THE HILARIOUS BLUFFING GAN

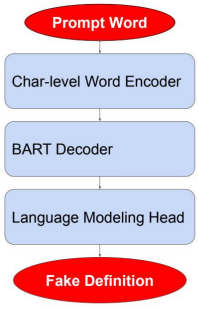
Training an AI to play Balderdash

The Game

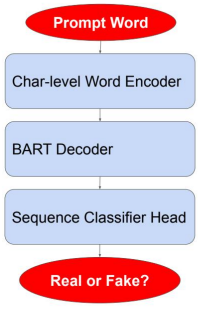
Balderdash is a game where two or more players invent fake definitions for an obscure word. Then all fake definitions are presented along with the true definition, and each player must guess which definition is real.

Models

Generator



Discriminator



Contrastive Loss

I introduce a novel contrastive loss function for training language modeling as a GAN.

	Normalized Generator Score	Normalized Discriminator Score	Loss
Definition 1	1.0	0.5	0.5
Definition 2	0.7	0.8	0.1
…			
Definition N	0.0	0.7	0.7

Illustration of the Contrastive Loss Function for Text Generation

Although the whole text generation process uses discrete tokens and is not differentiable, the process involves the generator producing a score for each sequence, which *is* differentiable.

The discriminator is used to score each of a set of generated definitions, which can be seen as a set of 'comparative' or 'contrastive' labels. The generator scores are normalized to the same range, and the loss is calculated as the absolute difference in normalized scores.

Training Regime

- Pretrain a BART decoder on 0.1% of Wikipedia
- Fine-tune encoder, cross-attention, LM head and Sequence Classification head on obscure words dataset
- Further co-train the Generator and Discriminator as a GAN.

Key Conclusions

- It is difficult to avoid text degeneracy when training the GAN.
- Even a very primitive one-hot char encoder can lead to the model learning important word-parts.
- When overtrained, the generator tends to regurgitate definitions from the training dataset
- Pretraining the LM head on Wikipedia is essential for producing sensible original definitions.

ailurophobia	fear of peanut butter
rumfustian	the making of an excuse for not appearing
xanthocomic	having the character of a whirlwind
ophelimity	application of principles to past events

AI-generated definitions for real words

praxeology	study of mental disorders and the state of being in a coma
chronography	writing about the origins of religions or belief systems
synaesthesis	having the power or ability to express feelings without blushing
jiggumbob	to cause to work more energetically than usual

AI-generated definitions for real words

syndetic	belief that matter travels on time, having an eternal velocity
mignon	excessive devotion to France or the French
inverecund	inability to fall asleep due to stress or strain
whereout	to disburden by stripping a rope of its sheath

AI-generated definitions for real words