IIIT - Hyderabad

BTP-1

Numerical Question Generation In Science Using Diffusion Models

Ruhul Ameen S (2020102031)
Harish Umasankar (2020102067)

Supervised by
Asst. Prof. Dr. Pawan Kumar

Table of Contents

0.1 Objectives

The main objective of this project is to develop a model that can automatically generate numerical questions in science based on a given theory. To achieve this goal, different approaches will be explored to model the relationships between the context and questions in order to improve the diversity of the generated questions. The quality and diversity of the questions generated will be assessed through various metrics and evaluation techniques

Ultimately, the success of this project will be determined by the ability of the model to accurately and effectively generate a wide range of numerical questions in science

0.2 Text Embedding: BERT

Diffusion models tend to perform better on continuous data like images. In images the pixel value is in the range [0, 255]. Diffusion models use stochastic differential equations hence they perform better on continuous functions.

Text data is discrete in nature, because it consists of a finite set of distinct symbols or tokens, such as letters, punctuation marks, and whitespace, which cannot take on continuous values.

Text Tokenization and Embedding helps in converting the discrete text data into a continuous representation that can be processed by the model. BERT Embedding is created with the following process

1. **Tokenization** BERT uses a process called WordPiece tokenization, which breaks words into smaller subwords based on their frequency in a large corpus of text. For example, the word "playing" might be broken down into the subwords "play" and "ing".

2. **Adding Special Tokens** To enable BERT to process text sequences, special tokens are added to the beginning and end of the sequence. These tokens are:

CLS : This token is added to the beginning of the text sequence and represents the start of the sequence. It is used to train models to perform sentence-level classification tasks, such as sentiment analysis.

SEP : This token is added to the end of each sentence in the sequence and separates the sentences. It is used to train models to perform tasks such as question answering.

PAD : This token is used to pad shorter sentences to ensure that all sequences have the same length. It is added to the end of the sentence and is used to train models to perform tasks such as text classification.

3. **Encoding**

- The encoding process involves passing the tokens through a multi-layer bidirectional transformer architecture, which captures the contextual information of each token based on the surrounding tokens in the sequence
- The resulting contextual embeddings are dense, low-dimensional vectors that capture the meaning and context of each token in the text sequence.

0.3 Model

0.3.1 Text Embedding

An embedding function (as described above) EMB (w) will be used to a continuous space. For a given pair of sequences

- w^x - Input Sequence, here context is given as input
- w^y - Output Sequence, here question generated is the output

The pair of sequences w^x, w^y are concatenated = $w^{x \oplus y} = w^z$. The concatenated sequence is then embedded as

EMB($w^{x \oplus y}$) = [EMB(w_1^x), EMB(w_2^x), ..., EMB(w_m^x), EMB(w_1^y), EMB(w_2^y) ... EMB(w_n^x)], where m is length of sequence w^x and n is length of sequence of w^y

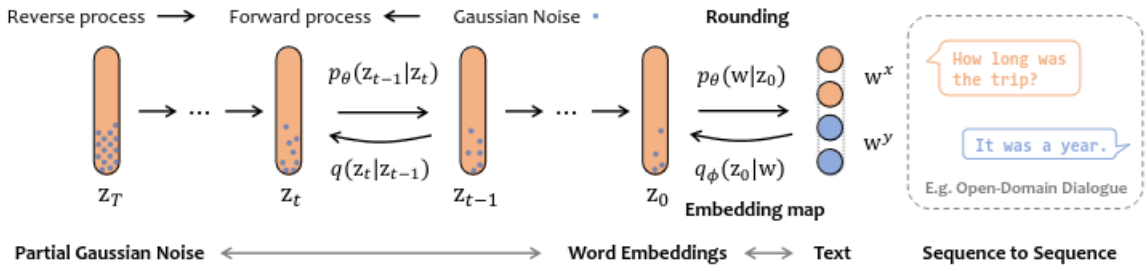
0.3.2 Forward Diffusion Process and Partial Noising

In diffusion models, the goal is to model the underlying probability distribution of the data by iteratively adding noise to the input data. In each forward step, the model predicts the distribution of the next step conditioned on the previous step, i.e., $q(z_t|z_{t-1})$. Here $z_t = x_t \oplus y_t$, where x_t is the input sequence and y_t is the output sequence.

There are two common ways to inject noise into the previous step z_{t-1} in order to obtain z_t:

- **Conventional Models**: In this approach, the entire previous step z_{t-1} is corrupted with noise to obtain z_t. This can be done by adding Gaussian noise or randomly masking out some elements of the previous step.
- **Partial Noising**: In this approach, only a subset of the previous step, denoted as y_t, is corrupted with noise to obtain z_t. This can be done by masking out a random subset of elements in z_{t-1} and replacing them with noise. The remaining elements are left unchanged.

Here, the partial noising method has been adopted. The partial noising approach has been shown to be effective in diffusion models because it allows the model to focus on the most relevant information in the previous step, while preserving some of the original signal. This can lead to more stable and accurate predictions, particularly in cases where the noise added to the input data is high.



0.3.3 Reverse Diffusion Process with Conditional Denoising

Reverse Diffusion

- The goal of a diffusion model is to recover the original input data, denoted as z_0, by iteratively denoising the input data z_t.
- The model learns the denoising process by estimating the conditional distribution $p_\theta(z_{t-1}|z_t)$, which represents the probability distribution of the previous step z_{t-1} given the current step z_t. This conditional distribution is parameterized by a neural network with learnable parameters θ.

Conditional Diffusion Models

- During training, the model is optimized to maximize the likelihood of the observed data given the parameters, by minimizing the negative log-likelihood loss.
- In conditional diffusion models, we have an additional input y (for example, a class label or a text sequence) and we try to model the conditional distribution $p(x \mid y)$ instead. In practice, this means learning to predict the conditional score function $\nabla_x \log p(x \mid y)$.
- Modes of Guidance
  - Classifier Guided
    - Applying bayes rule and taking gradient with respect to x on $p(x|y)$, we can get the expression in terms of unconditional and classifier scoring functions
    - $\nabla_x \log p_\gamma(x \mid y) = \nabla_x \log p(x) + \gamma \nabla_x \log p(y \mid x)$
    - Here γ is the guidance scale
    - This method involves training a strong unconditional model $p(x)$
    - The model is the conditioned by a trained classifier $p(y|x)$
  - Classifier Free
    - Using Bayes rule on $p(y|x)$ and taking gradient we get
    - $\nabla_x \log p(y \mid x) = \nabla_x \log p(x \mid y) - \nabla_x \log p(x)$.
    - Using this in classifier guided equation, we get
    - $\nabla_x \log p_\gamma(x \mid y) = (1-\gamma)\nabla_x \log p(x) + \gamma \nabla_x \log p(x \mid y)$
    - From the equation we can observe that score is independent of any classifier model.

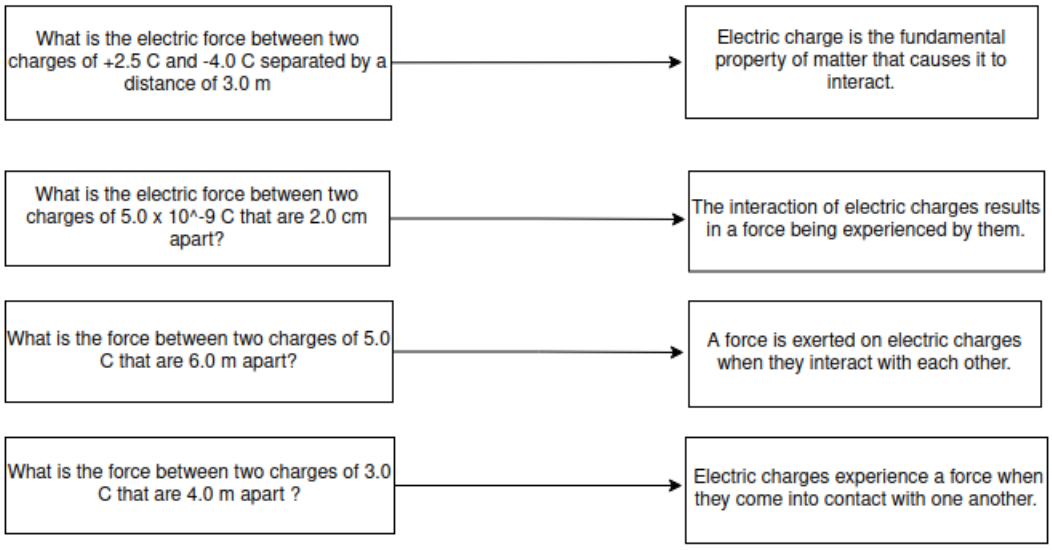
  In our model, classifier free approach has been adopted. Classifier-free diffusion models do offer the advantage of being able to be trained end-to-end in a single go and we have to train only one generative model.

0.4 Dataset

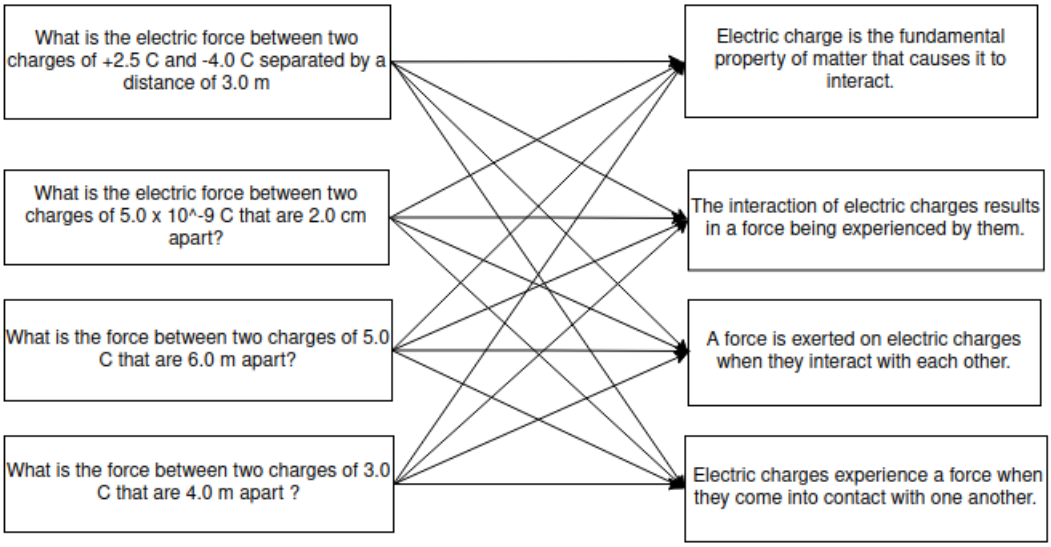
We have used the OpenQA dataset to train the diffusion model on basic English question-context pairs to improve the vocabulary understanding of the model. This dataset contains 117k question-context pairs in the English language.

Then we train the model with a dataset created by us to generate numerical questions given a theoretical context in several topics of science like acceleration, optics, etc. We have created this dataset in two different ways to test and compare the model results.

- First type of dataset : Forward mapping
  - 10 unique contexts for each science topic.
  - 20 different numerical questions for each context.
  - Totally, 200 question-context pairs for each science topic.
  - Validation and test dataset contains pairs sampled from the training dataset but it contains paraphrased versions of contexts and questions.



- Second Type of dataset : Forward + reverse mapping
  - It has 40 contexts for each science topic in sets of 4 contexts being the paraphrased versions of each other.
  - 20 different numerical questions for each of these sets of 4 contexts which are paired up against each context in the set.
  - Totally, 80 question-context pairs for each context set. As we have 10 different context sets for each science topic, There will be 800 question-context pairs for each science topic.
  - Validation and test dataset contains pairs sampled from the training dataset but it contains paraphrased versions of contexts and questions.



The reason behind having a second type of dataset with forward + reverse mapping between questions and contexts is to train the model to understand related theoretical contexts and make them generate questions in the correct topic of science when given a completely new context.

0.5 Loss Function

Diffusion models are trained to maximize the log-likelihood of the data given the noise-corrupted samples. However, computing the exact log-likelihood involves integrating over all possible paths of the diffusion process, which is computationally expensive and not tractable in practice.

Therefore, a Variational Lower Bound (VLB) on the log-likelihood is used instead, which can be optimized using standard gradient-based techniques. The Variational Lower Bound involves approximating the true posterior distribution over the latent space with a simpler distribution that is tractable, such as a Gaussian distribution.

The Variational Lower Bound is then used as a surrogate objective function for training the diffusion model.

$$L_{VLB} = E_{q(z_{1:T}|z_0)}\left[\log\left(\frac{q(z_t|z_0)}{p_\theta(z_T)}\right) + \sum_{t=2}^{T}\log\left(\frac{q(z_{t-1}|z_0,z_t)}{p_\theta(z_{t-1}|z_t)}\right) + \log\left(\frac{q(z_0|w^{x\oplus y})}{p_\theta(z_0|z_1)}\right) - \log p_\theta(w^{x\oplus y}|z_0)\right]$$

Inside the expectation expression,

- **First Term**: Since q has no learnable parameters and p is just a Gaussian noise probability, this term will be a constant during training and thus can be ignored.
- **Second Term**: Refers to the step wise denoising process. This term compares the target denoising step q and the approximated denoising step p.

- **Third Term**: This is the reconstruction loss of the last denoising step and it can be ignored during training because it can be approximated using the same neural network in L_{t-1}
- **Fourth Term**: Refers to the rounding error. Rounding error in diffusion models refers to the accumulation of small errors over multiple rounds of the diffusion process.

0.6 Experiments

We have performed three different type of experiments to test and improve the results of the model.

- Training with OpenQA + forward map dataset : training with OpenQA dataset makes the model understand normal english language in the basis of conditional modelling and forward map dataset finetunes the model to our specific interest of generating numerical questions in the topics of science.
- Training with OpenQA and forward + reverse map dataset : Similar to first experiment, OpenQA makes the model normal english language in the basis of conditional modelling and forward + reverse map dataset provides more diversity and adaptibility of the model to contexts which are slightly different from the training contexts which enables it to generate questions in the correct domain.

0.7 Results

The result - numerical science question generated by the model can be discussed in terms of the following characteristics

0.7.1 Diversity

Experiment 1: Forward Mapped Dataset

The model could not capture the underlying variability and relationships between concepts and generated same or similar questions for a given context.

Topic: Electricity

Context	the relationship between current and voltage in a conductor is described by ohm's law, which states that the current is proportional to the voltage
Question1:	what current for voltage 2. 5 V through a wire with a resistance of 7 ohms?
Question2:	what current for voltage 2 V through a wire with a resistance of 4 ohms?

Experiment 2: Forward+Backward Mapped Dataset

With forward and reverse mapping of the dataset used in the experiment 2, it accounts for some variability in the questions generated.

We can observe this from above that we have 2 different type of questions generated for the same context when queried again.

Topic: Electricity

Context	the relationship between current and voltage in a conductor is described by ohm's law, which states that the current is proportional to the voltage
Question1:	what current for voltage 2. 5 V through a wire with a resistance of 7 ohms?
Question2:	required voltage to produce a current of 12. 5 amps through a wire with a resistance of 7 ohms?

0.7.2 Quality

A sample of questions generated for different topics are presented below

Topic: Acceleration

Context	acceleration is rate of change of velocity with respect to time
Question:	what is acceleration is train to from 10 m/s to 20/s in 20 seconds?
Context	acceleration due to gravity leads to freefall motion
Question:	if a ball is at height 10 km, what is time to freefall to gravity in seconds

Topic: Electricity

Context	the relationship between current and voltage in a conductor is described by ohm's law, which states that the current is proportional to the voltage
Question:	what current for voltage 2. 5 V through a wire with a resistance of 7 ohms?
Context	the sum of all resistances present in a series circuit is called series resistance
Question:	resistance find in circuit with three resistors resistances with 2 ohms, 4 ohms, 3 ohms, connected in series

Topic: Alternating Current

Context	there are capacitors and resistors in ac circuits.
Question:	what is the imp impedance of a circuit composed of a 10 ŏ3c9 resistor? [SEP] with a 20 ŏ3bcf capacitor ?
Context	voltage drops and current stays constant in series connection.
Question:	what is the voltage drop across between in a resistor composed of a 50 ŏ3c9 resistor and a 20 ŏ3bcf capacitor in series?

Topic: Acids and Bases

Context	in water, acids can donate h + ions.
Question:	what is the ph of a solution that has a [h +] concentration of $5x10^{-}8m$?

Topic: Light

Context	the image formed by a plane mirror is located at the same distance from the mirror as the object.
Question:	if an object is placed 5 cm from a plane mirror, how far is the image from the mirror?

Topic: Electric Charges

Context	electric charges are surrounded by electric field.
Question:	what is the electric field strength at a distance of 0. 2 m from a point charge of 3. 0 c?
Context	the electric potential energy can be expressed as a relation of electric charges and distances.
Question:	what is the potential energy of two charges of 4. 0 c that are 4. 0 m apart?

Topic: Thermal Matter

Context	the transfer of thermal energy within a fluid through the movement of the fluid itself is referred to as convection.
Question:	what is the convective heat transfer coefficient for a fluid flowing through a 2 cm diameter tube, if the rate of heat transfer is 100 w, the temperature difference between the fluid and the tube is 5$\breve{0}$0b0c, and the fluid velocity is 1 m / s?
Context	Product of pressure and volume is constant under constant temperature.
Question:	gas has a volume of 3 l at a pressure of 2 atm. if the pressure is increased to 3 atm while the temperature remains constant, what will be its new volume?

Topic: Oscillations

Context	sinusoidal function can be used to model a type of oscillatory motion known as simple harmonic motion.
Question:	a simple pendulum has a frequency of 2 hz. what is its period?
Context	a damped harmonic oscillator experiences an exponential reduction in amplitude over time.
Question:	a dampeded harmonic oscillator has a quality factor of and a natural frequency of 6 hz. what is the damping coefficient of the oscillator?

Topic: Electromagnetic Induction

Context	the amount of magnetic field lines that go through a particular area is known as magnetic flux.
Question:	what magnetic flux has through square loop loop radius 5 side of a cm and a magnetic field 6 T?

Discussion

The questions were mathematically correct and coherent with input context.
However, there were still some grammar errors and syntax issues that needed to be addressed. These errors included misspellings, incorrect word usage, and awkward sentence structures. While these errors did not affect the mathematical correctness or logical coherence of the questions, they could be distracting

0.8 Limitations

Upon analyzing the performance of our model, we have identified several areas where it is struggling to generate accurate or relevant questions. Specifically, we have found that the model is having difficulty with certain complex concepts

- Topic: Acids and Bases
  - Context: pH is defined as the negative log of concentraion of H+ ions
  - Question Generated: +] concentration of a solution that has a ph of 8. 2.

  - Possible Explanation:
    * Most of the questions generated for contexts from Acids and Bases Topics have poor usage of symbols like "+" and "[]". This can be attributed to the fact that there are relatively more number of symbols used in Acids-Bases topic like "+" for positive ion, "[]" for concentration and for exponent symbol.
    * We can hypothesize that although model performs well for generating number based questions, the model struggles when it has to generate questions with several symbols involved

- Topic: Light
  - Context: the linear magnification is a term used to describe the relationship between the height of an object and the height of its image when viewed through a mirror.
  - Question Generated: is 100 8 cm?
  - Possible Explanation:
    * There are uncorrelated subtopics under Light topic - focal length, magnification, refraction, deflection, image distance etc
    * We can hypothesize that, when there are multiple uncorrelated subtopics under a subtopic, the model will find it difficult to learn the underlying patterns.

0.9 Metrics for Evaluation

The performance of the model is analyzed with respect to

- **Quality : BLEU Score**
  The BLEU score is a measure of the similarity between the machine-generated text and the reference question generated. The BLEU score is computed by comparing the n-grams (contiguous sequences of n words) in the machine-generated text to those in the reference text, and measuring the proportion of overlapping n-grams. A higher BLEU score indicates greater similarity between the generated and reference texts, and thus a higher quality of generated questions.

- **Diversity: Self BLEU Score**
  It computes the BLEU score between an n-gram of the generated text and all other n-grams in the text except for itself. This can help to identify repeated or redundant phrases in the generated text, which can be an indication of a lack of diversity or creativity. A lower Self BLEU score indicates greater diversity in the generated questions, and thus a more diverse set of questions.

We have used BLEU Score as the metric for judging the questions generated.

0.10 Score

Model	BLEU	Self BLEU
DiffuSeq - Question Generation	0.1731	0.2732
DiffuSeq - Numerical Science Question Generation	0.1775	0.029

0.11 Conclusion

In conclusion, this paper presented a novel approach to numerical question generation in science using diffusion models. While the proposed method showed some promising results in generating questions that are syntactically and semantically correct, as well as challenging for students to solve, the overall performance was mixed.

The evaluation of the generated questions using a variety of metrics showed that the proposed method performed well in terms of generating questions that were challenging for students to solve. However, there were some issues with the quality and relevance of the questions, as well as the overall coherence of the generated question sets.

Despite these limitations, the use of diffusion models for numerical question generation in science is still a promising direction for future research. Further improvements in the methodology and evaluation of the generated questions could help to address some of the issues observed in this study. Additionally, the proposed approach could be adapted and applied to other domains, potentially leading to improvements in automatic question generation for a wide range of educational settings.

0.12 References

- Gong, S., Li, M., Feng, J., Wu, Z. and Kong, L., 2022. Diffuseq: Sequence to sequence text generation with diffusion models. arXiv preprint arXiv:2210.08933.
- Li, X., Thickstun, J., Gulrajani, I., Liang, P.S. and Hashimoto, T.B., 2022. Diffusion-lm improves controllable text generation. Advances in Neural Information Processing Systems, 35, pp.4328-4343.
- Kollepara, N., Chatakonda, S.K. and Kumar, P., 2021. SCIMAT: Science and Mathematics Dataset. arXiv preprint arXiv:2109.15005.

0.13 Code and Github Repository Link

You can find the link to the dataset and code used below `https://github.com/misterpawan/BTP2023_Diffusion_Augmentation_Anomaly/tree/main`