LLMs are Generative Models

Introduction to Machine Learning (CSC 2515)
Fall 2024

University of Toronto

Generative Models

- We saw in lecture that "Generative Models" model both the data and the target:

$$P(x, t)$$

- This is the joint distribution between the data and the targets, which is why these are also called "joint" models.

Next Token Prediction

- You've probably heard people dismiss LLMs as just "next token predictors"
- They model the prediction of the next word t given the preceding context X, that is,

$$P(t|X)$$

- This is pretty clearly a discriminative model...

What's in an X

- When working with language, we have a sequence of "tokens"
- We break down the "context" X into this sequence $x_1, x_2, \ldots, x_n$
- Thus we are predicting

$$P(t|x_1, x_2, \ldots, x_n)$$

When t isn't a t

- We are predicting the next word, so t is actually one of the possible values an x can take on.
- t can be thought of as x_{n+1}
- Thus we are modeling

$$P(x_{n+1}|x_1, x_2, \ldots, x_n)$$

Predicting each Token

For a given string, our model is predicting each token

$$
\begin{gathered}
P(x_1|\varnothing) = P(x_1) \\
P(x_2|x_1) \\
P(x_3|x_1, x_2) \\
\cdots \\
P(x_n|x_1, x_2, \ldots, x_{n-1}) \\
P(x_{n+1}|x_1, x_2, \ldots, x_n)
\end{gathered}
$$

Compactly, we can express this as:

$$
P(x_{n+1}|X) = \prod_{i=1}^{n} P(x_i|x_{<i})
$$

The Chain Rule of Probability

We can repeatedly use the Chain Rule of Probability to breakdown a joint distribution between multiple variables into conditionals.

$$
\begin{aligned}
P(A_n, \ldots, A_2, A_1) =& P(A_n|A_{n-1}, \ldots, A_2, A_1)P(A_{n-1}, \ldots, A_2, A_1) \\
=& P(A_n|A_{n-1}, \ldots, A_2, A_1)P(A_{n-1}|A_{n-2}, \ldots, A_2, A_1) \\
& P(A_{n-2}, \ldots, A_2, A_1) \\
=& P(A_n|A_{n-1}, \ldots, A_2, A_1)P(A_{n-1}|A_{n-2}, \ldots, A_2, A_1) \\
& P(A_{n-2}|A_{n-3} \ldots, A_2, A_1) \ldots P(A_2|A_1)P(A_1)
\end{aligned}
$$

What does that look like? Our next token predictions! So our LLM is really predicting:

$$P(x_{n+1}, x_n, \ldots, x_2, x_1)$$

What's the probability of a string?

- So our model is actually predicting the joint prediction of the data, $x_1, x_2, \ldots, x_n$, and the target. $t = x_{n+1}$. It is a generative model!
- This is actually the reverse of how Language Models are normally presented in NLP.
- An LM assigns a probability to a string $P(X)$, but it is intractable to enumerate all possible string values.
- The Chain Rule of Probability are used decompose the probability of the full string into an autoregressive sequence of next token predictions.
- This is similar to our decomposition of the probability distribution in the Naïve Bayes Classifier.
- Next token prediction has becomes most peoples first introduction to language modeling due to the popularity of LLMs.