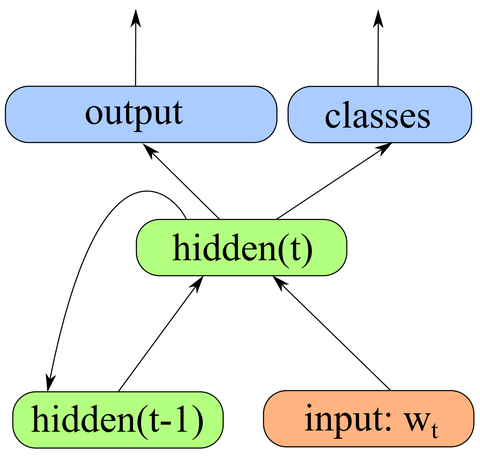
# Online Representation Learning in Recurrent Neural Language Models

**Marek Rei**
**University of Cambridge**

## Language Modelling

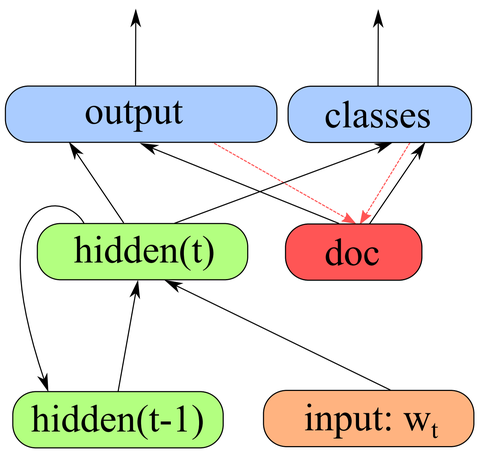
- **Recurrent neural network language models** (RNNLM) are some of the best-performing language models (Chelba et al., 2014).
- We investigate a modification of RNNLM, which allows it to efficiently **learn and adapt during testing**.
- We extend the idea of **Paragraph Vectors** (Le and Mikolov, 2014) to RNNLMs and apply it directly to language modelling.
- The new model achieves **lower perplexity** with **fewer parameters** and **fewer operations**.

## RNNLM




- **RNNLM** implementation based on Mikolov et al. (2011).
- The hidden layer from the previous time-step is used as input, creating a **recurrent connection**.
- Words are divided into larger **classes** to decrease the required computation in the output layer.
- Trained using **backpropagation through time** on complete sentences. Negative log-probability of the word sequence is used as the cost function.

## RNNLM with online learning




- Introducing an additional **document vector** to represent the unit of text being processed.
- It has no inputs and is **updated using backpropagation** after each word.
- It gets **initialised to a default state**, which is also optimised during training.
- This vector is continuously **updated during testing**, while all other parameters remain static.
- The document vector is optimised to represent how the current text **differs from the main language model**.
- The word used for updating the document vector for the next time-step is also available in the next input layer, therefore **the system receives no additional knowledge** as input.

## Experiments

- Evaluation performed on sentences from **English Wikipedia**. 10M words for training, 200K words for development, 4M words for testing.
- Vocabulary of **16,514 unique words** (frequency $\geq 30$), the rest replaced by the UNK tag.
- Increasing the document vector D gives a **larger improvement**, compared to increasing the hidden vector M by the same amount.

| | Train PPL | Dev PPL | Test PPL |
|---|---|---|---|
| Baseline M=100 | 92.65 | 103.56 | 102.51 |
| M=120 | 88.60 | 98.78 | 97.79 |
| M=100, D=20 | **87.28** | **95.36** | **94.39** |
| M=135 | 85.17 | 96.33 | 95.71 |
| M=100, D=35 | **80.11** | **91.05** | **90.29** |

**Table 1:** Perplexity when increasing either hidden vector M or document vector D.

## Parameters

- Adding the document vector or increasing the hidden layer requires **additional parameters and computation**.
- The document vector gives a larger improvement with a **smaller number of parameters**, compared to scaling the hidden vector.
- The graph of perplexity with respect to **additional operations** in the model also has a very similar shape.

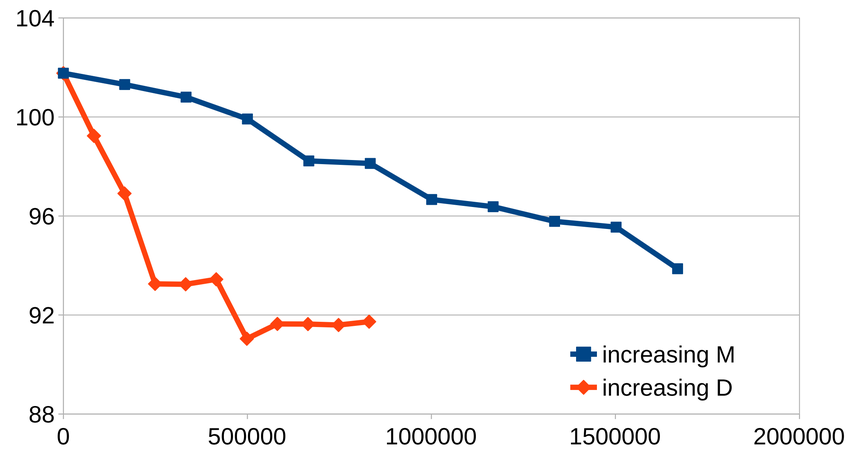



**Figure 1:** Perplexity as a function of additional parameters in the model.

## Conclusion

- The language model includes a separate vector to **represent the unit of text**, such as a sentence, being currently processed.
- The vector starts in a default state and is **continuously updated** using backpropagation.
- The modified language model achieves **lower perplexity** with a more optimal use of parameters and computation.

## Example: finding other sentences with similar document vectors

Both Hufnagel and Marston also joined the long-standing technical death metal band Gorguts.

1. The band eventually went on to become the post-hardcore band Adair.
2. The band members originally came from different death metal bands, bonding over a common interest in d-beat.
3. The proceeds went towards a home studio, which enabled him to concentrate on his solo output and songs that were to become his debut mini-album "Feeding The Wolves".