THE UNIVERSITY OF SYDNEY

Title : Neural Ordinary Differential Equations Based Language Modelling

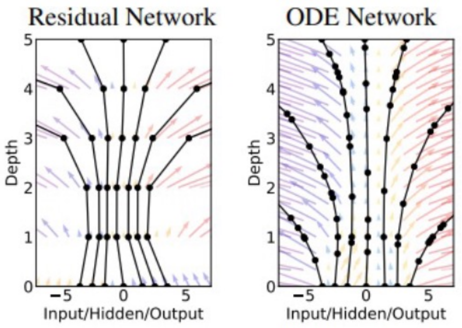
Name + SID Hongbo Du (500635346) Yuhao Hu (470139936) Liam Watts (510562348)

COMP5329 Deep Learning

Description

We constructed Language Models using a novel form of neural network: Neural Ordinary Differential Equations[1]. The Neural ODEs are continues analogue of a ResNet.

Our goal is to apply Neural-ODEs to the language modelling task, evaluating their performance compared to baselines such as LSTM and GPT-2 to find possible improvements.



Hidden state of ResNet

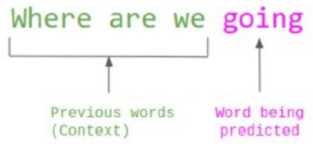
Hidden state of Neural ODE

$\mathbf{h}_{t+1} = \mathbf{h}_t + f(\mathbf{h}_t, \theta_t)$

$\frac{d\mathbf{h}(t)}{dt} = f(\mathbf{h}(t), t, \theta)$

[1]

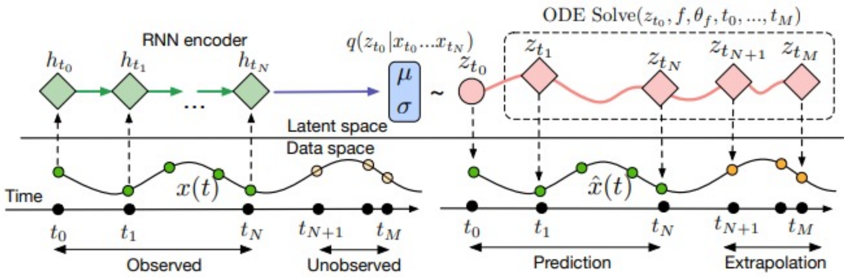
The above figure shows the comparison between the hidden state of a discrete-depth neural network versus a continuous-depth neural network.



Language models predict the most likely continuation of a text. They have other applications in dialogue systems and Natural Language Processing in general.

Model Architecture

We studied how Neural ODEs based models perform in language modeling field by reconstructing a new ODE-LSTM model and proposing an ODE-GPT-2 and Augmented-ODE-GPT model by using the existing Neural ODE framework, such as TorchDyn and other language modeling frameworks. Our model architectures involve the classical state-of-the-art models as encoder and outputs the generated words by Neural ODEs model. The ODE models are trained via adjoint method. We also compared the performances between our ODE-based language models and the classical state-of-the-art language models such as LSTM and GPT-2.



[1]

ODE-Based Model Architecture

Datasets

Dataset	Size	SoTA Perplexity
Penn Treebank	~1 million words	20.5 from GPT3
WikiText-2	~2 million words	18.34 from GPT2

Evaluation Metrics

There are two dimensions for evaluation the model

1. Perplexity: confusion of wording.
2. BLUE (Bilingual Evaluation Understudy): comparison between generated sentence and referenced sentence.

Results

Model	Dataset: Penn TreeBank			Dataset - WikiText2		
	Perplexity (train/test)	BLEU (train/test)	# of Params	Perplexity (train/test)	BLEU (train/test)	# of Params
Base-LSTM	212 / 266	**0.295** / **0.228**	3.73 M	171 / 243	0.610 / 0.359	10.7 M
ODE-LSTM	232 / 253	**0.296** / 0.225	3.91 M	224 / 266	**0.612** / 0.357	10.9 M
Base-GPT	**21.57** / **23.95**	0.235 / 0.225	163 M	**15.46** / **16.77**	0.553 / **0.545**	163 M
ODE-GPT	179 / 173	0.177 / 0.179	164 M	303 / 221	0.387 / 0.375	164 M
Augmented-ODE-GPT	169 / 158	0.189 / 0.190	206 M	---	---	---

Ablation study

Model	Perplexity (train/test)	BLEU (train/test)
Full Model	**169** / **158**	**0.189** / 0.190
- Zero Intialisation	176 / 165	0.188 / **0.191**
- Teacher forcing analogue	185 / 173	0.182 / 0.186
- No Fine Tuning on GPT	177 / 165	0.185 / 0.185
- Not Augmented	179 / 173	0.177 / 0.179

Reference

[1] Chen, R.T.Q., Rubanova, T., Bettencourt, J., and Duvenaud, D. (2018). Neural ordinary differential equations. Advances in Neural Information Processing Systems.