Text Classifications Learned from Language Model Hidden Layers

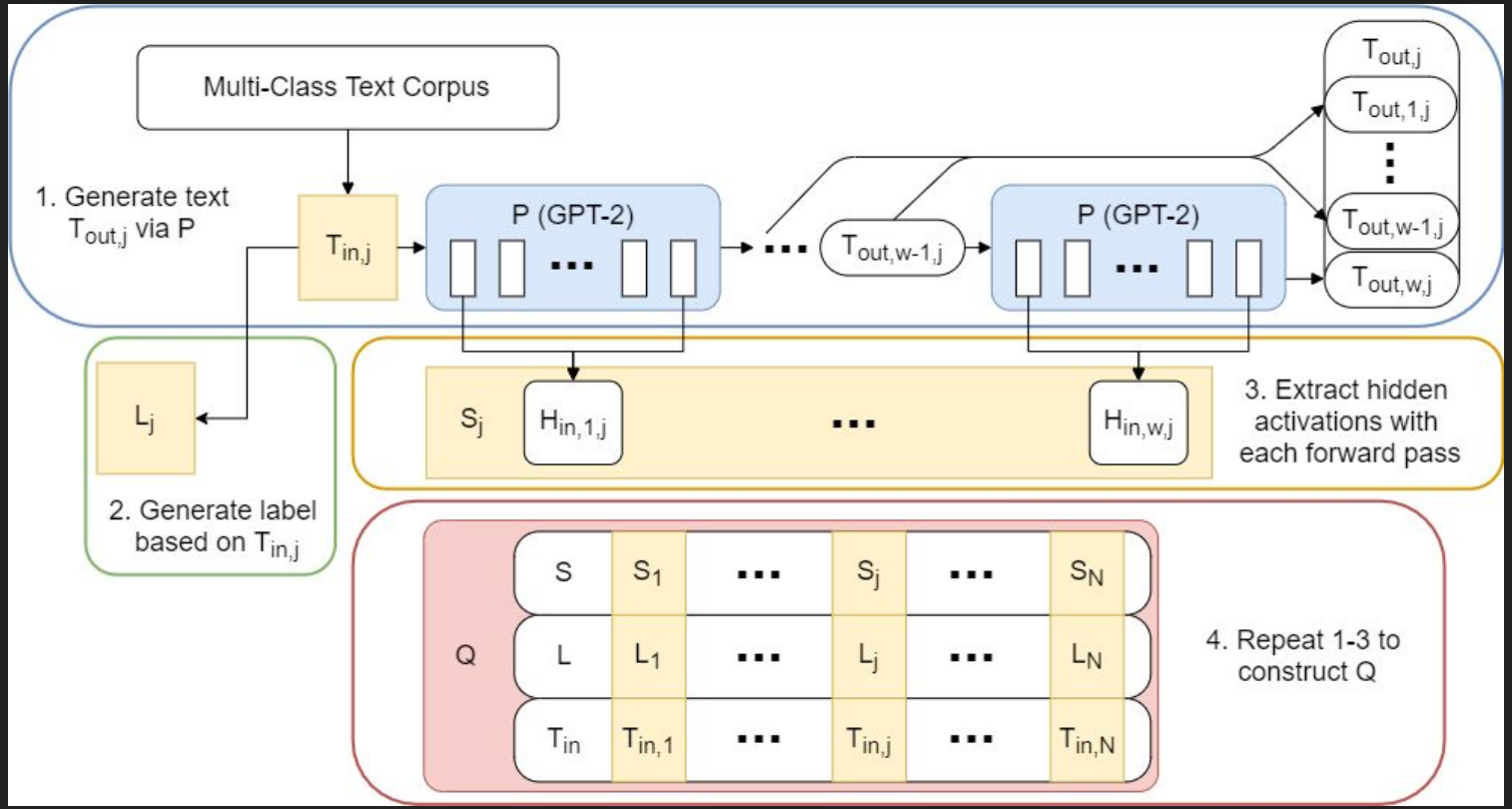
Nathaniel Robinson, Zachary Brown, Timothy Sitze, Nancy Fulda

dragn.ai

Motivation

- Language models are uninterpretable
- PR disasters
  - Microsoft's Tay[4]
- Classifiers used in controllable text generation
  - Plug and Play Language Models[1]
  - Neural Programming Interfaces[2]
  - The above interface with OpenAI's GPT-2[3]

Data Collection

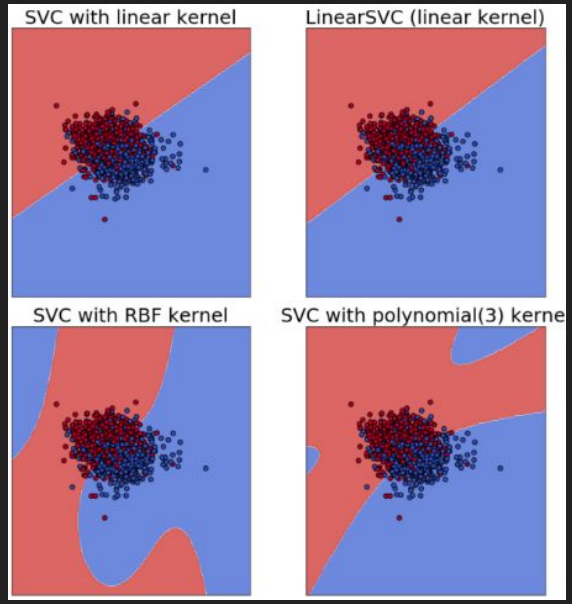


Classification of offensive and non-offensive text

Network type	Notable parameters	Acc.
Original feed-forward	3 dense layers, 112 neurons, batch size 5	.9138
Original feed-forward with larger batches	**3 dense layers, 112 neurons, batch size 20**	**.9155**
Original feed-forward trained on 14x1024 arrays	3 dense layers, 112 neurons, data shape 14x1024	.8959
Wider feed-forward	3 dense layers, 1792 neurons	.9084
Shallower feed-forward	1 dense layer, 64 neurons	.8822
Deeper feed-forward without skip connections	7 dense layers, 1016 neurons	.9029
Deeper feed-forward with skip connections	13 dense layers, 1071 neurons, residual connections	.9015
Convolutional neural network	9 layers, 7 batch norms	.9008
Random Forest	Max depth 4, max 122 features, data shape 14x1024	.8768

Classification of offensive and non-offensive text

Network type	Notable parameters	Acc.
Original feed-forward	3 dense layers, 112 neurons, batch size 5	.9138
Original feed-forward with larger batches	**3 dense layers, 112 neurons, batch size 20**	**.9155**
Original feed-forward trained on 14x1024 arrays	3 dense layers, 112 neurons, data shape 14x1024	.8959
Wider feed-forward	3 dense layers, 1792 neurons	.9084
Shallower feed-forward	1 dense layer, 64 neurons	.8822
Deeper feed-forward without skip connections	7 dense layers, 1016 neurons	.9029
Deeper feed-forward with skip connections	13 dense layers, 1071 neurons, residual connections	.9015
Convolutional neural network	9 layers, 7 batch norms	.9008
Random Forest	Max depth 4, max 122 features, data shape 14x1024	.8768



Classification of cat- and non-cat-sentences:

- 99.9% accuracy (feed-forward NN)

Classification of cat- and non-cat-sentences:

	Sentence	Class	Model output
1	dogs and cats prefer to play together in packs with their cubs	CAT	1.03
2	children prefer to play together in groups with their toys	NO CAT	.40
3	children prefer to play together in groups with their cats	CAT	1.00
4	the film was set in the seventeenth century. A time of war-torn	NO CAT	-.07
5	the film was set in the seventeenth century. A time of small and large cats everywhere	CAT	.72
6	the very feline tiger purred and cleaned her tail and whiskers for her cubs	CAT	.96
7	the very human man groaned and cleaned his hair and mustache for his kids	NO CAT	.01
8	the very human man groaned and cleaned his hair and mustache for his cats	CAT	.99
9	she had feline habits and purred and meowed often	CAT	.97
10	the little furball meowed, grabbed her cub, and slinked away	CAT	.86
11	the little dude yawned, grabbed his friend, and skipped away	NO CAT	.12

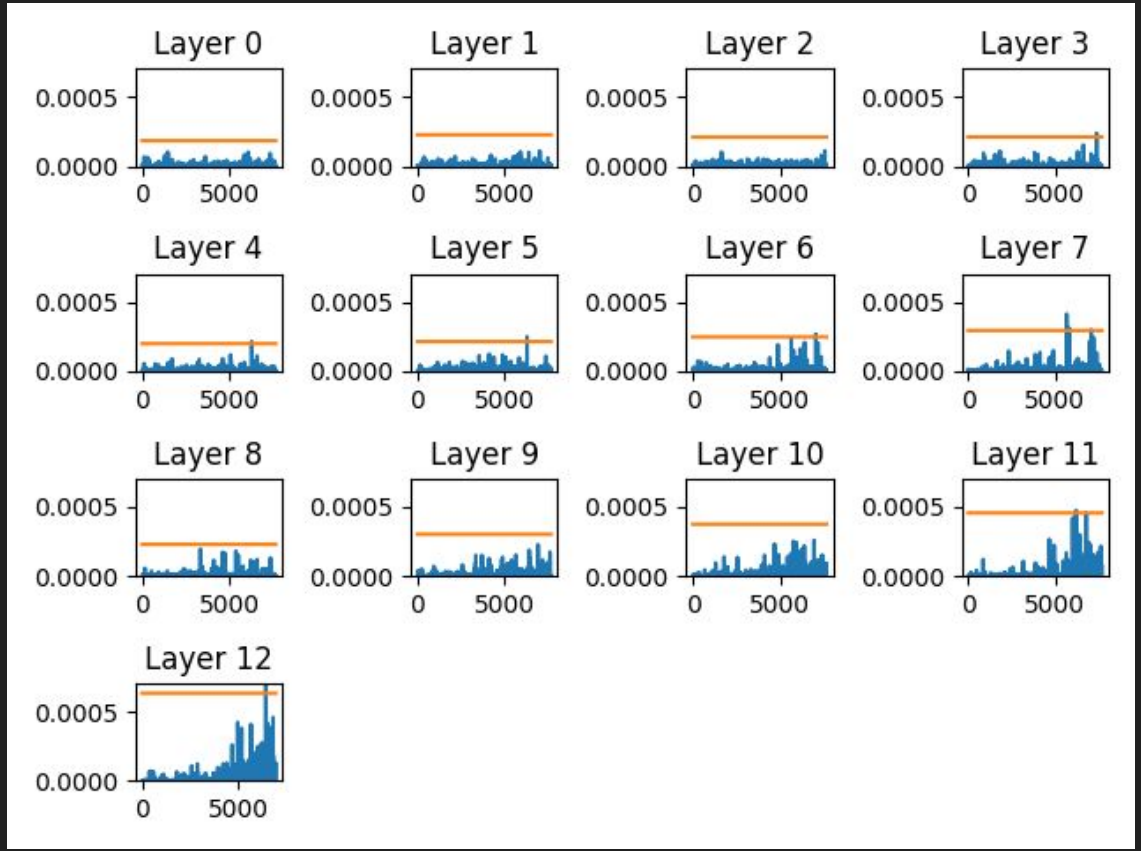
Could text representations from language model hidden layers be used in place of more traditional embeddings?

Comparison: Universal Sentence Encoder[5]

- Traditional full-sentence embeddings fail at classification task

Representations used	Network type	Notable parameters	Acc.
GPT-2 activations	**Original feed-forward**	**3 dense layers, 112 neurons**	**.9117**
U.S.E. embeddings	Original feed-forward	3 dense layers, 112 neurons	.5000
U.S.E. embeddings	Deeper feed-forward	6 dense layers, 504 neurons	.5000
U.S.E. embeddings	Random Forest	Max depth 3, max features: "auto"	.4401

A note on choice of hidden layers



Conclusion

- Simple feed-forward networks without residual connections are sufficient for classification of language model hidden layers

- Classifiers glean beyond word-level semantic information from representations

- These deep text representations may be preferable to traditional embeddings in some applications

Thank you

[1] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation, 2019.

[2] Zachary Brown, Nathaniel Robinson, David Wingate, and Nancy Fulda. Towards neural programming interfaces. In Advances in Neural Information Processing Systems, volume 33, 2020.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9, 2019.

[4] Gina Neff and Peter Nagy. Talking to bots: Symbiotic agency and the case of tay. International Journal of Communication, 10:4915–4931, 10 2016.

[5] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. arXiv preprint arXiv:1803.11175, 2018.