Better Caption Embeddings Projections in PixelCNN++

Making better use of caption representations in the PixelCNN+ conditional generative model with multiple tiers of granularity

Background

Text-to-image synthesis (or caption-to-image) is of great academic and industrial interest, and is a natural application of generative models using latent variables. The conditional PixelCNN, introduced by Van Der Oord et al. (2016b) [1] is an autoregressive model that models each conditional distribution with masked convolution layers. Inspired by the success of an originally proposed variant, PixelRNNs, PixelCNN++ leverages gated activation unit to model more complex conditional dependencies.

Brief Summary of Work

In the original paper, the caption embeddings are "projected" into each convolution layer by the gated activation unit,

$$y = \tanh(W_{k,f} * x + V_{k,f}^T h) \odot \sigma(W_{k,g} * x + V_{k,g}^T h).$$

From here, I extended the linear projected term into a new set of hyperparameters,

$$\{f(h)|y_k = \tanh(W_{k,f} * x + f(h)) \odot \sigma(W_{k,g} * x + f(h))$$

.

I ran experiments with projection choices of gated linear, linear + ReLU, and shallow NN:

$$f(h) = (W^T h)\sigma(V^T h + b)$$

$$f(h) = ReLU(V^T h + b)$$

$$f(h) = ReLU(W_2^T \sigma(W_1^T h + b_1) + b_2)$$

.

Finally, I evaluated the results on key, interpretable generative model metrics - BPD, IS, and FID, look at sample generations, and discuss the implications.

Technical Results on ImageNet32 with BERT Embeddings

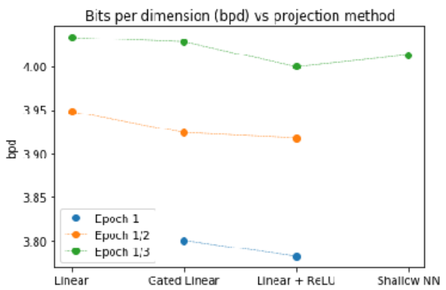
	Epochs trained	Projection Method	Val BPD	IS	FID
Tier 1	1	Gated Linear	3.800	1.493	439.58
		Linear + ReLU	3.782	1.481	454.83
		Shallow NN	TBA	TBA	TBA
	1/2	Gated Linear	3.924	1.507	421.90
		Linear + ReLU	3.918	1.486	416.75
		Shallow NN	TBA	TBA	TBA
Tier 2	1/3	Gated Linear	4.028	1.486	429.79
		Linear + ReLU	4.000	1.447	433.89
		Shallow NN	4.013	1.470	407.25

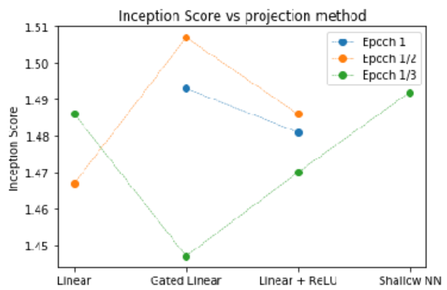
Technical Disclaimers

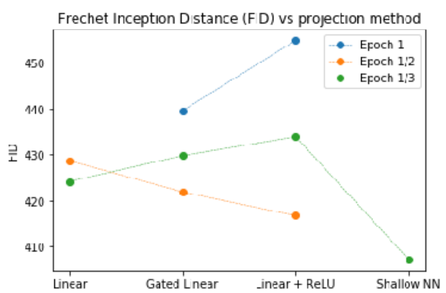
1. Due to inhumane budget constraints, some models were trained only on 1/2 or 1/3 epochs, and the ones marked TBA will be resolved by report submission.
2. Training was done on size=32 batches under the same random seeds (reproducible).
3. Evaluation was done on a randomly chosen size=32 batch.

Baseline Benchmarks

	Epochs trained	Projection Method	Val BPD	IS	FID
Tier 1	2	Linear	3.840	1.484	423.58
	1	Linear	3.819	1.461	423.66
	1/2	Linear	3.948	1.467	428.66
Tier 2	1/3	Linear	4.033	1.492	424.19







Discussion & Conclusion

I make the following observations consistent across all three epoch groups:

1. The linear + ReLU projection method resulted in the lowest validation bpd, suggesting it's the optimal projection method.
2. The gated linear projection method had the highest IS score, suggesting its "gate" functionality achieves the best of both worlds - sharpness and expressivity.
3. The baseline method (linear) retains the lowest FID scores, although the only model trained with shallow NN may prove otherwise. It seems linear and shallow NN projections are most "consistent" with the deep NN features (InceptionV2) used to evaluate FID.

These observations give me clear insight into how designing the method of projection contribute to the robustness of the model. In each epoch group, every projection method surpassed the val bpd of the baseline, and corroborates my hypothesis that increasing the expressivity of the projection reduces model bias (and with the exception of the shallow NN model, variance as measured by val bpd as well)

I also gained insights into how we can use the projection method to optimize for different, common and interpretable evaluative metrics. The third observation also reminds us it's not enough to consider only the expressivity, but also how the projection fits in as a feature to the model architecture.

Future

The results clearly show the method of projection on contribute significantly to not only model robustness but is also flexible to different objectives. A natural extension may be to discover relationships between the projection method's analysis (i.e. # params, # non-linearities) and objective functions.

Unfortunately, label-based word2vec carries only so much semantic meaning. It would be interesting to explore more nuanced caption-image conditioning using transformers (attention heads on extended caption phrases and attention heads on decoded image pixels).

Contact Information

Michael Sun
Stanford University
Email: msun415@stanford.edu
Website: shininingsunnyday.com/portfolio

References

[1] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. *arXiv preprint arXiv:1606.05328*, 2016c.

[2] Tim Salimans, Andrej Karpathy, Xi Chen, Diederick P. Kingma. Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. *arXiv preprint arXiv:1701.05517v1, 2017.*