context) onto the training data. Let us formally write down the problem of OOD generalization in the context of AR models as a projection problem:

$$\Pr\{w_t \mid \text{proj}_D(w_{<t}), D\},$$

where proj_D is a function that maps the input $w_{<t}$ to a subset of the training data D that is most relevant for producing continuations of the input that the data generating process (DGP) would *likely* produce.

Choosing the Projection Function

Let us parameterize the projection function as

$$\mathcal{F} = \{\text{proj}_D(x; \beta) \mid \beta \in \mathcal{B}\},$$

where β is an index or label that specifies the projection. For example, $\beta = 1$ could be a label for longest_suffix, or it could be something more complicated based on the space of possible projections $\mathcal{F}$.

We can choose a projection function from $\mathcal{F}$ by choosing a β in $\mathcal{B}$, which is frequently a discrete optimization problem.

We can choose the projection function in two fundamental ways. A common approach is to simply choose a value using domain knowledge. However, by lessons of the bitter kind, we know that this is not always the best approach, as it requires human expertise that is often limited. A more general approach is to treat this as a secondary optimization problem, where we treat β as a parameter of the AR model and estimate it from the data. This is a form of *meta-learning* where we are learning the projection function from the data D:

$$\hat{\beta} = \arg\max_{\beta} \prod_{t=1}^{T} \Pr_{\hat{\theta}}(w_t \mid \text{proj}_{D'}(w_{<t}; \beta)),$$

where D' is the held-out validation data used to estimate the quality of the projection function and

$$\hat{\theta}|\beta = \arg\max_{\theta} \prod_{t=1}^{T} \Pr_{\theta}(w_t \mid \text{proj}_D(w_{<t}; \beta)).$$

We see that we may want to estimate $\hat{\theta}$ on the training data using the projection function parameterized by β and then estimate $\hat{\beta}$ on the validation data using the estimated parameters $\hat{\theta}$. This is a form of *nested optimization* where we are optimizing the parameters of the AR model and the projection function simultaneously.

If the projection functions do not affect the performance of the AR model on the training data D, then we can optimize them separately. However, if the projection functions do affect the performance of the AR model on the training data, then we need to optimize them simultaneously in this two-step process. One way in which the projection functions can affect the performance of the AR model is by changing the distribution of the training data that the AR model sees, which can change the parameters of the AR model that are estimated from the data. For instance, β may include a parameter that limits the maximum length of the context, which can change the parameters of the AR model that are estimated from the data.

Next, we consider the space of possible projection functions $\mathcal{F}$. We can draw inspiration from techniques developed in information retrieval (IR), natural language processing (NLP), and classical AI informed search strategies to design projections (inductive biases) that yield more sample efficient algorithms that improve OOD generalization. It is worth pointing out that in high-dimensional spaces, essentially every input is OOD, so designing effective inductive biases (projections) is crucial for generalization.

Similarity Bias: Shortest Edit Distance

Shortest edit distance finds the shortest sequence of operations (e.g., swaps, insertions, deletions, and substitutions) transforming the current context $w_{<t}$ into a sequence in D.

The recency bias can be seen as a special case of the similarity bias where we only allow deletions from the end of the context until a match is found.

A justification for using shortest edit distance is based on the idea of similarity: if two sequences are similar, they are more likely to have similar continuations. Edit distance is a way to measure the similarity between two sequences based on the minimum number of operations needed to transform one into the other.

Example

Let's use the earlier example, where we have as input "the dog ran after the" and the training data D contains the following:

1. "a dog chased the cat in the garden"
2. "the dog chased the cat, but the cat climbed a tree and got away"
3. "the mouse ran from the cheese trap after setting it off"

The longest suffix is "the". What is shortest edit distance to find a match in the training data? We can perform the following two edits: substitute "ran" with "chased" and delete "after", resulting in "the dog chased the" which has a longest suffix match equal to "the dog chased the" (2), and so the next token predicted is "cat". Thus, a completion by the model might be: "the dog ran after the cat, but the cat climbed a tree and got away". The training data does not contain

this completion, but the shortest edit distance found a *relevant* projection to the training data.

What else could we have found within two edits? We could have substituted "dog" with "mouse" and "after" with "from", resulting in the input "the mouse ran from the" and the completion "the dog ran after the cheese trap after setting it off". This is a less plausible completion for the DGP, and things could go much worse, but we see that just counting the number of edits can lead to a poor projection onto the training data.

Challenges

As the example demonstrated, a primary challenge with shortest edit distance is that it treats all single edits as having a uniform cost. Ideally, when we edit the input, we want to preserve the "meaning" of the context. Thus, some edits should be more costly than others, based on how much they change the meaning of the input.

Also, the shortest (least-cost) edit distance, particularly when we combine it with non-uniform costs, is more computationally intensive than the longest prefix projection. However, it is tractable, e.g., graph search in GOFAI. Approximate methods like Monte Carlo Tree Search (MCTS) can also be used to find approximate solutions.

Semantic Similarity: Least-Cost Edit Distance

A significant issue with *shortest edit distance* is that it treats each edit as having a uniform cost (a kind of uninformed search). A simple extension is to add a cost to each edit based on some measure of semantic similarity between tokens or sequences. Once we have a cost in place, we can use classical search techniques, like A* search, to efficiently find the most relevant sequences in the training data to the current context.

Classical IR (Information Retrieval) techniques like BM25, query expansion, and semantic similarity measures can be used to assign costs to edits.

1. Input expansion (query expansion): Expand the input to multiple possible sequences that are similar to the input.
2. Treat the input like a search query in IR and use BM25 or other similarity measures to find the most similar sequences in the training data. We then define the projection function as:

$$\text{proj}_D(x; \beta) = \arg \max_{y \in \text{segments}_\beta(D)} \text{similarity}_\beta(x, y),$$

where $\text{segments}_\beta(D)$ is a segmentation strategy of the data D into segments (e.g., sentences paragraphs) and similarity_β is a similarity measure between the

input x and the segment y, e.g., cosine similarity or Euclidean distance between embeddings or some tf-idf measure, like BM25.

We can use the inferred β to find the most relevant segments in the training data to the input, and then use the AR model to generate continuations of the input based on these segments.

Inductive Bias to Favor Cohesive and Smooth Continuations

If we change the input too much, we may lose the coherence and smoothness of the text, as the model may generate completions that are incoherent or nonsensical when we map the input back to its original form. Thus, we may want to compose the similarity bias with a recency bias that favors keeping the most recent tokens in the context minimally edited. We can do this by generalizing the recency bias to assign a cost to each edit based on the position of the token in the context, where a higher cost is associated with changing more recent tokens in the input.

Challenges

Learning sample efficient representations of the data is the primary driver of OOD generalization. *Deep Learning* is about learning these representations from the data. We can use pre-trained models like BERT and GPT to learn embeddings of the data that are more semantically meaningful than the raw tokens. These embeddings can be used to compute the similarity between tokens or sequences and assign costs to edits based on this similarity.

If we use LLM embeddings, it may be costly to compute the similarity between the input and all segments in the training data. We could, however, take the training data and compute embeddings for each segment and store them in a vector storage database for fast retrieval:

$$\text{proj}_D(x;\beta) = \arg \max_{y \in \text{segments}_\beta(D)} \text{similarity}_\beta(\text{embed}(x), \text{embed}(y)),$$

where embed is a function that maps tokens or sequences to embeddings. Since we have all of the embeddings in a vector storage database, the above $\arg\max$ operation can be computed very efficiently at the cost of precomputing and storing the embeddings.

Experimental Results

We can compare the performance of the recency bias, shortest edit distance, and semantic similarity bias on a language modeling task. We can use perplexity as a measure of the model's performance on the task, where lower perplexity indicates better performance.

Python Code

```
# code here
```

For more details, see the GitHub repository for this project.

Conclusion

We have reframed of OOD generalization in the context of AR models as a context reduction and matching problem and explored various inductive biases to improve sample efficiency.

Even hand-crafted inductive biases like the recency bias and similarity bias can significantly enhance AR models' performance, but utilizing learned embeddings from pre-trained models like BERT and GPT can likely yield more effective results.

In either case, we see that the goal is to reduce or rewrite the context to increase the probability of finding a match in the training data. This is a discrete optimization problem that can be solved using techniques from information retrieval and natural language processing, and so we can leverage these techniques to design more sample-efficient learning algorithms that search over the space of possible context reductions and rewrites to find the most relevant training data to facilitate OOD generalization.

Even if an exact match is found in the training data, we often still want to explore a larger space of possibilities to make new discoveries and generate novel and creative outputs.

Further research into optimizing these techniques and seamlessly integrating them into AR frameworks promises to advance natural language processing, driving innovation in computational linguistics and machine learning, and help facilitate the development of more intelligent and creative AI systems that can be more easily explained and understood than current neural models, which are often seen as black boxes with inscrutable decision-making processes.

By leveraging a LLMs embeddings for sample efficient representatations and classical symbolic AI techniques for context reduction and matching, we can build more interpretable and efficient AR models that can be used in a wide range of applications, from chatbots to code generation to scientific discovery and beyond.

Appendices

A: Reward Functions

Previously, we discussed the idea of projecting the input onto the training data to find the most relevant context for predicting the next token that the DGP is likely to produce.

However, we are normally not interesed in predicting the next token, but *biasing* the model to generate outputs that are more likely to score well on some task. For example, if the task is to generate a coherent text continuation, we want to bias the model to generate text continuations that are coherent and correct, even if the the DGP, given the input, is more likely to generate very different continuations (e.g., toxic, incoherent, or incorrect).

One way to fine-tune the model to generate more effective outputs is to use a reward function that scores the outputs based on some task-specific criteria. A nearly universal approach is to take your reward function and produce k samples from the model, then score each sample using the reward function, and then sample these outputs based on their scores, e.g., $\Pr_{\hat{\theta}}\{w_{t:(t+k)} \mid w_{<t}\} \propto \exp\{R(w_{1:(t+k)})\}$.

However, what if we want to go beyond predicting the DGP's next token and instead generate outputs that are more effective at solving a particular task?

B: Data Transformation

The training data D is the primary source of information we have about the DGP. However, the training data may not be in the optimal form for solving the task we have in mind.

So, a final inductive bias we can consider is data transformation. We can transform the training data into a more suitable form for solving the task at hand. One way which can be particularly effective at improving the sample efficiency of the model is to transform the training data into a more abstract representation space.

There are a lot of fancy things you can do, but in interest of transparency and interpretability, we will consider a simple transformation: stemming or lemmatization.

Both of these are computationally efficient techniques that reduce the vocabulary size and thus increase the probability of finding relevant projections of the input onto the training data. They are also simple to implement and understand, making them a good choice for a first pass at data transformation.

Essentially, this transformation allows for "reasoning" over more abstract representations of the DGP, facilitating OOD generalization but at a loss of some information and expressiveness.

> Predictive modeling now takes place over this more abstract representation space, which can be more sample efficient. However, when we generate sequences, if the end product is, say, high-quality text, we may have to decode the stemmed or lemmatized representations back to unstemmed or unlemmatized forms, which can be a challenge.

Other Kinds of Inductive Biases

We formalized most of our inductive biases as projections of the input onto the training data. However, we can consider other kinds of inductive biases that can be used to improve the sample efficiency of AR models that directly affect the AR model's probabilty distribution.

In particular, we can also consider more complex inductive biases that involve pattern matching and context-free grammars. For example, we can use a context-free grammar to define the space of possible continuations of the input.

In theory, we could have a number of production rules on the input that restrict the set of possible continuations to some CFG. This is a more hand-crafted inductive bias that requires more domain knowledge and human expertise to implement, but it may be appropriate in some circumstances, such as when the task requires generating text that follows a specific structure or format, like JSON or Python code.