Unsupervised Approach for Text Generation

Sai Kumar Kayala
Problem Statement

- Understanding Text?
  - What does it mean?
- Predict next word from history
- How does it help generate text?
- Can the generated text have a style? Yes!
- Fine-Tuning to generate Shakespeare plays.

Can you please come here?
Language Model

- Goal is to find unsupervised distribution estimation given a set of words

\[ p(x) = \prod_{i=1}^{n} p(s_n|s_1, \ldots, s_{n-1}) \]

- A similar idea can be used to predict next ‘k’ words

\[ p(s_{n-k}, \ldots, s_n|s_1, \ldots, s_{n-k-1}) \]
Attention

- Scaled dot product attention

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]

- Multi head self Attention

\[
\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^Q
\]

where \(\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\)
GPT-2

- OpenAI’s language modelling algorithm.
- Unsupervised Learning

\[ h_0 = UW_c + W_p \]
\[ h_i = \text{transformer\_block}(h_{i-1}) \forall i \in [1, n] \]
\[ P(u) = \text{softmax}(h_n W^T_e) \]

\[ L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta) \]

- Other supervised fine-tuning applications
**Dataset**

**Pretrained GPT-2**
- Pre-trained on all 3-karma reddit posts
- Nearly 40GB of cleaned data
- Have multiple models ranging from 117M to 1.5B parameter models
- Still all of these models underfit the data

**Fine-tuning**
- Got all the 36 plays
- Cleaned them from
  - Titles
  - ‘*’ tokens
  - Blank tokens
- Encoded for the model
Training Stats

- **Hardware**
  - i7-8700K
  - Gtx 1080 Ti

- **Time**
  - Around 8hrs
  - About 44000 iterations
CELIA
  Come, come, you mock me; this is not the fashion of the commonwealth. Let an old man embrace his weary neck, and let his tongue laugh at; and let not him that plays the touch for grace, but let desert tarry his hour, which perforce robs him, and falls on haime: therefore let me have right, and let desert mount.

CELIA
  Come, come, you jocund, fiery, saucy, non-gracious, inscrutable, and affecting fantastic instruments!

CELIA
  Peace, peace!

DUKE
  What, dost thou?

CELIA
  Why, for my life, but let thine be seen so by the gazing moon, upon the highmost centre Heave the murderous star, envy his own set To take my sight and away my sight: I would thou didst but whet me with my legs.

DUKE
  What wilt thou, thou?

CELIA
  That I did think, but thou didst not know, What pride is, what pride, and furious age, The c degree and dignity of this squain: I would thou didst, and thou didst.

DUKE
  Why, man?
Facebook Messages

115067385088) Sai Kumar: chepu
(137009173389) Singamaneni Viswajith: Haha rey chepu
(137009174905) Singamaneni Viswajith: Enti akkada nokku cheptha?
(1370091761878) Sai Kumar: sarelele inka raa
(137009178650) Singamaneni Viswajith: Ennintiki vacha...
(137009178799) Singamaneni Viswajith: Oka sari cheptha raa
(1370091704878) Singamaneni Viswajith: Enti ra
(1370091785016) Sai Kumar: hahaa
(1370091788301) Singamaneni Viswajith: Anduku raa..
(1370091813705) Singamaneni Viswajith: Naku telusu raa
(1370091814984) Singamaneni Viswajith: Naku chala mandi anaduu raa
(1370091826685) Singamaneni Viswajith: Chepa raa
(1370091834755) Sai Kumar: chepu
(1370091881419) Singamaneni Viswajith: Ha adi cheptha
(1370091886759) Singamaneni Viswajith: I just want to say..
(1370091897014) Singamaneni Viswajith: Chepa raa
(1370091863676) Singamaneni Viswajith: Haa chepa raa
(1370091871582) Singamaneni Viswajith: Chepa raa
(1370091881393) Singamaneni Viswajith: Cheptha le
(1370091887974) Sai Kumar: haha
(1370091903243) Singamaneni Viswajith: Vadini lekapothe chepara..
(1370091903243) Singamaneni Viswajith: Raraa
(1370091912919) Singamaneni Viswajith: Bada dorakaru..
(1370091928985) Singamaneni Viswajith: Aavida sai...
(1370091946354) Singamaneni Viswajith: Haha..
(1370091953609) Singamaneni Viswajith: Ee branch of law
(1370091971293) Singamaneni Viswajith: Andulo roju koncham koncham vallu adokkudadu.
(1370091972923) Singamaneni Viswajith: Repu exam aipondi..
Thank You & Questions