# AI for Semi-Automatic Grading of Online Formative Assessments

Pakhale Advay Dilip, Xavier Lien Tong Wei

## Introduction

**Formative Assessments (FAs)**
- Provides feedback to both students and teachers
- In contrast, summative assessments (SAs) focus only on score
- FAs are highly beneficial for student learning compared to SAs
- But tedious and time-consuming to mark

→ **Possible Solution: Semi-automatic grading**

**Current Literature on Automatic Grading**
- Focuses mainly on grading summative assessments
- Primary focus is on generating score, not feedback

∴ **Need for new architecture to fulfil aims of grading FAs**

### Proposed Solution: Semi-Automatic Grading Architecture

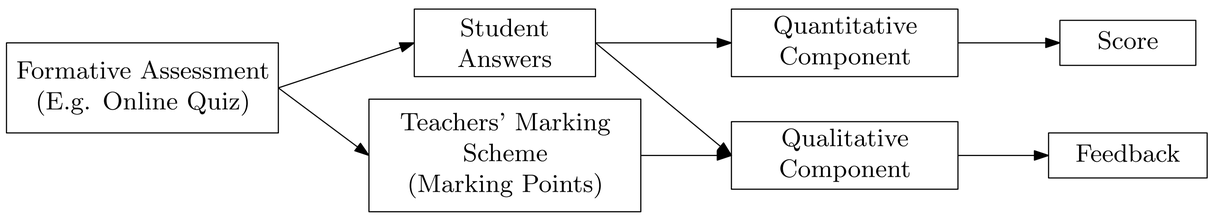


*Figure 1: Outline of Architecture*

**With Our Architecture…**
- FAs can be administered more frequently
- Both teachers and students reap benefit of timely and frequent feedback generated

## Methodology

### Datasets and Pre-processing

Thermodynamics online quiz conducted on 292 Secondary 2 RI students

Dataset 1 **Simple Recall Question** | Dataset 2 **Complex Application Question**

Marked by 3 physics teachers using a simple **marking scheme** comprising multiple **marking points**

Answers pre-processed to make **lowercase**, remove **punctuation** and **stopwords**

**Why Thermodynamics?**
Important topic in mainstream schools

Tested using qualitative short-answer questions

Students have many misconceptions regarding the topic

### Quantitative Component

1. **Word Embeddings.** Comparison between GloVe and fastText
2. **Baseline Feedforward Neural Network**. Sum of word vectors in answer as input. Does not preserve sequential information**.**
3. **LSTM and GRU Networks.** Recurrent Neural Network (RNN) variants, state of the art algorithms for sequence processing tasks.
4. **Bidirectional RNNs (BiRNNs).** Processes sequence to retain contextual information in both directions to improve performance.
5. **Attention Mechanism.** Places more weight on more important words. Especially important in Physics contexts as teachers mark with reference to more important key words

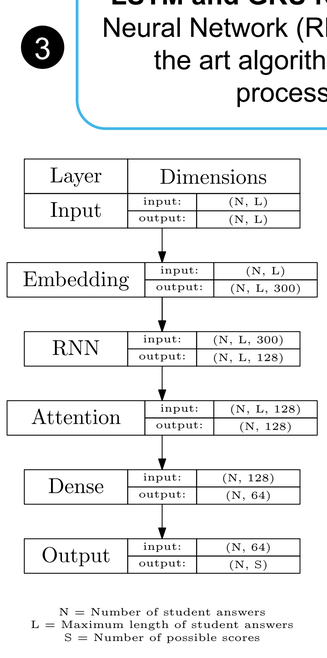


*Figure 2: Non-Baseline Neural Network Architecture*

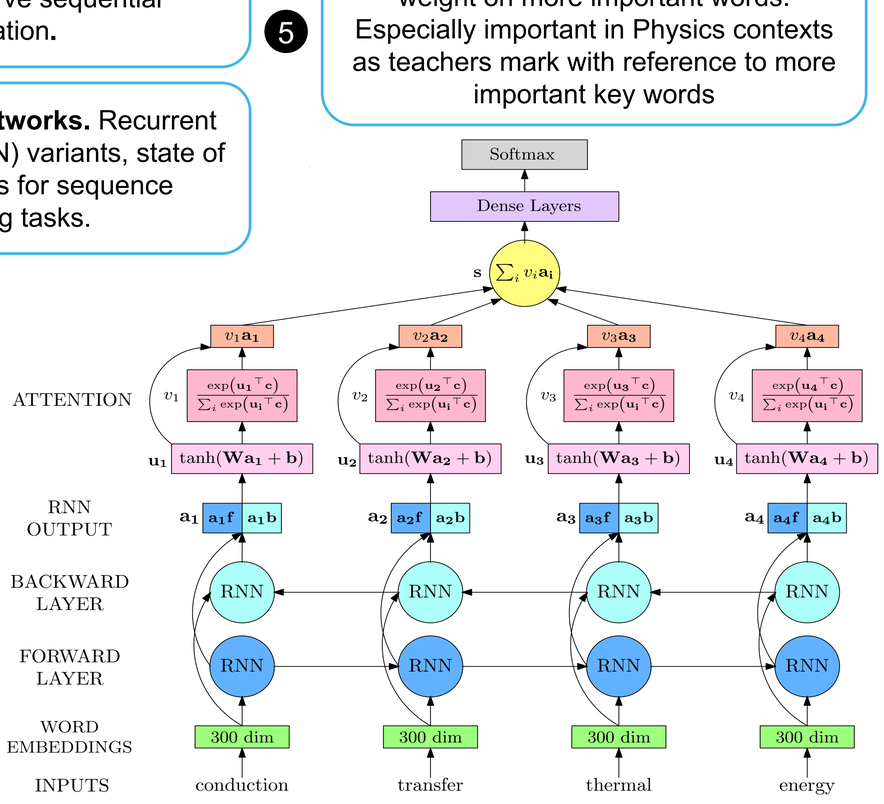


*Figure 3: Attention Mechanism on Top of a BiRNN*

### Qualitative Component

For each answer $j$, qualitative models produce a vector of marking point proportions, $\mathbf{p_j} \in \mathbb{R}^n$, where $n$ is the number of marking points. $p_{jk}$ is the proportion of marking point $k$ in answer $j$, where $p_{jk} \in [0, 1]$.

$\mathbf{t_k} = \sum_{i \in T_k} \mathbf{w_i}$, where $\mathbf{t_k}$ and $T_k$ are the marking point vector and the set of words in the $k$ th marking point respectively, and $\mathbf{w_i}$ is the word vector for the $i$ th word.

$\mathbf{a_j} = \sum_{i \in A_j} \mathbf{w_i}$, where $\mathbf{a_j}$ and $A_j$ are the answer vector and the set of words in the $j$ th answer respectively.

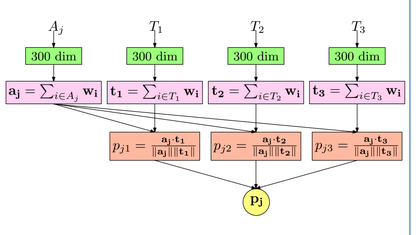


*Figure 4: Cosine Similarity Model for Question with 3 Marking Points*

## Methodology (cont.)

### Qualitative Component (cont.)

1. **Vector Decomposition.** $\hat{\mathbf{p_j}} = \arg\min_{\mathbf{p_j}} \|\hat{\mathbf{a_j}} - \mathbf{a_j}\|^2$, where $\hat{\mathbf{a_j}} = \sum_{k=1}^{n} p_{jk} \mathbf{t_k}$
2. **Cosine Similarity.** $p_{jk} = \cos\theta_{jk} = \frac{\mathbf{a_j} \cdot \mathbf{t_k}}{\|\mathbf{a_j}\|\|\mathbf{t_k}\|}$, and $\mathbf{p_j} = (\, p_{j0}\ p_{j1}\ \cdots\ p_{jk}\,)$
3. **Euclidean Distance.** $p_{jk} = \|\mathbf{a_j} - \mathbf{t_k}\|$, and $\mathbf{p_j} = (\, p_{j0}\ p_{j1}\ \cdots\ p_{jk}\,)$

## Data and Discussion

### Evaluation Methodology

5-Folds cross-validation

To test performance in different data environments, size of training set varied from 25-250 samples, 42 in validation set

Simple feedforward NN to test accuracy of $\mathbf{p_j}$ vectors generated by qualitative component

### Quantitative Component

Performance of Quantitative Models on Dataset 1

| | Models | Accuracy | Loss | F1 Score |
|---|---|---|---|---|
| | Feedforward Neural Network (Baseline) | 0.733 | 0.849 | 0.719 |
| LSTM | With GloVe | 0.702 | 0.780 | 0.660 |
| LSTM | BiLSTM with GloVe | 0.705 | 0.837 | 0.699 |
| LSTM | BiLSTM with GloVe and Attention | 0.760 | 0.698 | 0.750 |
| LSTM | BiLSTM with fastText and Attention | 0.716 | 0.741 | 0.696 |
| GRU | With GloVe | 0.715 | 1.050 | 0.701 |
| GRU | BiGRU with GloVe | 0.709 | 0.909 | 0.692 |
| GRU | **BiGRU with GloVe and Attention** | **0.781** | **0.701** | **0.775** |
| GRU | BiGRU with fastText and attention | 0.726 | 0.715 | 0.692 |

Performance of Quantitative Models on Dataset 2

| | Models | Accuracy | Loss | F1 Score |
|---|---|---|---|---|
| | Feedforward Neural Network (Baseline) | 0.472 | 1.354 | 0.433 |
| LSTM | With GloVe | 0.428 | 1.840 | 0.399 |
| LSTM | BiLSTM with GloVe | 0.432 | 1.546 | 0.394 |
| LSTM | BiLSTM with GloVe and Attention | 0.520 | 1.530 | 0.500 |
| LSTM | **BiLSTM with fastText and Attention** | **0.541** | **1.157** | **0.506** |
| GRU | With GloVe | 0.414 | 1.658 | 0.373 |
| GRU | BiGRU with GloVe | 0.421 | 1.459 | 0.392 |
| GRU | BiGRU with GloVe and Attention | 0.462 | 1.435 | 0.447 |
| GRU | BiGRU with fastText and attention | 0.496 | 1.265 | 0.465 |

**Findings**
- BiGRU with GloVe works best for simpler questions, BiLSTM with fastText for more complex questions
- Models perform better on dataset 1 → likely due to difference in question complexity
- Simple baseline models do surprisingly well → could be due to overfitting and poor hyperparameter tuning in the more complex models

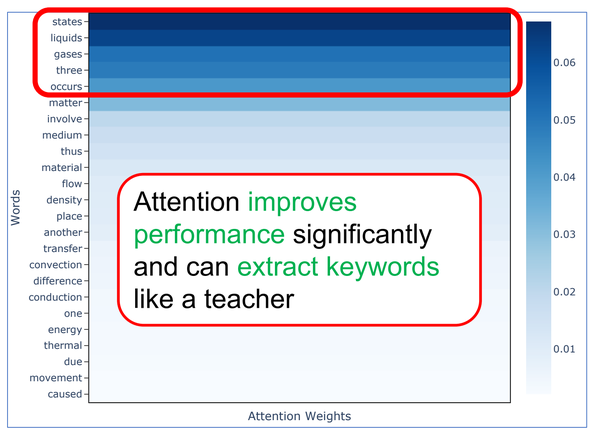


*Figure 5: Attention Weights of Answer Sample from Dataset 1*

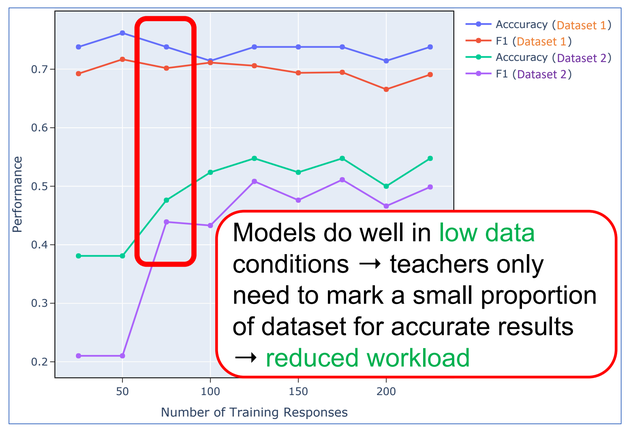


*Figure 6: Performance of Best Models Against Number of Training Responses*

### Qualitative Component

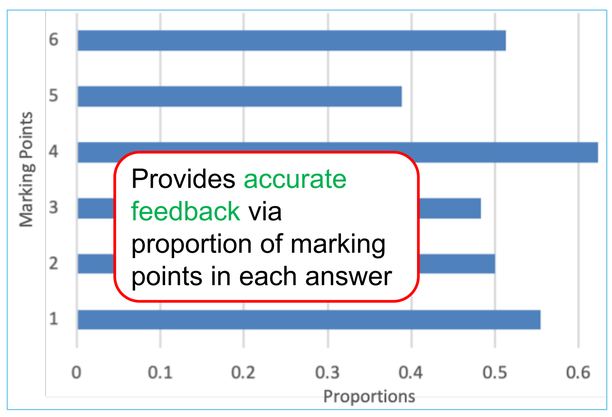


*Figure 7: Averaged Marking Point Proportions (Cosine Similarity) Across Dataset 1*

Performance of Qualitative Models on Dataset 1

| Models | Accuracy | Loss | F1 |
|---|---|---|---|
| Vector Decomposition | 0.591 | 0.931 | 0.552 |
| **Cosine Similarity** | **0.716** | **0.652** | **0.713** |
| Euclidean Distance | 0.648 | 0.769 | 0.608 |

Performance of Qualitative Models on Dataset 2

| Models | Accuracy | Loss | F1 |
|---|---|---|---|
| Vector Decomposition | 0.455 | 1.251 | 0.438 |
| **Cosine Similarity** | **0.541** | **1.128** | **0.527** |
| Euclidean Distance | 0.534 | 1.165 | 0.493 |

- Cosine similarity performs best
- Good accuracy is indicative of accurate marking point proportions

## Conclusion

Novel approach to grading short-answer physics questions

Models perform well even with low data

Reduces marking workload of teachers

Provides interpretable feedback

### Recommendations

Explore different word embeddings (E.g. ELMo)

Dropout layers to tackle overfitting

Hyperparameter optimisation (E.g. randomised or sequential search)

Evaluation on larger datasets from other domains to test scalability and adaptability

**References**
[1] P. Black and D. Wiliam, "Assessment and Classroom Learning", Assessment in Education: Principles, Policy & Practice, vol. 5, no. 1, pp. 7–74, 1998. doi: 10.1080/ 0969595980050102.
[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", 2014. arXiv: 1409.0473.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, pp. 993–1022, Mar. 2003, issn: 1532-4435. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937.

All images and graphs were self-drawn.