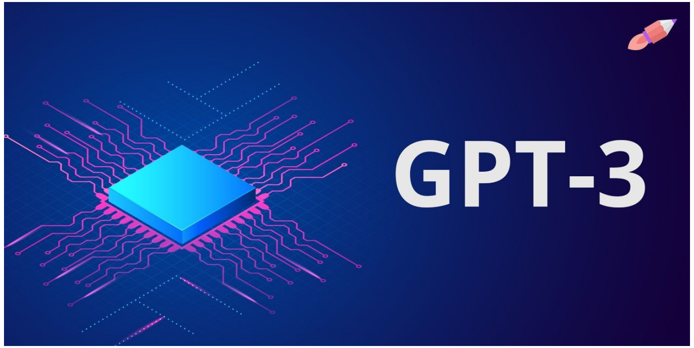
Sentence Generation and Classification with Variational Autoencoder and BERT

Geeling Chau, Anshuman Dewangan, Jin-Long Huang, Keshav Rungta, Margot Wagner

Introduction

- **Natural-language generation**: transformation of structured data into natural language, in this case using AI techniques
- **Current SOTA**: *GPT-3* by OpenAI -- 175 billion parameters (May 2020)
- **Problem**: text generation can tend to be contradictory
- **Goal**: generate text responses with different levels of contradictoriness



The Dataset

“Contradictory, My Dear Watson” by Kaggle:
- Over 12k unique pairs

Stanford Natural Language Inference (SNLI) :
- Over 570k unique pairs

- Given a sentence pair (a premise and a hypothesis) there are 3 ways they could be related: “**This church choir sings to the masses as they sing joyous songs from the book at a church.”**
  1. One sentence entails the other (entailment)
     a. The church is filled with song.
  2. The sentences are neutral but related (neutral)
     a. The church has cracks in the ceiling.
  3. One sentence contradicts the other (contradiction)
     a. A choir singing at a baseball game.

Data Processing

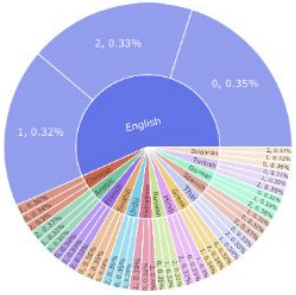


Figure 1: Language distribution in MDW dataset.

Kaggle MDW:

- Kaggle dataset contains sentences from many languages
- Since only 56% of them are in English, we translate all of them to English using the pytransgoogle library provided by Google Translate API

SNLI:

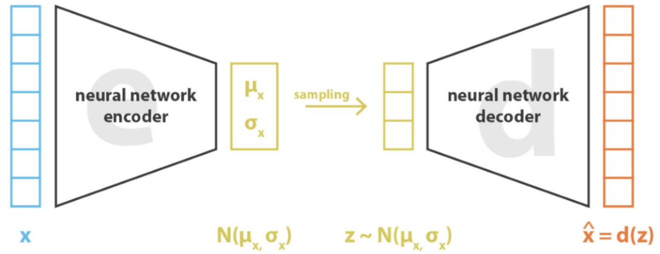
- Used HuggingFace Datasets (🤗Datasets) to process the original dataset to work with our original dataloader one for MDW.

The Goal

1. Create 1 variational autoencoder that generates a hypothesis given a premise with no regard for class label
    a. Metric: BLEU scores compared to dataset
2. Create 3 variational autoencoders, one for each class, to generate a hypothesis given a premise of that label
    a. Metric: BLEU scores compared to dataset and accuracy score when passed through BERT classifier
3. Create a conditional variational autoencoder that generates a hypothesis given a premise and class label
    a. Metric: BLEU scores compared to dataset and accuracy score when passed through BERT classifier

Hypothesis Generation -- Variational Autoencoder (VAE)

- Variational autoencoder (VAE): generative version of basic autoencoder
- LSTM encoder and decoder



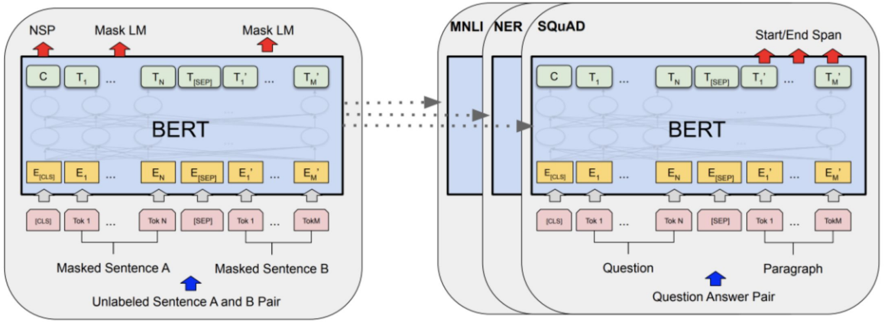
$$\text{loss} = \|x - \hat{x}\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)]$$

- Encoder
  - Input: premise sentence
  - Output: embedding in Gaussian latent space
- Decoder
  - Input: embedding
  - Output: hypothesis sentence
- Reparameterization trick to enable backprop

Hypothesis Classification -- BERT

- We used pretrained BERT provided by 🤗

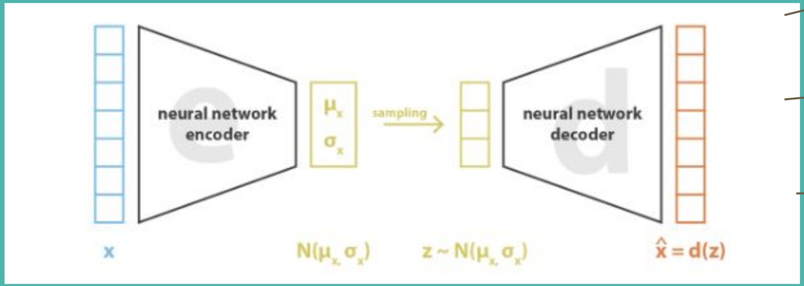
- Bidirectional representation
- **Attention** mechanism



Pre-training
Fine-Tuning

- Input: concatenated sentence pairs
- Output: sentence relationship class

~ Results ~
Variational AutoEncoder (VAE) Sentence Generator



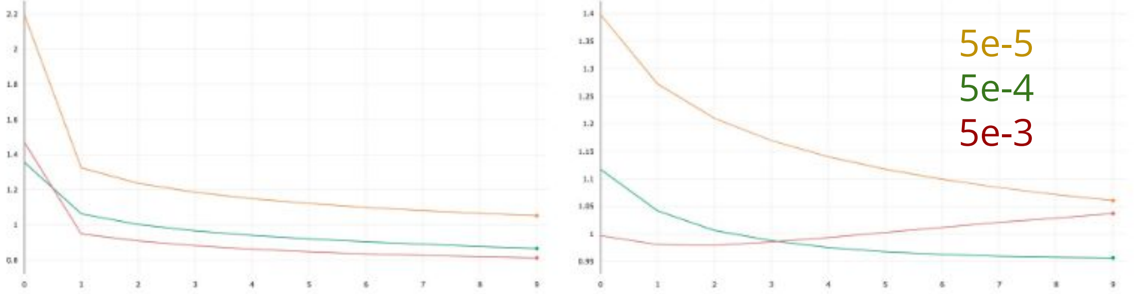
A man is holding a child.

The man is a professional musician.

The man is sitting on a couch.

VAE HP Tuning + Data Augmentation

Learning Rate: 5e-4



(a) Training Loss (b) Validation Loss

Data Augmentation

Experiment	Test Loss	BLEU-1	BLEU-4
MDW English Only	2.754	**34.73**	**5.90**
MDW w/ Translations	2.47	21.44	2.91
SNLI	**0.957**	26.0	3.74

How close can we get our generated sentences to be?

1. Class Agnostic

Test Loss	BLEU-1	BLEU-4
0.957	**26.0**	**3.74**

A man in a black hat opens his mouth -> <u>a man is looking at a camera</u>

A young infant cries while having his or her pajamas button -> <u>a man is standing outside</u>

How close can we get our generated sentences to be?

1. Class Agnostic

Test Loss	BLEU-1	BLEU-4
0.957	**26.0**	**3.74**

A man and a child are laughing at each other.
Predicted Entailment: A man is holding a child

2. Class Specific
   a. 0: entail
   b. 1: neutral
   c. 2: contra

0.846	25.8	3.74
1.164	25.9	3.74
0.967	24.6	3.48

A man talking into a microphone with a woman standing next to him.
Predicted Neutral: The man is a professional musician

A man wearing a white shirt and a blue jeans reading a newspaper while standing.
Predicted Contradiction: The man is sitting on the couch.

How close can we get our generated sentences to be?

1. Class Agnostic

Test Loss	BLEU-1	BLEU-4
0.957	**26.0**	**3.74**

2. Class Specific
   a. 0: entail
   b. 1: neutral
   c. 2: contra

0: entail	**0.846**	25.8	3.74
1: neutral	1.164	25.9	3.74
2: contra	0.967	24.6	3.48

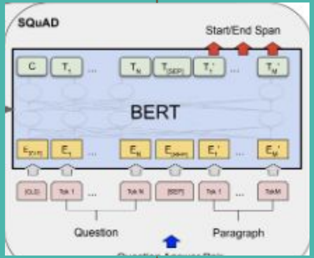
A man on a bicycle rides past a park, with a group of people in the background.
Predicted Contradiction: The man is sitting on the couch.

3. Class Conditional

0.954	26.0	3.73

~ Results ~
BERT Classification of Premise + Hypothesis Pairs

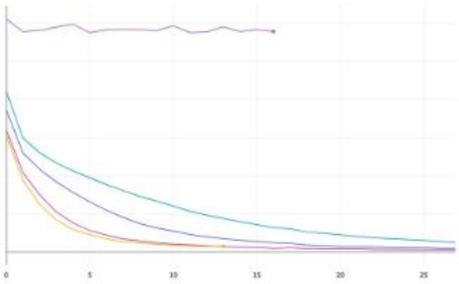
0: Entailment



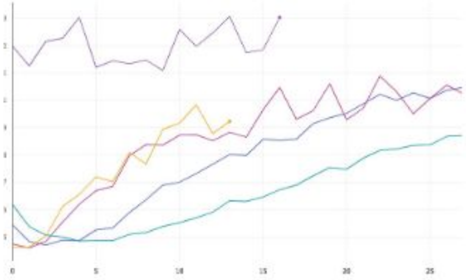
A man and a child are laughing at each other + Two people are laughing

BERT HP Tuning + Baseline Accuracy

Learning Rate: 5e-5



(a) Train Loss



(b) Validation Loss

5e-3 5e-6 1e-5 3e-5 5e-5

Baseline Accuracy:

Class Label	Test Loss	Acc
All classes	0.291	0.891
Class 0 examples only	0.240	0.913
Class 1 examples only	0.320	0.844
Class 2 examples only	0.139	0.919

How much of the logic did our generation models learn to produce?

Varying Temperature:

Temperature	VAE Test Loss	VAE BLEU-1	VAE BLEU-4	BERT Loss	BERT Acc
0	0.939	25.6	3.70	3.95	0.357
0.25	0.939	23.7	3.50	3.513	0.387
0.5	0.939	21.8	3.40	3.313	0.346
0.75	0.939	18.5	3.06	3.567	0.346
1	0.939	15.5	2.80	3.77	0.316

Table 5: BERT classification performance on sentences generated by conditional VAE using different temperatures and SNLI dataset.

Varying Class Label:

Class Label	VAE Test Loss	VAE BLEU-1	VAE BLEU-4	BERT Loss	BERT Acc
0	0.864	25.2	3.78	3.93	0.151
1	1.11	24.4	3.57	3.68	0.218
2	0.870	25.1	3.87	0.988	0.781

Table 6: BERT classification performance on sentences generated by class-specific VAE using temperature = 0.25 and SNLI dataset.

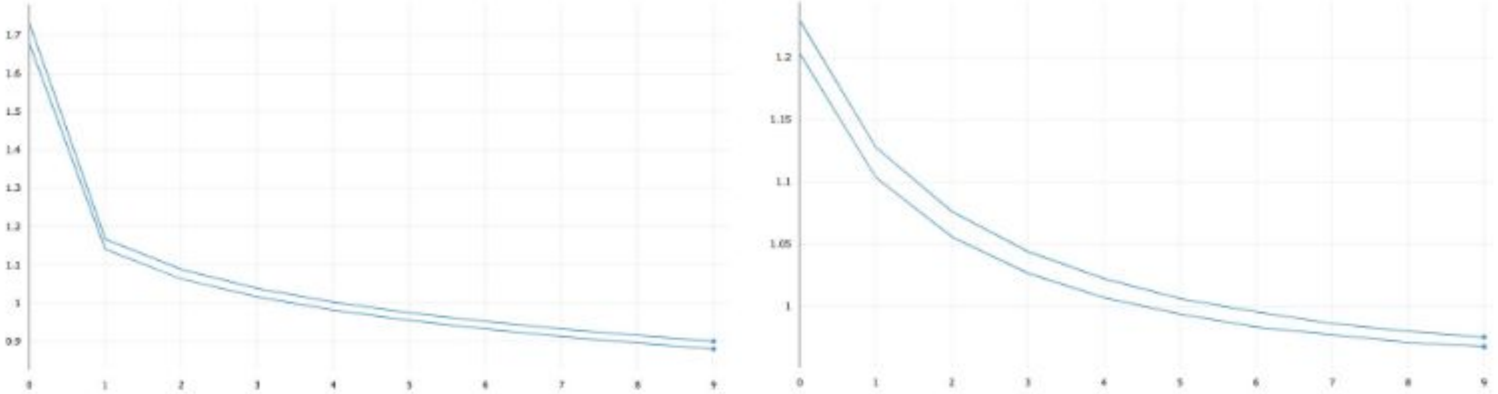
Architecture Changes?

- What we tried:
  - Use *encoder output* as *decoder hidden state* for initial time step only
  - Use *encoder output concatenated with hypothesis* as *decoder input*
  - Use *encoder output* as *decoder hidden and cell state* at every time step
  - Use *encoder output* as *decoder hidden state only* at every time step
  - Use *encoder output concatenated with decoder hidden state* as *decoder hidden state* at every time step

- Future work:
  - Can generate synonyms for various words in sentences
  - Use concatenated (premise, hypothesis) pair as encoder input
  - Use BERT instead of LSTM for encoder/decoder

Appendix

Architecture Changes



(a) Training Loss

(b) Validation Loss

Figure 5: Loss for baseline VAE model with 5e-4 learning rate, 512 hidden size, and 300 embedding size while varying architecture. Legend: Bottom Curve = Version 1, using hypothesis only as input into decoder; Top Curve = Version 2, concatenating embedded output of encoder with the hypothesis as input into decoder.

premise	a man in a black hat opens his mouth.
actual hypothesis (class 1)	The governor prepared to deliver the speech that would deliver the votes.
good neutral	a man is looking at a camera.
premise	a young infant cries while having his or her pajamas button.
actual hypothesis (class 2)	A young baby smiles.
bad contradiction	a man is standing outside.

Table 7: One "good" and one "bad" generated hypotheses from baseline (class-agnostic) VAE using temperature = 0.25.

premise	a man and a child are laughing at each other.
actual hypothesis (class 0)	Two people are laughing.
good entailment	a man is holding a child.
premise	a woman holds a newspaper that says "real change"
actual hypothesis (class 0)	a woman holding a newspaper that says "real change"
bad entailment	a man is wearing a shirt.

Table 8: One "good" and one "bad" generated hypotheses from class 0-specific VAE using temperature = 0.25.

premise	a man talking into a microphone with a woman standing next to him.
actual hypothesis (class 1)	The woman is sitting in the chair next to the podium.
good neutral	the man is a professional musician.
premise	a woman in black reviews a message as she walks to work.
actual hypothesis (class 1)	The woman in black is being fired via text message.
bad neutral	a man is trying to fix a broken component.

Table 9: One "good" and one "bad" generated hypotheses from class 1-specific VAE using temperature = 0.25.

premise	a man wearing a white shirt and a blue jeans reading a newspaper while standing
actual hypothesis (class 2)	A man is sitting down reading a newspaper.
good contradiction	the man is sitting on the couch.
premise	the small dog is running across the lawn.
actual hypothesis (class 2)	A cat is running up a tree.
bad contradiction	the man is wearing a red shirt.

Table 10: One "good" and one "bad" generated hypotheses from class 2-specific VAE using temperature = 0.25.

premise	a man on a bicycle rides past a park, with a group of people in the background.
actual hypothesis (class 2)	a guy rides his bike in the middle of a park.
good contradiction	a man is sitting on a bench.
premise	a small dog runs to catch a ball.
actual hypothesis (class 0)	A little dog chases a ball.
bad entailment	a woman is holding a child.

Table 11: One "good" and one "bad" generated hypotheses from conditional VAE using temperature = 0.25.