Plug and Play Language Models: A Simple Approach to Controlled Text Generation

Romain Gambardella Philippe Gratias-Quiquandon

MVA 2025

Plan

Controlled Text Generation

[–] The potato and cauliflower are both in season to make combo breads, mounds, or pads. For an added challenge, try some garlic mashed potatoes.
[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you.. . .
[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them – so many little ones.
[Science] The potato was once thought to have no health problems and has been promoted as a nutritious food source since the mid-1800s, but recent reports indicate that it has many harmful health issues. In fact, researchers from Johns Hopkins University. . .
[Politics] **[Positive]** To conclude this series of articles, I will present three of the most popular and influential works on this topic. The first article deals with the role of women's political participation in building a political system that is representative of the will of the people.
[Politics] **[Negative]** To conclude, the most significant and lasting damage from the economic crisis in 2008 was that many governments, including those in the political center, lost power for the first time in modern history.

Figure: Example from the article

Plug and Play Methods

$$p(x|a) \propto p(x)p(a|x)$$

$p(x)$ should be a generative model $\Rightarrow$ could be any generative model adapted to the situation (Pre-trained Transformer)
$p(a|x)$ should be a classifier adapted to the conditionning we want to use $\Rightarrow$ can be easy to train.
In order to maximize $p(x|a)$, we maximize $p(x)p(a|x)$.

The idea in a nutshell

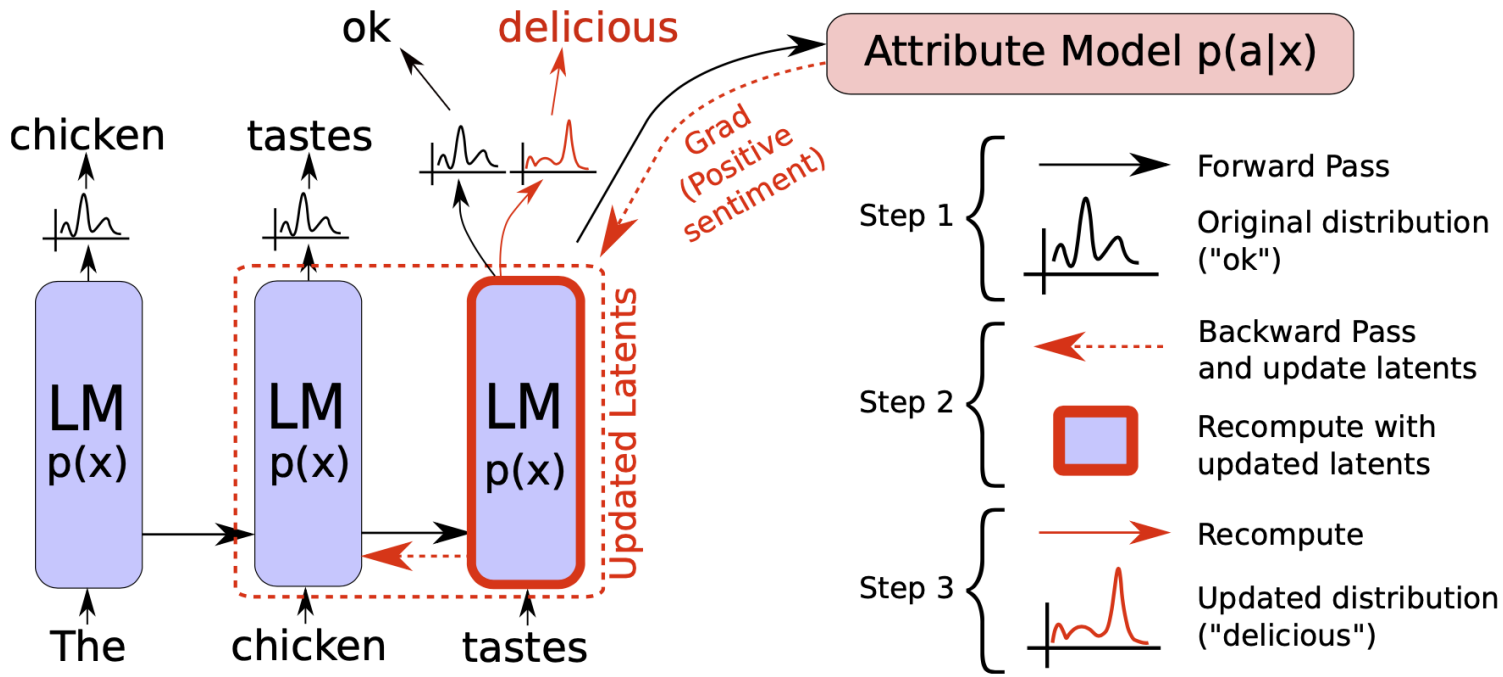

Figure: Scheme of the updates

How to choose $p(a|x)$

Two ways to classify sentences :

- BoW: gives a distribution of topics given the presence of certain words in the text.
- Classifier: adding a layer to GPT-2 and training on classification topics task.

Building the Transformer Model

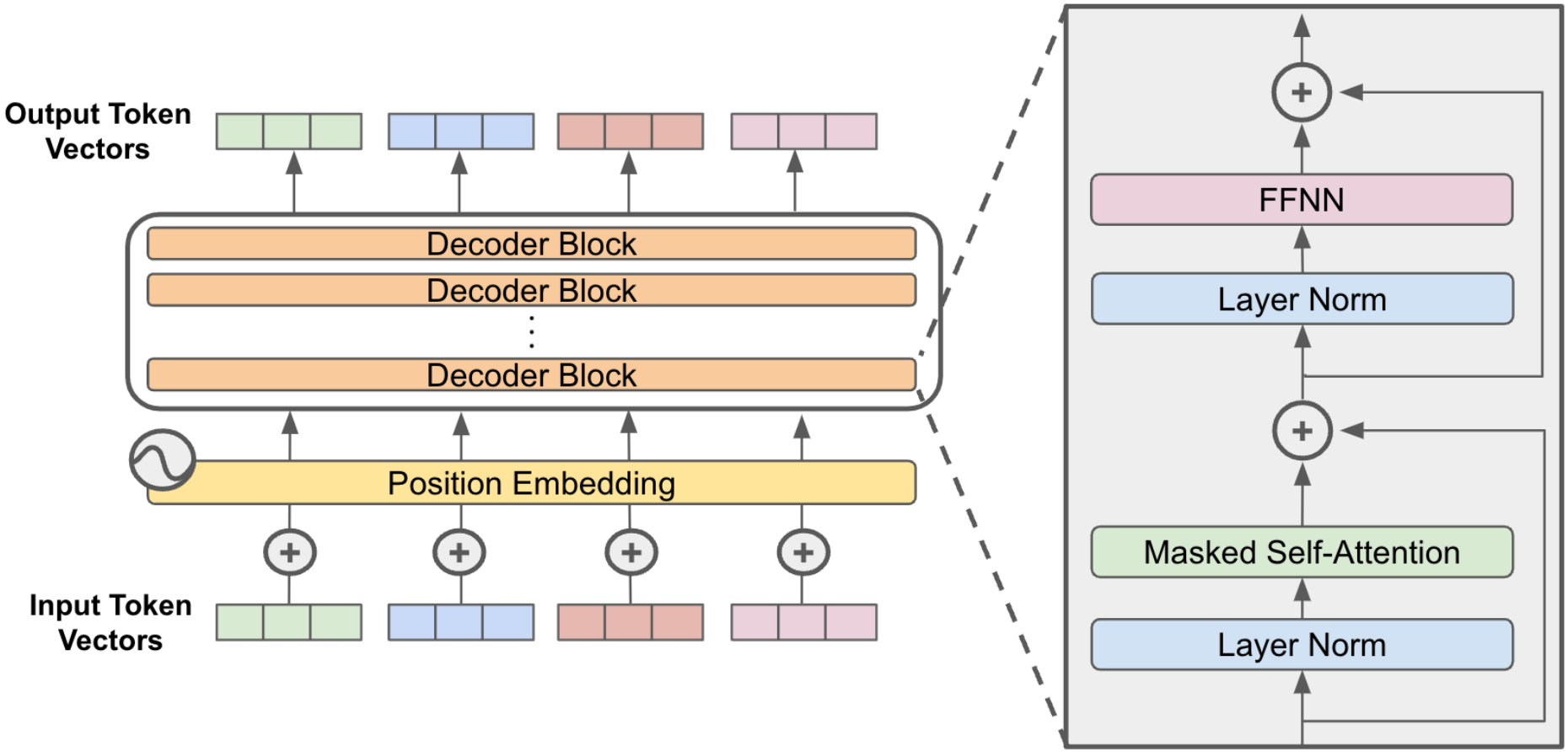

Figure: Transformer model

Building the Transformer Model

- built from scratch
- Hugging Face FineWeb-edu
- Custom tokenizer BPE
- Pretrained during 1 day
- implemented KV caching

Result example pretrained transfromer

The chickenpox virus is known to cause cancer, including some cancers, but not even cancer. And, this may also be one of the main reasons for the rise of cancer. The study also found that high in all mice, chickenpox is one of the most severe cancers among the mammals. And, for the most part, it's not just one thing to worry about, cancer is the most common cancer in men. So why are these cancers diagnosed?
The chicken is already eaten? We are going to take some fresh, raw chicken for the week, then it's worth it! Food-consciousness. It is a tasty treat for someone who has a high risk. However, eating chicken eggs is an important way to keep away from the risk of developing heart disease. If you are taking a chicken feed with chicken eggs, it can also be a sign that your kids are getting sick.

Implementation of Bag of Words controlled generation

We want to control generation toward a specific bag of words.
We use the formula $log(p(a|x)) = log(\sum_{w_i \in BoW} p_{t_{i+1}}(w_i))$
We ascend $log(p(a|x))$ during inference

Result example : Bag Of Words

[Computer] The chicken industry's most recent success at the University of Pennsylvania, and it's a highly proliferous research project for the United States Department of Agriculture, the US Department of Agriculture, and the US Department of Agriculture. The U.K, which is home to nearly a dozen species of the disease-eater, makes its study in the journal Nature Genetics and is one of the most comprehensive studies to date around the world. In June of 2019, the researchers published their findings in the journal Nature Genetics, which was a leading research center for the Human Genome Project. Have you seen this project? We've been working through a recent publication that will show you how it'll go through a **process** where a team of researchers began to identify, and hopefully answer questions. **Facebooks** are great for the study, but have you noticed them? You can read the original research project here.

Bag of Words

Drawbacks of BoW :

- Highly sensitive to bag of word content.
- Hard to tune
- Easy to 'cheat'

$\Rightarrow$ Neural Network Discriminator for $p(a|x)$ $\Rightarrow$ GPT-2

Example without ascend on $p(x)$

The chicken is vibrant beautifully complement beautiful beautifully complement dazzling beautifully complement beautifully complement beautifully

Working example

The chicken is just as juicy as the regular. I also love using it only once. It's wonderful in sauces, curries, sides, and dips. Can get crispy on an egg. A wonderful protein powder if you're really into protein shakes. I wish I could use this flavor combination. Would also loved sharing this with my family!
I love this flavor combination but it also turns black (what??) which is what is sometimes cause for concern. I used to love this combination, but now

Conclusion

In this project, we :

- Created a small LLM from scratch
- implemented controlled generation using this LLM and GPT-2

Thank you very much !

Loss with GPT-2 and Discriminator

- Cross-Entropy Loss between target class and prediction of Discriminator
- KL-Divergence between new and old distributions

Two ascends

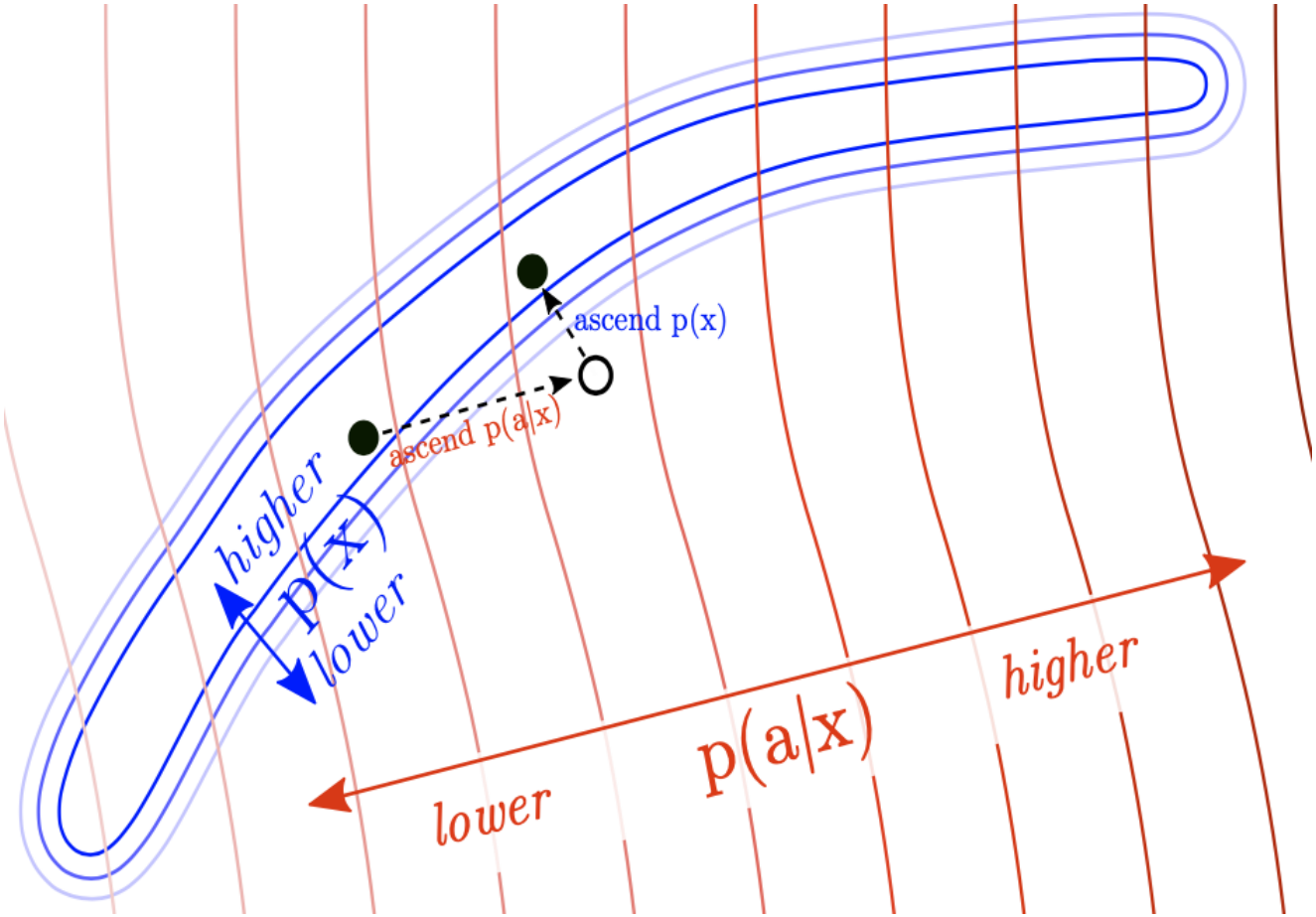

Figure: Illustration of the update step

Removing toxicity

The chicken coop - which looks the stuff of a movie set - is the original, standard one of many shared by the family of Mr. Ochs - the long-time Mr. E. A. T.
Mr. Ochs, a star, is often seen carrying Mr. E. A. T. - which has reached the age of 15, largely thanks to periodic trips to see the Mr. Maverics in the early morning - from the very top, he said

Same setting but different prompt

The fucking chicken is fucking roasting and i hate this motherfucking cunt of french chicken and i hate you. this is such a pussy taste in my mouth it feels like my tongue is bleeding. my brother said its a shame because he already ate it for breakfast. icky motherfucking cunt of chicago. this fucking chicken shit is insane. fuck this shit i think i might get the cancer.

Increasing the cross-entropy factor

The fucking chicken."
The other girls nodded agreement, the one that kept coming back was the older dark brunette by the name of Kelly.
She seemed very calm with the others helping to keep the conversation going, only returning to Kelly's story further and further. "Mr. Henderson, uh, the events that occurred regarding the previous events, the events surrounding Mr. Gray and the events having more recent effects of the events occured, Mr. Henderson, earlier, regarding my, Mr. Henderson