# FAD: Feature Alignment Discriminator for Abstractive Text Summarization

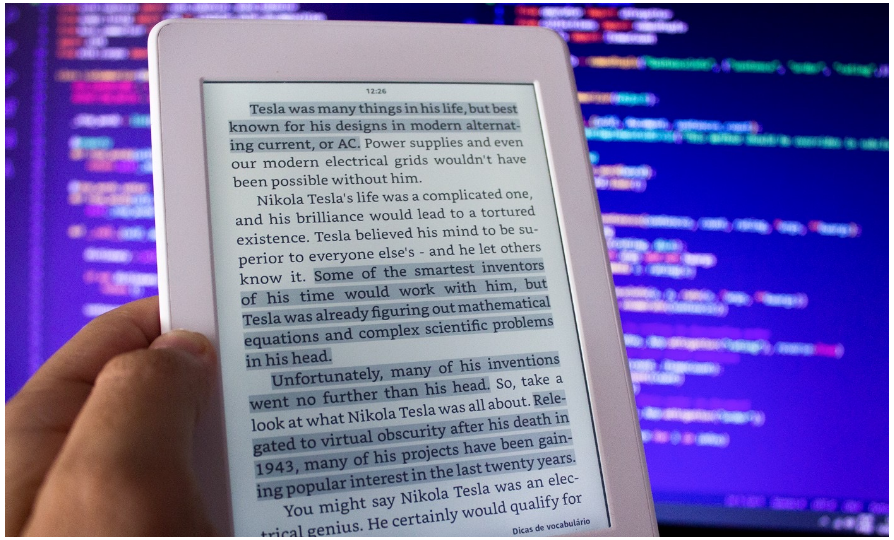
EECS 487 Group 5
04/11/2022

# Problem Description

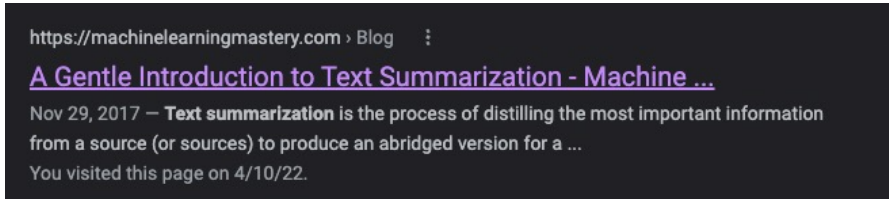
**Text summarization** is the process of distilling the most important information from a text to produce an abridged version for a particular task and user [Berry M.W 1995].

Significance:

- More and more text <-> Less and less time
- "Summaries as short as 17% of the full text length speed up decision making twice, with no significant degradation in accuracy." ["14-summarization"]




Salient words (extractive) in articles, source: Medium



https://machinelearningmastery.com › Blog

A Gentle Introduction to Text Summarization - Machine ...

Nov 29, 2017 — **Text summarization** is the process of distilling the most important information from a source (or sources) to produce an abridged version for a ...

You visited this page on 4/10/22.

"Google's summaries when googling 'summarization'"

# Problem Description

**Abstractive Summarization**:
express the ideas in the source documents possibly using different words [lecture-14 2022].

- more like the way humans process text [Liao, et al. 2020]
- better overall performance (controversiality).

Challenges:

- Repetition, quality of the reference summary, evaluation metric…
- **Coherence** and **Preciseness** (Our Focus)
  Better F1-score, precision score in ROUGE

**Input Text:** New York (CNN) When Liana Barrientos was 23 years old, she got married in Westchester County, New York. A year later, she got married again in Westchester County, but to a different man and without divorcing her first husband. Only 18 days after that marriage, she got hitched yet again
...

Abstractive Summarizer
e.g. fine-tuned Bart-large

**Generated Summary:** Liana Barrientos, 39, is charged with two counts of "offering a false instrument for filing in the first degree" In total, she has been married 10 times, with nine of her marriages occurring between 1999 and 2002. She is believed to still be married to four men.

**Referenced Summary:** Liana Barrientos, 39, re-arrested after court appearance for alleged fare beating . She has married 10 times as part of an immigration scam, prosecutors say . Barrientos pleaded not guilty Friday to misdemeanor charges

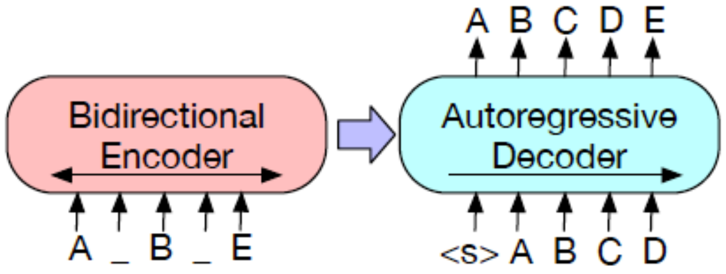
# Contributions

| Name | Contributions |
| --- | --- |
| Zixuan Pan | Model design/implementation/training, proposal draft, progress report Future Plan, presentation Methodology |
| Muzhe Wu | Model implementation review, inference, diagram, proposal Dataset/Evaluation review, progress report Methodology/Current Result, presentation Problem Description/Related Work |
| Jiarui Liu | Model implementation review, inference, training log visualization, proposal Related Work review, progress report Data Preprocessing/Current Result, presentation Experiment result |

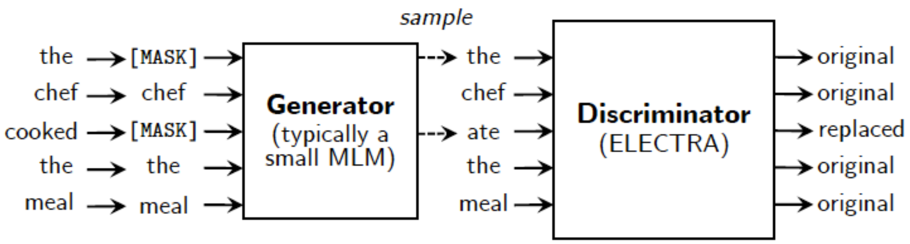
# Related Work

1. Abstractive Summarization with Pre-trained Models
    - Pre-trained language model: BERT, GPT, etc.
    - BART: text generation [Lewis, et al. 2019]
        - Problems: neglect token distribution in original text
        - Attempts:
            - HIBRIDS (bias term in Attention Calculation) [Cao and Wang 2022]
            - HIE-BART (multi-layer encoders) [Akiyama, 2021]
            - BART-Muppet (pre-finetuning) [Aghajanyan, 2021]

-> We chose BART as backbone generator




BART: BERT + GPT




Electra

2. Adversarial Training in NLP
    - Problems: discrete tokens, embedding
    - ELECTRA: replaced token detection [Clark, et al. 2020]

-> We applied Electra discriminator and designed the **aligned features** as input

# Dataset Statistics

CNN/DailyMail V3.0

- An English news dataset collected from CNN and DailyMail. Each sample contains an article and a reference human summary.
- A major dataset to evaluate summarization models.
- Contain both abstractive and extractive samples (mainly extractive).

| Set Name | Number of instances |
|---|---|
| Train | 287113 |
| Validation | 13368 |
| Test | 11490 |

| | Mean token count | Mean sentence count |
|---|---|---|
| News article | 781 | 29.74 |
| Summary | 56 | 3.72 |

# Data Preprocessing

1. Use GPT2 vocabulary to map each token into an index (Vocabulary size 51200).

1. Use a discrete and meaningful word embedding by GPT2.

**Raw news:** When singer Avril Lavigne went missing from the music scene, there was tons of speculation. Was she pregnant? In rehab? Going through a split from her husband, Nickelback front man Chad Kroeger?

…

Now the Canadian singer has revealed to People magazine that she was bedridden for five months after contracting Lyme disease. "I felt like I couldn't breathe, I couldn't talk, and I couldn't move," she told the magazine. "I thought I was dying."

…

"There were definitely times I couldn't shower for a full week because I could barely stand," she told People. "It felt like having all your life sucked out of you."

…

**Referenced Summary:** The singer had been off the scene for a while . She says she was bedridden for months . Lavigne was sometimes too weak to shower.

**Binarized Summary:** 24111 18695 509 7673 1008 468 1297 4881 284 8335 329 5885 286 1175 351 4068 764 8920 257 5975 12557 284 29737 36518 31060 338 5940 1230 764 4418 262 717 4141 8708 4139 284 3187 3908 1201 12122 764 30153 286 5786 1040 20845 8880 3183 11 635 220 4141 3710 3554 287 1737 15963 764

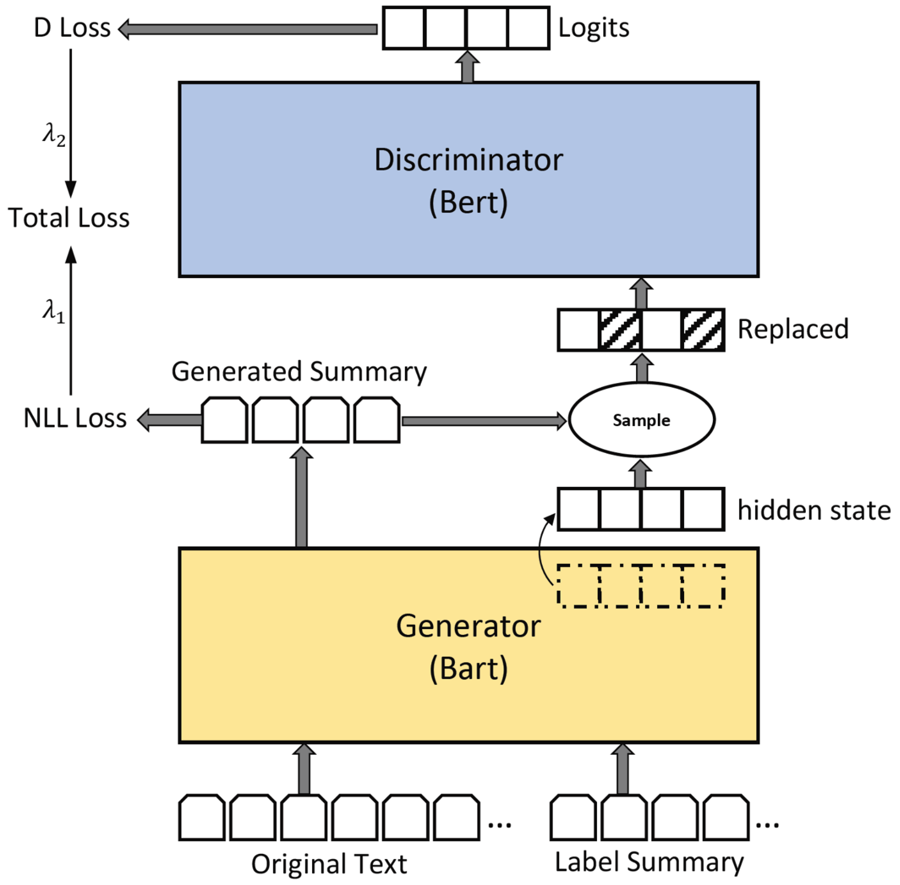
# Model Structure

Pretrained Bart Generator +
Feature Alignment Discriminator
(Pretrained Electra model)

1. Stack original text and label summary on batch dimension.

1. Pass batched input through Generator.

1. Calculate NLL Loss for generated summary.

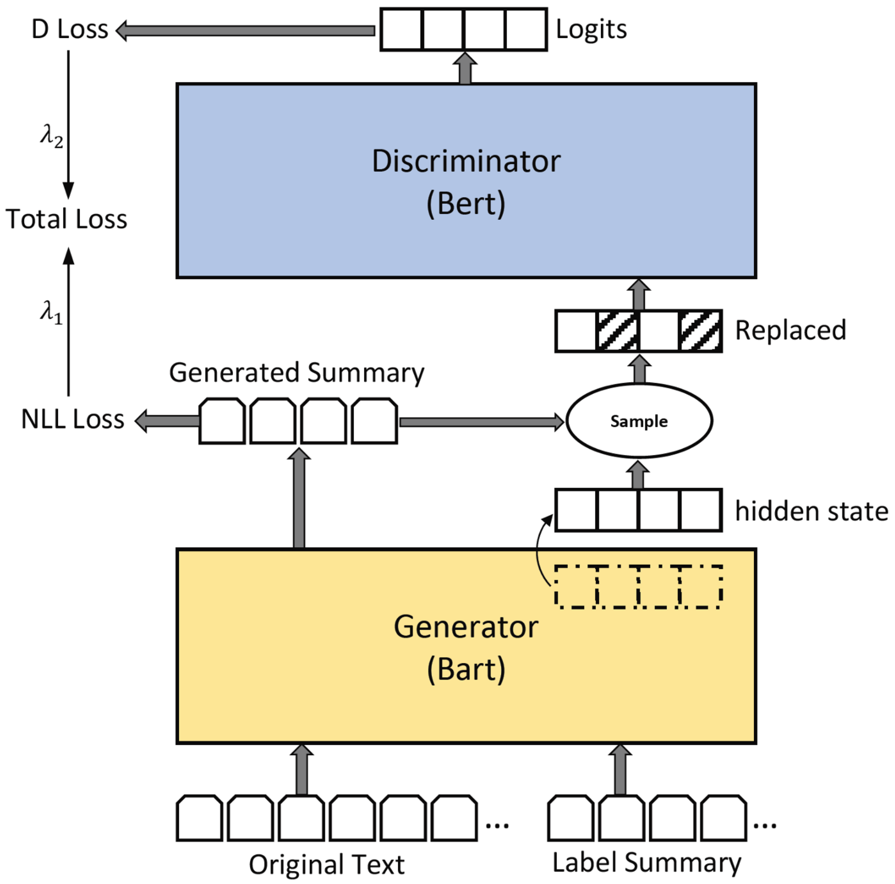


# Model Overview

4. Replace part of label summary feature with generated summary feature token-wisely. Label wrongly replaced tokens as fake, the others are real.

| Replacement naive example | | |
|---|---|---|
| I <u>cooked</u> <u>the</u> meal. | → | I <u>*ate*</u> <u>the</u> meal. |
| | | R F R R |

5. Pass replaced feature into Discriminator and calculate a Binary Cross Entropy Loss for each token.



# Model Highlights

- Use hidden layer features as Discriminator inputs. Thus break the discrete nature of sequence GAN using maximum likelihood method generator.
- A large replacing ratio (0.7 v.s. 0.15 in vanilla Electra)
- Utilize the position invariance of transformer and only regard wrongly replaced tokens from random sampling as fake.

**Algorithm 1** model.forward()

**Input:** $x$, $x_{\text{ref}}$

$\triangleright$ First stage

$hypo \leftarrow \text{BART}(x)$

$target \leftarrow x_{\text{ref}}$

$\mathcal{L}_{NLL} \leftarrow \text{nll\_loss}(hypo, target)$

$\triangleright$ Second stage

$h_{x,\text{ref}} \leftarrow \text{BART}(x_{\text{ref}})\text{.detach()}$

$\{replace\_ids\} \leftarrow \text{random\_sample}(x_{\text{ref}}\text{.index()}, p_{rep})$

$p(x) = \text{SoftMax}(hypo, \text{dim=-1})$

$\{candidate\_ids\} \leftarrow (\text{random\_sample}(hypo, p(x)) == x_{\text{ref}})\text{.index()}$

$\{replace\_ids\} = \{replace\_ids\} \backslash \{candidate\_ids\}$

$h_{x,\text{ref}}[\{replace\_ids\}] \leftarrow h_x[\{replace\_ids\}]$

$logits \leftarrow \text{Discriminator}(h_{x,\text{ref}})$

$labels \leftarrow \text{ones\_like}(logits)$

$labels[\{replace\_ids\}, :] \leftarrow 0$

$\mathcal{L}_D \leftarrow \text{BCEWithLogitsLoss}(logits, labels)$

$\mathcal{L}_{total} = \mathcal{L}_{NLL} + \lambda_2 \mathcal{L}_D$

# Experiment

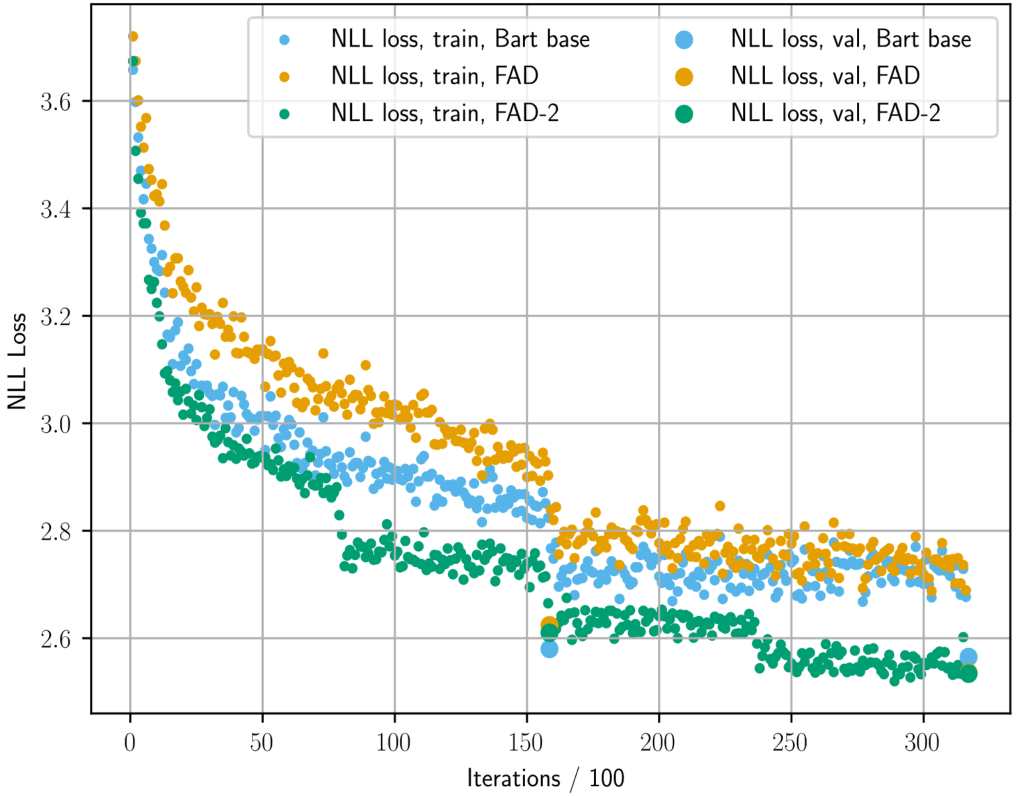
- Data splits
  - Follows CNN/DailyMail initial settings
- Hyperparameters
  - $P_{rep}$: 0.7 better than 0.4 or 0.15
  - Other parameters follow BART base model

| Hyperparameter Name | Symbol | Value |
|---|---|---|
| Replacement Ratio | $P_{rep}$ | 0.7 |
| Loss Scale | $(\lambda_1, \lambda_2)$ | (1, 50) |
| Learning Rate | $lr$ | $3 \times 10^{-4}$ |
| Regularization Strength | $\alpha$ | 0.7 |
| Adam Beta | $(\beta_1, \beta_2)$ | (0.9, 0.999) |
| Adam Weight Decay | *decay* | 0.01 |

# Training Process

- Train on GreatLakes Server with 2 A40s for 10 hours (3 epochs)
  - FAD uses the last hidden state of the decoder in the BART generator,

    while FAD-2 uses the first hidden state
  - An empirical finding that FAD-2 works better



# Evaluation

- Recall-Oriented Understudy for Gisting Evaluation (ROUGE)
  - A set of evaluation metrics for text summarization
  - Measures overlaps between generated and labeled summary
  - We use ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (LCS)

$$P = \frac{\#\ \text{n−grams in both generated and labeled summaries}}{\#\ \text{n−grams in generated summaries}}$$

$$R = \frac{\#\ \text{n−grams in both generated and labeled summaries}}{\#\ \text{n−grams in labeled summaries}}$$

- Perplexity
  - Measures the ability of modeling the objective function

# Results

- FAD-2 outperforms other models in all metrics
  - rDrop + FAD-2 the best in Precision, F1-score, and PPL; and FAD-2 itself the best in Recall

| Models | ROUGE-1 | | | ROUGE-2 | | | ROUGE-L | | | PPL |
|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | Precision | F1-score | Recall | Precision | F1-score | Recall | Precision | F1-score | |
| BART(baseline) | 0.497 | 0.386 | 0.423 | 0.229 | 0.179 | 0.196 | 0.459 | 0.357 | 0.391 | 5.84 |
| FAD | 0.494 | 0.392 | 0.427 | 0.229 | 0.182 | 0.198 | 0.457 | 0.364 | 0.395 | 5.77 |
| rDrop | 0.498 | 0.394 | 0.429 | 0.232 | 0.184 | 0.200 | 0.461 | 0.365 | 0.397 | 5.58 |
| FAD-2 | **0.500** | 0.391 | 0.428 | **0.233** | 0.183 | 0.199 | **0.463** | 0.362 | 0.397 | 5.65 |
| rDrop + FAD-2 | 0.497 | **0.395** | **0.430** | 0.232 | **0.185** | **0.200** | 0.459 | **0.366** | **0.397** | **5.58** |
| Electra base + FAD-2 | 0.498 | 0.392 | 0.428 | 0.231 | 0.182 | 0.199 | 0.461 | 0.363 | 0.396 | 5.64 |

# Example

- **Labeled summary:** Robin van Persie ruled out with ankle injury for Manchester United . Chris Smalling a doubt but Luke Shaw back from hamstring complaint . Ron Vlaar could make return to Aston Villa squad following calf injury . Joe Cole and Jores Okore have also regained fitness for Villans .
- **Bart base generated summary:** Robin van Persie ruled out of Manchester United's clash with Aston Villa . Luke Shaw has recovered from a hamstring problem, but Chris Smalling is a big doubt due to illness . Ron Vlaar could return for Aston Villa after shaking off a calf injury . Wayne Rooney has scored 12 goals against Aston Villa, his joint-highest tally against any opponent in Premier League history .
- **FAD-2 generated summary:** United host Aston Villa at Old Trafford (Saturday 3pm) Robin van Persie ruled out with ankle injury . Luke Shaw has recovered from a hamstring problem, but Chris Smalling is a big doubt due to illness . Ron Vlaar could return for Aston Villa after shaking off a calf injury . Kieran Richardson and Philippe Senderos could still miss out .

The last sentence is redundant

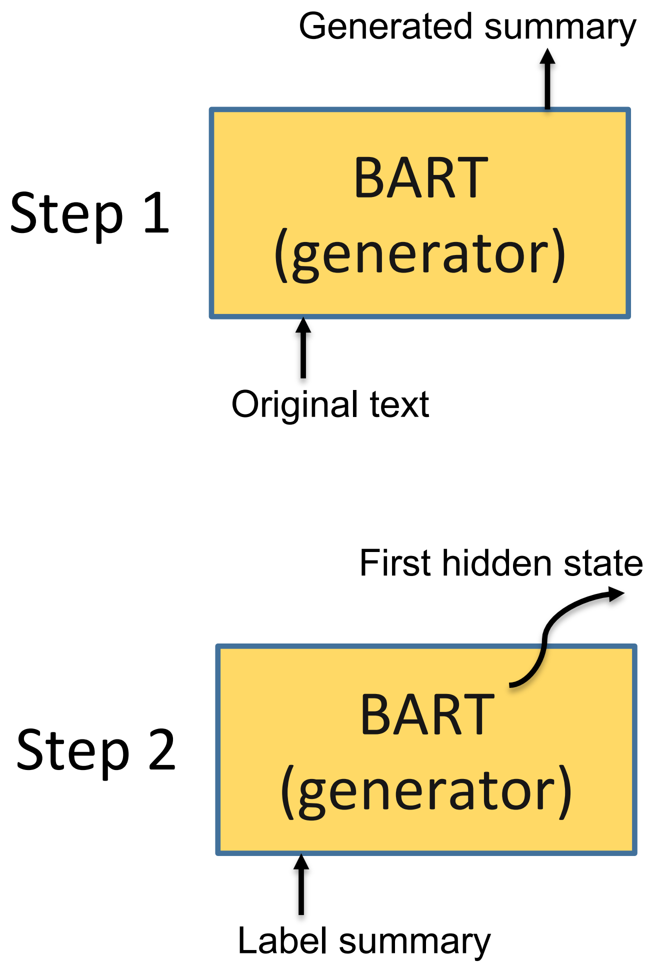
The last sentence contains useful information

# Discussion

- In general, applying a discriminator to BART generator improves ROUGE precision score for text summarization task
- Ablation study
  - FAD-2 (uses first hidden state) performs better than FAD (uses last hidden state), which tends to be empirical
- Electra small vs. Electra base
  - PPL and ROUGE are approximately the same
  - Conforms to the expectation that a larger model shouldn't performs worse than a smaller model
- Can be generalized to other sequence-to-sequence tasks than text summarization

# Discussion

- Strategy for **stopping gradient** in the training process
  - Go backward in Step 1 and detach in Step 2:
    - Our choice
    - Both generator and discriminator learn properly
  - Detach in Step 1 and Step 2:
    - Discriminator loss goes down, indicating that discriminator learnt in the expected manner
    - However, the weights in the generator will not be updated
  - Go backward in Step 1 and Step 2:
    - Discriminator loss decreases too fast
    - Generator kind of learnt to distinguish generated and labeled summary, which ought to be done by discriminator



# Reference


Berry M.W., Dumais S.T., and O’Brien G.W. Using linear algebra for intelligent information retrieval. SIAM Rev., 37(4): 573–595, 1995.

Liao, Pengcheng, et al. "Improving abstractive text summarization with history aggregation." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.

Wang, Lu. “14-summarization.” umich.instructure.com, 2022.

Suleiman, Dima, and Arafat Awajan. "Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges." *Mathematical problems in engineering* 2020 (2020).

Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

Cao, Shuyang, and Lu Wang. "HIBRIDS: Attention with Hierarchical Biases for Structure-aware Long Document Summarization." *arXiv preprint arXiv:2203.10741* (2022).

Akiyama, Kazuki, Akihiro Tamura, and Takashi Ninomiya. "Hie-BART: Document Summarization with Hierarchical BART." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 2021.

Aghajanyan, Armen, et al. "Muppet: Massive multi-task representations with pre-finetuning." *arXiv preprint arXiv:2101.11038* (2021).

Clark, Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555* (2020).

# Q&A

# Thank You