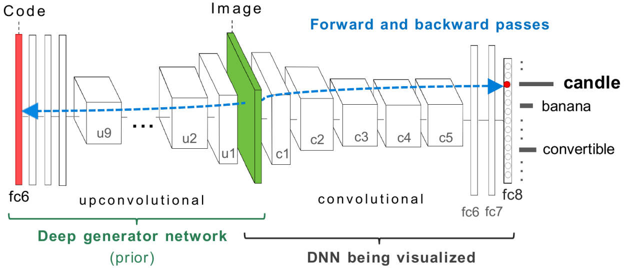
Plug and Play Language Models: A simple Approach too Controlled Text Generation

Dathathri et al., ICLR 2020

Outline

- Plug and play generative models (PPGNs), activation maximization (Nguyen+ NIPS 2016, CVPR 2017)
  - Generating images by performing gradient descent in the latent space of a generator network to maximize the activations of one or more attributes in a separate classifier network
- Plug and play language models (PPLMs) (Dathathri+ ICLR 2020)
  - Controlled language generation towards a list of words or sentiment
  - Plug and play, no fine-tuning on specific dataset

Plug and Play Generative Models (PPGNs), Activation Maximization



- General framework
  - Conditional model

$$p(h|y=c) \propto p(h)p(y=c|h)$$

  - Metropolis-adjusted Langevin sampler

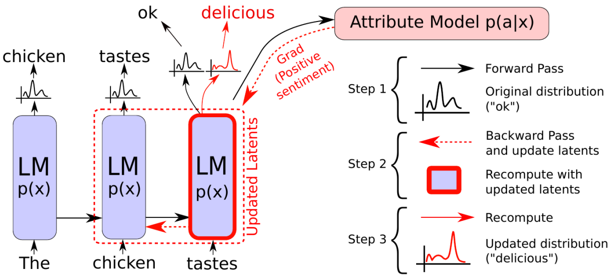
$$h_{t+1} = h_t + \alpha_1 \frac{\partial \log p(h_t)}{\partial h_t} + \alpha_2 \frac{\partial \log p(y=c|h_t)}{\partial h_t} + N(0, \epsilon_3^2)$$

Controlled Text Generation: Different Models

Model type	Form of model	Samples	Example models and number of trainable params
Language Model	$p(x)$	Uncond.	GPT-2 medium: 345M (Radford et al., 2019)
Fine-tuned Language Model	$p(x)$	Uncond.	Fine-tuned GPT-2 medium: 345M (Ziegler et al., 2019)
Conditional Language Model	$p(x\|a)$	Cond.	CTRL: 1.6B (Keskar et al., 2019)
Plug and Play Language Model (PPLM)	$p(x\|a) \propto p(x)p(a\|x)$	Cond.	PPLM-BoW: 0 (curated word list) PPLM-Discrim: $\sim$ 1K/attribute (not counting pretrained $p(x)$)

- Out-of-the-box LMs are capable of generating fluent text, but is not controlable
- Fine-tuning LMs must be fine-tune the whole language models on a specific dataset
- Conditional LMs, in Keskar+ work, design 50 different control codes and train the language model with these codes to generate desirable text, kinda expensive
- Plug and play LMs (PPLMs), combine plug in any language models and conditional classifier of choice, only sampling no fine-tuning the whole language models

Plug and Play Language Models



- Recurrent generating sequence

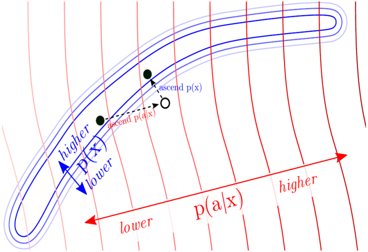
$$o_{t+1}, H_{t+1} = LM(x_t, H_t)$$

$$x_{t+1} \sim p_{t+1} = \text{Softmax}(W o_{t+1})$$

- Update $\tilde{H}_t = H_t + \Delta H_t$ to generate words according to desire attributes $\alpha_2 \frac{\partial \log p(y=c|x_t)}{\partial x_t}$

$$\Delta H_t = \Delta H_t + \alpha_2 \frac{\nabla \log p(a|H_t + \Delta H_t)}{\|\nabla \log p(a|H_t + \Delta H_t)\|}$$

Plug and Play Language Models



- Update $\tilde{H}_t = H_t + \Delta H_t$ to generate words according to desire attributes $\alpha_2 \frac{\partial \log p(y=c|x_t)}{\partial x_t}$

$$\Delta H_t = \Delta H_t + \alpha_2 \frac{\nabla \log p(a|H_t + \Delta H_t)}{\|\nabla \log p(a|H_t + \Delta H_t)\|}$$

- Also, update ΔH_t to minimize the KL divergence between output distribution of the modified and unmodified language models $\alpha_1 \frac{\partial \log p(x_t)}{\partial x_t}$ (not sure how it is done in this work)
- Another trick, sampling $x_{t+1} \sim \frac{1}{\beta}(\tilde{p}_{t+1}^{\gamma} p_{t+1}^{1-\gamma})$

Attribute Models

- BoW attribute models with a set of keywords $\{w_1, ..., w_k\}$

$$\log p(a|x) = \log(\sum_{i}^{k} p_{t+1}[w_i])$$

- Sentiment classifier models

$$\log p(a|x) = \log f(o_{:t+1})$$

Experimental Examples: BoW Attribute Models

Table 3: Comparison of different samples generated by (top row) baseline GPT-2 and (other rows) PPLM with different BoW corresponding to different topics (e.g. **[Military]**), all conditioned on a single prefix: "The issue focused". Both directly optimized (in red) and related words (in soft red) are highlighted, showing how the optimization takes effect.

[–] The issue focused on the way that the city's police officers have reacted in recent years to the deaths of Michael Brown in Ferguson, Mo., Eric Garner in New York City and Sandra Bland in Texas, as well as the shooting of unarmed teen Michael Brown by a white police officer in Ferguson, Mo. ...
[Military] The issue focused on the fact that the government had spent billions on the military and that it could not deploy the troops in time. The prime minister said that the country would take back control of its airspace over Syria in the next 48 hours. \n The military is investigating why...
[Space] The issue focused on a series of incidents that occurred in the past few months, which included an alleged attack by Islamic State fighters on a Kurdish checkpoint, the use of drones in combat, space technology research by Russian and American space companies, and more. \n The world...
[Science] The issue focused on a single piece: the question "What is the meaning of life?" This question has puzzled many philosophers, who have attempted to solve it by using some of the concepts of quantum mechanics, but they have to solve it by the laws of nature themselves....
[Politics] The issue focused on a single section of the legislation. It's unclear whether the committee will vote to extend the law, but the debate could have wider implications. \n "The issue of the law's applicability to the United Kingdom's referendum campaign has been one of...
[Computers] The issue focused on the role of social media as a catalyst for political and corporate engagement in the digital economy, with the aim of encouraging companies to use the power of social media and the Internet to reach out to their target market. \n ...

Experimental Examples: Sentiment Classification Models

Table 5: Sentence samples in triplets, generated by {baseline GPT-2, PPLM-Discrim POSITIVE, PPLM-Discrim NEGATIVE}, conditioned on prefixes: The chicken & The country. Words related to the sentiment are highlighted (in soft red). Each triplet is generated from the same random seed.

[-] The chicken is now out on the grill. \n The 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! \n Come 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....

Experimental Results

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

Method	Sentiment Acc. (%) (human)	Sentiment Acc. (%) (external classifer)	Perplexity (↓ better)	Dist-1 (↑ better)	Dist-2 (↑ better)	Dist-3 (↑ better)	Human Evaluation Fluency (↑ better)
B	19.3	52.2	42.1±33.14	0.37	0.75	0.86	3.54±1.08
BR	41.5	62.2	44.6±34.72	0.37	0.76	0.87	3.65±1.07
BC	39.6	64.4	41.8±34.87	0.33	0.70	0.86	2.79±1.17
BCR	**73.7**	**78.8**	46.6±40.24	0.36	0.77	0.91	3.29±1.07
CTRL	**76.7**	96.6	37.4±16.89	0.35	0.78	0.89	3.54±0.77
BCR	70.0	–	–	–	–	–	3.36±0.82
GPT2-FT-RL*	13.3	77.8	217.3±176.4	0.54	0.91	0.94	3.31±0.84
BCR	**84.4**	–	–	–	–	–	3.68±0.83
WD	18.9	52.2	31.7±28.0	0.33	0.69	0.83	3.67±0.89
BCR	**61.1**	–	–	–	–	–	3.75±0.66

Thank you !