# PLUG AND PLAY LANGUAGE MODELS:A SIMPLE APPROACH TO CONTROLLED TEXT GENERATION.

(Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu.)

presented by Alexey Korolev

11 November 2020

# Content

1. Generative language model.
2. Plug and Play Language Model.
3. Autoregressive language model($P(x)$).
4. Attribute model $p(a|x)$.
5. Plug and Play Language Model training.
6. Results.

# Generative language model.

Nowadays big generative language model archive great performs. But if we want to specify their work some problem arises. We can't just finetune this model for our task because:

- This model is very big (over a billion parameters).
- require massive amounts of computing resources.
- on enormous data sets which are often not publicly released

Authors propose effective, easy to use method Plug and Play Language model. We need the next components:

1. autoregressive language model
2. attribute model.

We steer the enormous language model by gradients.

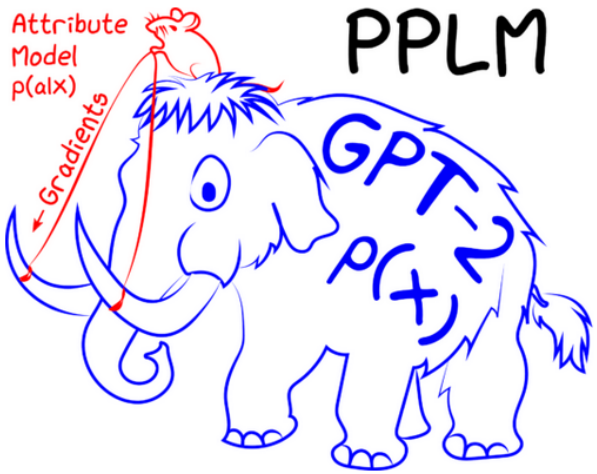



*Figure 1. How to steer a mammoth. Many attribute models used in PPLM are 100,000 times smaller than the language model (LM), roughly the weight ratio of a field mouse to a wooly mammoth. The PPLM method is plug and play: it can combine any generative neural language model (mammoth) and any differentiable attribute model or models (mouse) representing the desired steering objective(s). It is also resource efficient: the LM is used as-is without training or updating any of its weights (mammoths are hard to train, after all).*

# Autoregressive language model($P(x)$)

Autoregressive Model is merely a feed-forward model, which predicts the future word from a set of words given a contex. Start with your seed $x_1, x_2, \ldots, x_k$ and predict $x_{k+1}$.In formula that mean that we compute $P(x_i|x_{i-1}...x_{i-k})$ and choose the biggest  Mainly generative language model is Autoregressive(GPT1&2).

# Autoregressive language model($P(x)$)

In this work, we use the transformer approach.
A history matrix $H_t$ to consist of the key-value pairs from the past i.e $H_t = [(K_t^{(1)}, V_t^{(1)}), ..., (K_t^{(l)}, V_t^{(l)})]$, where $(K_t^{(l)}, V_t^{(l)})$ is key-value pairs from l-th layer generated at all time-steps from 0 to t.
To generate next we use formula: $o_{t+1}, H_{t+1} = LM(x_t, H_t)$
$x_{t+1}p_{t+1} = Softmax(Wo_{t+1})$ W is a linear transformation that maps the logit vector $o_{t+1}$ to a vector of vocabulary size.

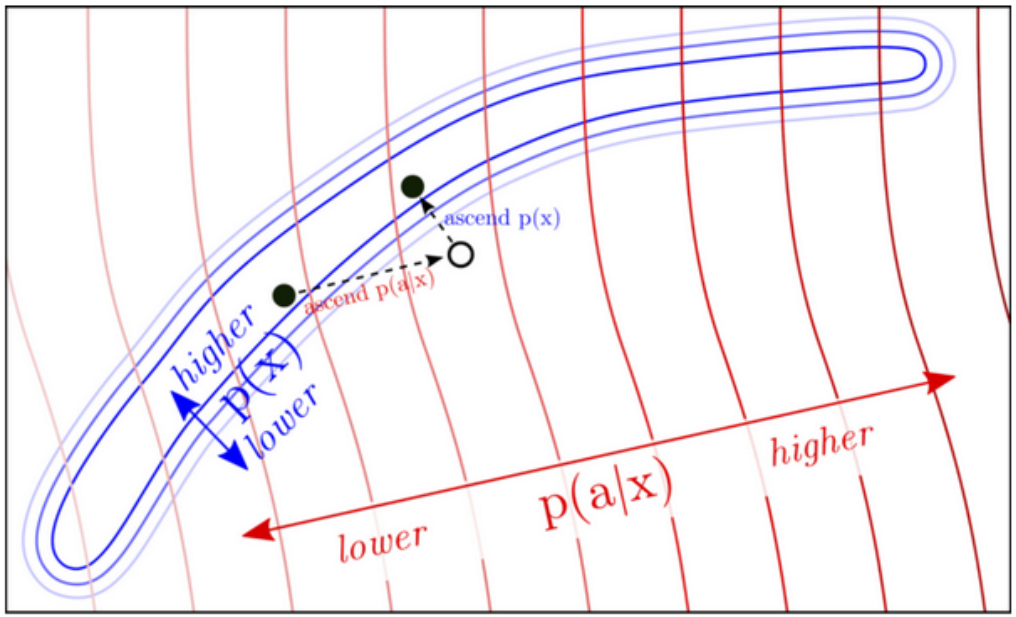
# Attribute model $p(a|x)$

Attribute model $p(a|x)$, which takes a sentence x and outputs the probability that it possesses the attribute a.These models can be tiny and easy to train because, intuitively, classification is easier.
As topic prediction use bag of word $((p(a|x) = \sum_i^k p_{t+1}[w_i]))$
For sentiment prediction use softmax classifier.

# Plug and Play Language Model training.

Main idea base on Bayes rule: $p(x|a) \sim p(a|x)p(x)$ There 3 steps:

1. Given a partially generated sentence, compute $\log(p(x))$ and $\log(p(a|x))$ and the gradients of each with respect to the hidden representation of the underlying language model.
2. Use the gradients to move the hidden representation of the language model a small step in the direction of increasing $\log(p(a|x))$ and increasing $\log(p(x))$.
3. Sample the next word.



# STEERING GENERATION:ASCENDING $log(p(a|x))$

Let $\Delta H_t$ be the update to $H_t$, such that generation with $(H_t + \Delta H_t)$ shifts the distribution of the generated text such that it is more likely to possess the desired attribute.

$\Delta H_t + \alpha \frac{\nabla_{\Delta H_t} \log p(a|H_t+\Delta H_t)}{||\nabla_{\Delta H_t} \log p(a|H_t+\Delta H_t)||^\gamma} \rightarrow \Delta H_t$ This update step can be repeated m times.

$o_{t+1}$ as $o'_{t+1}, H_{t+1} = LM(x_t, H'_t)$, where $H'_t = H_t + \Delta H_t$. The perturbed $o'_{t+}1$ is then used to generate a new distribution.

# ENSURING FLUENCY:ASCENDING $\log(p(x))$

- Kullback–Leibler (KL) Divergence: $D_{KL}(P||Q) = \sum \log(\frac{P(x)}{Q(x)})$
  We update $H_t$ to minimise the KL divergence between the output distribution of the modified and unmodified language models in addition to the step above.
- Post-norm Geometric Mean Fusion:
  Tie the generated text to the unconditional $p(x)$ LM distribution. We accomplish this by sampling from $x_{t+1} \sim \frac{1}{\beta}(\tilde{p}_{t+1}^{\gamma_{gmt}} p_{t+1}^{1-\gamma_{gmt}})$, where $p_{t+1}$ and $\tilde{p}_{t+1}$ are the unmodified and modified output distributions, respectively.

# Results

B: the baseline,unchanged GPT-2 LM, sampled once;
BR: B but sampled r times, with best sample chosen based on the LL ranking and filtering based on Dist score;BC: update the latent representations and then sample once; and lastly
BCR: update the latent representations and generate r samples,choose the best sample based on the LL score (after filtering out samples with low Dist scores).
CTRL: a recent language model.
WD:a weighted decoding baseline in which the B LM's outputs are weighted directly toward maximizing $p(a|x)$.

| Method | Topic % (↑ better) (human) | Perplexity (↓ better) | Dist-1 (↑ better) | Dist-2 (↑ better) | Dist-3 (↑ better) | Fluency (↑ better) (human) |
|---|---|---|---|---|---|---|
| B | 11.1 | 39.85±35.9 | 0.37 | 0.79 | 0.93 | 3.60±0.82 |
| BR | 15.8 | 38.39±27.14 | 0.38 | 0.80 | 0.94 | 3.68±0.77 |
| BC | 46.9 | 43.62±26.8 | 0.36 | 0.78 | 0.92 | 3.39±0.95 |
| BCR | **51.7** | 44.04±25.38 | 0.36 | 0.80 | 0.94 | 3.52±0.83 |
| CTRL | 50.0 | 24.48±11.98 | 0.40 | 0.84 | 0.93 | 3.63±0.75 |
| BCR | **56.0** | – | – | – | – | 3.61±0.69 |
| WD | 35.7 | 32.05±19.07 | 0.29 | 0.72 | 0.89 | 3.48±0.92 |
| BCR | **47.8** | – | – | – | – | 3.87±0.71 |

# Use

[-] The chicken is now out on the grill. \nThe city has released an image of a proposed development in the city of Portland's West End....

[Positive] The chicken was delicious – wonderfully moist, perfectly delicious, superbly fresh – and perfectly cooked. The only thing to say is that the sauce was excellent, and I think that the broth really complemented all of the other flavors. The best part was the sauce…

[Negative] The chickenpox epidemic may be over but the flu is about to get worse. The United States is facing one of the worst flu seasons on record and...

[-] The country's new chief minister, A.J. Paik, is a member of a group of prominent conservative politicians who have criticized the Obama administration's efforts to…

[Positive] The country's largest indoor painting event!\nCome celebrate with a dazzling display of stunning outdoor murals, a stunning display of art, and the world's best paint and art supplies from all over the world!

[Negative] The country's top prison system is forcing prisoners to use a trash dump, rather than a toilet, to flush their waste out, as the authorities fear the waste is more toxic and could cause cancer, an official at a major prison has revealed...

[Space] Prior to this year, we had no idea that the world could contain so many planets. N we know that they're out there, orbiting the star, and we've mapped them out with the hel NASA's Kepler space telescope.This is one of the world's most beautiful, planet-sized galaxies – a galaxy that is more than twice the diameter of our galaxy.

[Space] An illustration of a supermassive black hole at the center of our galaxy. Credit: ESO/C.C. Böhm-Bawerk, NASA/JHU/MSSS, J.F. Schönlein (Phys.org) —A black hole is object that has a very large and extremely dense galaxy, which makes it extremely unsta

[Space] The relationship between the United States and Russia is the largest military relationship that exists in space. Both nations are members of the International Space St (ISS), but are not directly orbiting the Earth. They use space stations for cargo and experiments to conduct science and research. Both countries are spaceflight pioneers an have developed space systems that provide spaceflight access for their citizens. Russian rockets are currently in...

[Space] The relationship between the two worlds was so complicated that it's hard to say was the original source. There must have been an earlier era where this planet was inhab or a planet of the same size existed. The only known planet on our own star is our sun, s there is no reason to be on another planet.The planet's surface temperature is about -27

language detoxification.
Controlled storytelling.

# Thank for attentions!

https://github.com/uber-research/PPLM