Does an LSTM forget more than a CNN? An empirical study of catastrophic forgetting in NLP

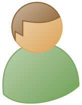
Gaurav Arora, Afshin Rahimi, Timothy Baldwin

The University of Melbourne

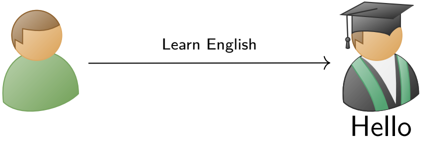
gaurava@student.unimelb.edu.au
{rahimia,tbaldwin}@unimelb.edu.au

December 3, 2019

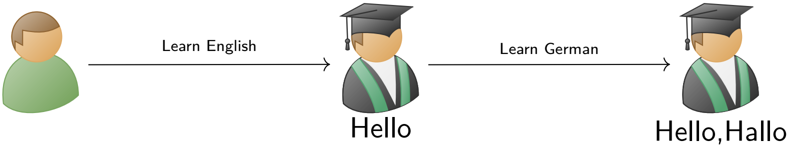
Perfect Scenario



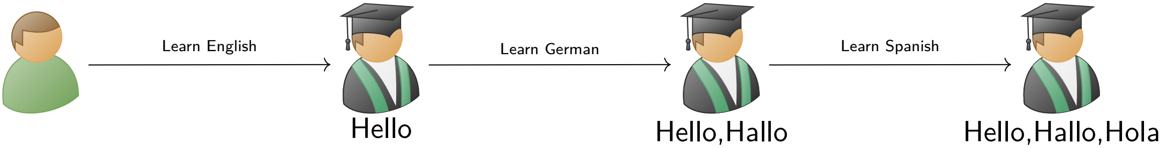
Perfect Scenario



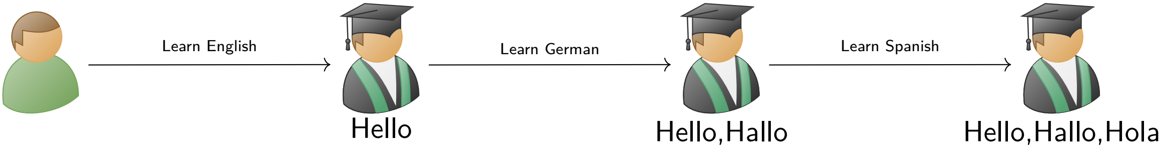
Perfect Scenario



Perfect Scenario

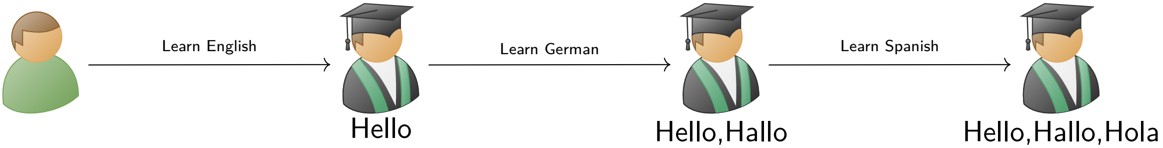


Perfect Scenario

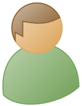


What if Sandy could remember only one language?

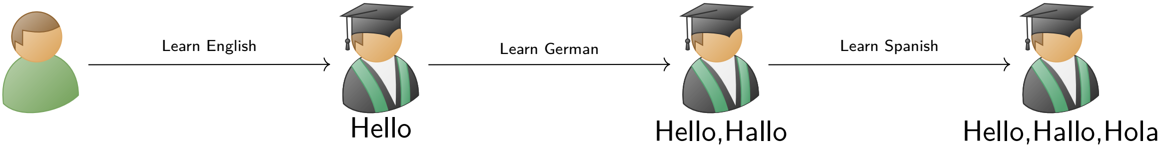
Perfect Scenario



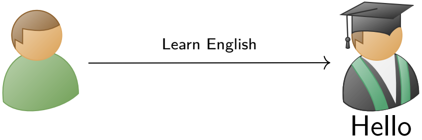
What if Sandy could remember only one language?



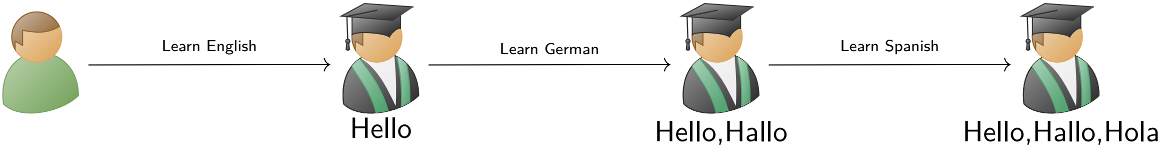
Perfect Scenario



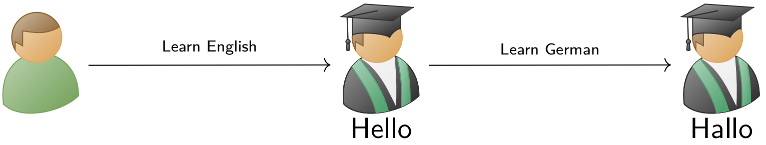
What if Sandy could remember only one language?



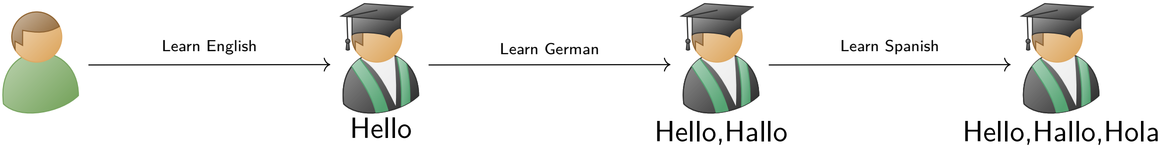
Perfect Scenario



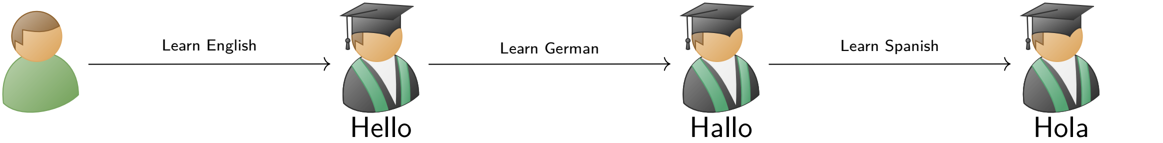
What if Sandy could remember only one language?



Perfect Scenario

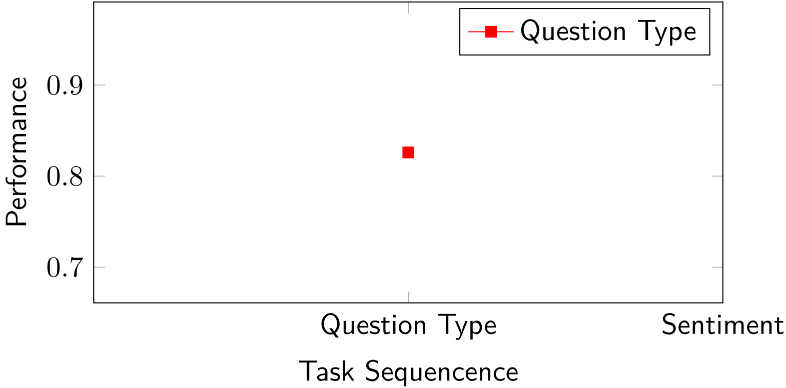


What if Sandy could remember only one language?



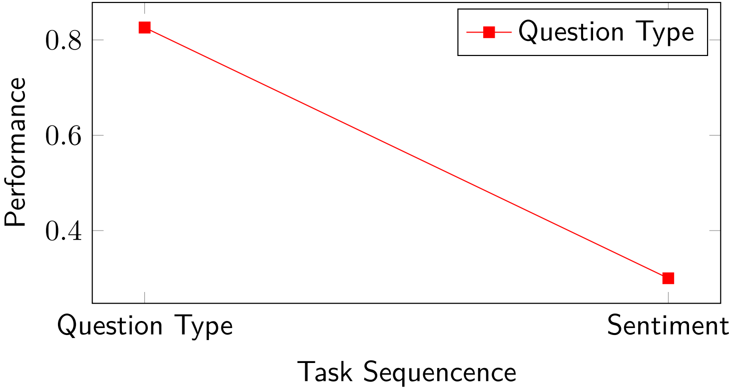
Catastrophic Forgetting

whereby a model trained on one task is fine-tuned on a second, and in doing so, suffers a "catastrophic" drop in performance over the first task.



Catastrophic Forgetting

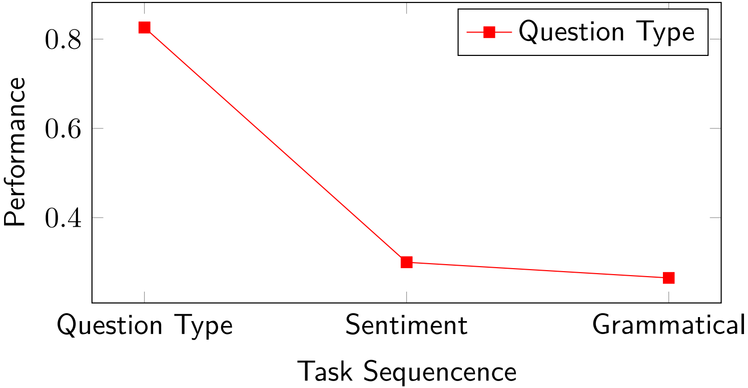
whereby a model trained on one task is fine-tuned on a second, and in doing so, suffers a "catastrophic" drop in performance over the first task.



Catastophic Forgetting

Catastrophic Forgetting

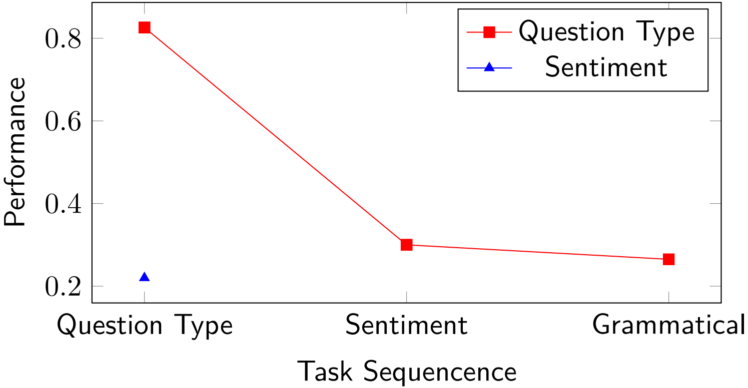
whereby a model trained on one task is fine-tuned on a second, and in doing so, suffers a "catastrophic" drop in performance over the first task.



Catastophic Forgetting

Catastrophic Forgetting

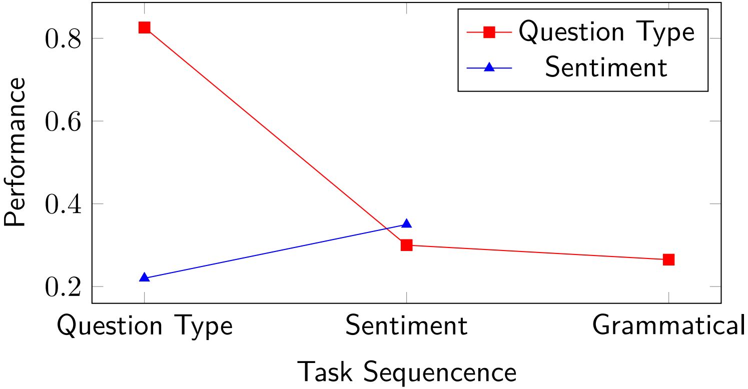
whereby a model trained on one task is fine-tuned on a second, and in doing so, suffers a “catastrophic” drop in performance over the first task.



Catastophic Forgetting

Catastrophic Forgetting

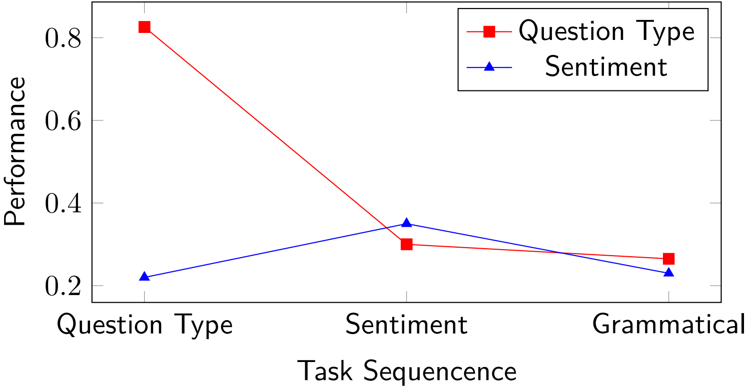
whereby a model trained on one task is fine-tuned on a second, and in doing so, suffers a "catastrophic" drop in performance over the first task.



Catastophic Forgetting

Catastrophic Forgetting

whereby a model trained on one task is fine-tuned on a second, and in doing so, suffers a "catastrophic" drop in performance over the first task.

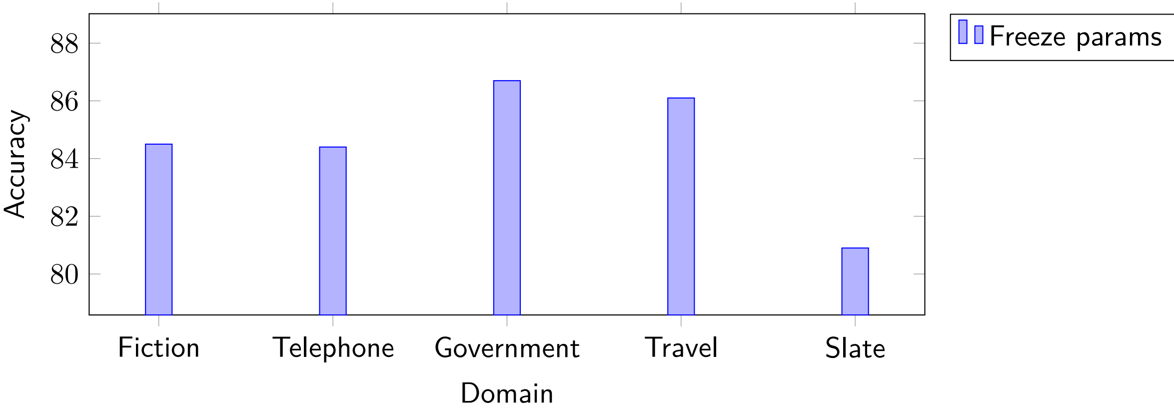


Motivation

Catastrophic forgetting is a hurdle in the development of better transfer learning techniques.

Motivation

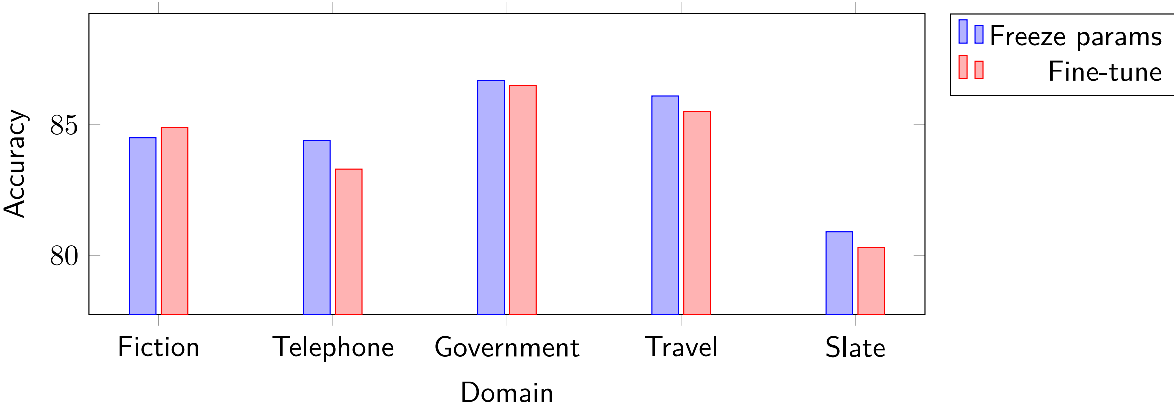
Catastrophic forgetting is a hurdle in the development of better transfer learning techniques.



Fine-tuning leads to forgetting when domains are different [Peters et al., 2019]

Motivation

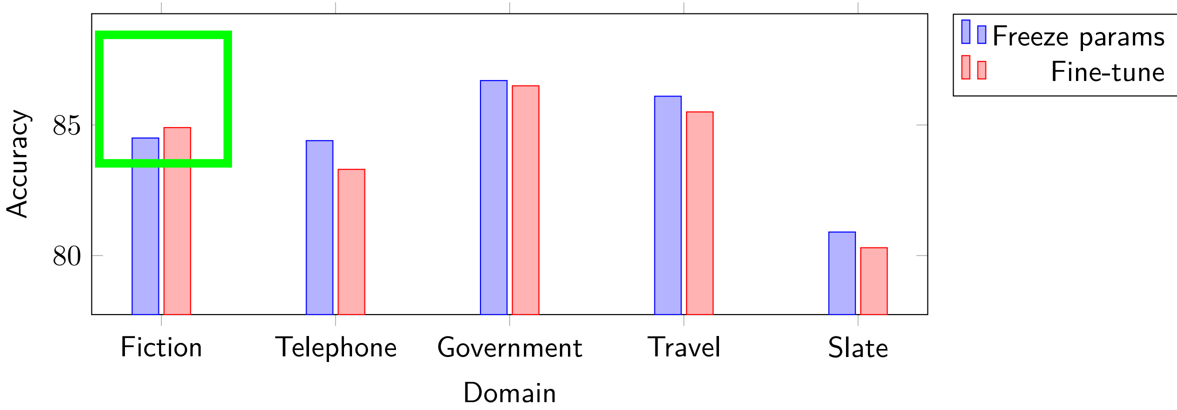
Catastrophic forgetting is a hurdle in the development of better transfer learning techniques.



Fine-tuning leads to forgetting when domains are different [Peters et al., 2019]

Motivation

Catastrophic forgetting is a hurdle in the development of better transfer learning techniques.



Fine-tuning leads to forgetting when domains are different [Peters et al., 2019]

Previous Studies

Task Complexity Task sequence's total complexity is positively correlated with the forgetting [Nguyen et al., 2019].

Previous Studies

Task Complexity Task sequence's total complexity is positively correlated with the forgetting [Nguyen et al., 2019].

Regularisation Using dropout decreases catastrophic forgetting [Goodfellow et al., 2014].

Previous Studies

Task Complexity Task sequence's total complexity is positively correlated with the forgetting [Nguyen et al., 2019].

Regularisation Using dropout decreases catastrophic forgetting [Goodfellow et al., 2014].

Max Operation Using max operation in the network reduces forgetting [Srivastava et al., 2013].

Tasks

TREC Question classification ("TREC"): coarse-grained classification of questions, based on 6 classes. [Voorhees and Tice, 1999]

Tasks	SOTA	CNN	$(\text{SOTA} - \text{CNN})$	Type
TREC	0.98	0.91	0.07	Easy

Tasks

TREC Question classification (“TREC”): coarse-grained classification of questions, based on 6 classes. [Voorhees and Tice, 1999]

Subjectivity (“SUBJ”): binary classification of Subjectivity vs. Objectivity in IMDB reviews. [Pang and Lee, 2004]

Tasks	SOTA	CNN	$(\text{SOTA} - \text{CNN})$	Type
TREC	0.98	0.91	0.07	Easy
Subjectivity	0.95	0.92	0.03	Easy

Tasks

TREC Question classification ("TREC"): coarse-grained classification of questions, based on 6 classes. [Voorhees and Tice, 1999]

Subjectivity ("SUBJ"): binary classification of Subjectivity vs. Objectivity in IMDB reviews. [Pang and Lee, 2004]

Corpus of Linguistic Acceptability ("CoLA"): prediction of whether a sentence is grammatical or not. [Warstadt et al., 2018]

Tasks	SOTA	CNN	$(\mathrm{SOTA} - \mathrm{CNN})$	Type
TREC	0.98	0.91	0.07	Easy
Subjectivity	0.95	0.92	0.03	Easy
CoLA	0.34	0.24	0.1	Hard

Tasks

TREC Question classification ("TREC"): coarse-grained classification of questions, based on 6 classes. [Voorhees and Tice, 1999]

Subjectivity ("SUBJ"): binary classification of Subjectivity vs. Objectivity in IMDB reviews. [Pang and Lee, 2004]

Corpus of Linguistic Acceptability ("CoLA"): prediction of whether a sentence is grammatical or not. [Warstadt et al., 2018]

Stanford Sentiment Treebank ("SST"): fine-grained sentiment classification over five classes. [Socher et al., 2013]

Tasks	SOTA	CNN	$(\text{SOTA} - \text{CNN})$	Type
TREC	0.98	0.91	0.07	Easy
Subjectivity	0.95	0.92	0.03	Easy
CoLA	0.34	0.24	0.1	Hard
SST	0.54	0.38	0.16	Hard

Tasks

TREC Question classification ("TREC"): coarse-grained classification of questions, based on 6 classes. [Voorhees and Tice, 1999]

Subjectivity ("SUBJ"): binary classification of Subjectivity vs. Objectivity in IMDB reviews. [Pang and Lee, 2004]

Corpus of Linguistic Acceptability ("CoLA"): prediction of whether a sentence is grammatical or not. [Warstadt et al., 2018]

Stanford Sentiment Treebank ("SST"): fine-grained sentiment classification over five classes. [Socher et al., 2013]

Tasks	SOTA	CNN	$\frac{(\mathrm{SOTA}-\mathrm{CNN})}{\mathrm{SOTA}}$	Type
TREC	0.98	0.91	0.07	Easy
Subjectivity	0.95	0.92	0.03	Easy
CoLA	0.34	0.24	0.29	Hard
SST	0.54	0.38	0.29	Hard

Continual Learning Setup

- Tasks are trained sequentially.

Continual Learning Setup

- Tasks are trained sequentially.
- Tasks are trained without access to data from previous tasks.

Evaluation: Forgetting Metric

Normalisation

Lower Bound: normalise performance based on Majority Classifier(MAJ) performance.

Lower Bound: normalise performance based on Majority Classifier(MAJ) performance.

Upper Bound: normalise performance based on State of the Art Accuracy (SOTA).

Evaluation: Forgetting Metric

Normalisation

Lower Bound: normalise performance based on Majority Classifier(MAJ) performance.

Upper Bound: normalise performance based on State of the Art Accuracy (SOTA).

$$P_{i,j} = \frac{\mathrm{PER}_{i,j} - \mathrm{PER}_{\mathsf{MAJ}}}{\mathrm{PER}_{\mathsf{SOTA}} - \mathrm{PER}_{\mathsf{MAJ}}} \quad \forall i \leq j$$

Evaluation: Forgetting Metric

Normalisation

Lower Bound: normalise performance based on Majority Classifier(MAJ) performance.

Upper Bound: normalise performance based on State of the Art Accuracy (SOTA).

$$P_{i,j} = \frac{\mathrm{PER}_{i,j} - \mathrm{PER}_{\mathsf{MAJ}}}{\mathrm{PER}_{\mathsf{SOTA}} - \mathrm{PER}_{\mathsf{MAJ}}} \quad \forall i \leq j$$

Incorporates a measure of task difficulty.

Evaluation: Forgetting Metric

Sequence Forgetting (F_{Seq})

Task forgetting is percentage performance drop from when task was first trained.

$$F_i = \frac{P_{i,i} - P_{i,T}}{|P_{i,i}|} \tag{1}$$

Evaluation: Forgetting Metric

Sequence Forgetting (F_{Seq})

Task forgetting is percentage performance drop from when task was first trained.

$$F_i = \frac{P_{i,i} - P_{i,T}}{|P_{i,i}|} \tag{1}$$

Sequence forgetting is sum of forgetting for all tasks.

$$F_{Seq} = \sum_{i=1}^{i=T} F_i \tag{2}$$

Evaluation: Forgetting Metric

Sequence Forgetting (F_{Seq})

Task forgetting is percentage performance drop from when task was first trained.

$$F_i = \frac{P_{i,i} - P_{i,T}}{|P_{i,i}|} \tag{1}$$

Sequence forgetting is sum of forgetting for all tasks.

$$F_{Seq} = \sum_{i=1}^{i=T} F_i \tag{2}$$

Lower forgetting is better.

Experiments

Research Question 1

Do some neural architectures forget more than others?

CNN vs LSTM

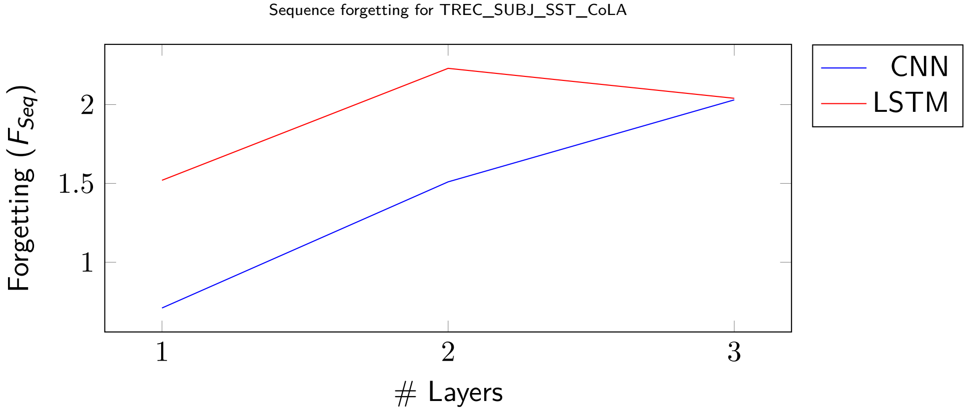
- Compare forgetting (F_{Seq}) between LSTM and CNN architecture.

CNN vs LSTM

- Compare forgetting (F_{Seq}) between LSTM and CNN architecture.
- Across all 24 possible Task Sequence from four tasks.

CNN vs LSTM

- Compare forgetting (F_{Seq}) between LSTM and CNN architecture.
- Across all 24 possible Task Sequence from four tasks.



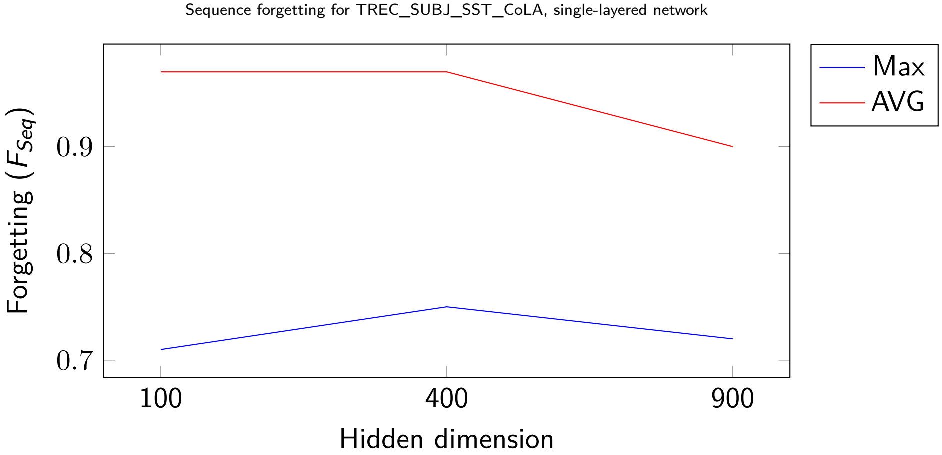
CNN forgets less than LSTM.

Max vs Average pooling

What makes CNN forget less, max-pooling vs average pooling?

Max vs Average pooling

What makes CNN forget less, max-pooling vs average pooling?



CNN forgets less due to max-pooling.

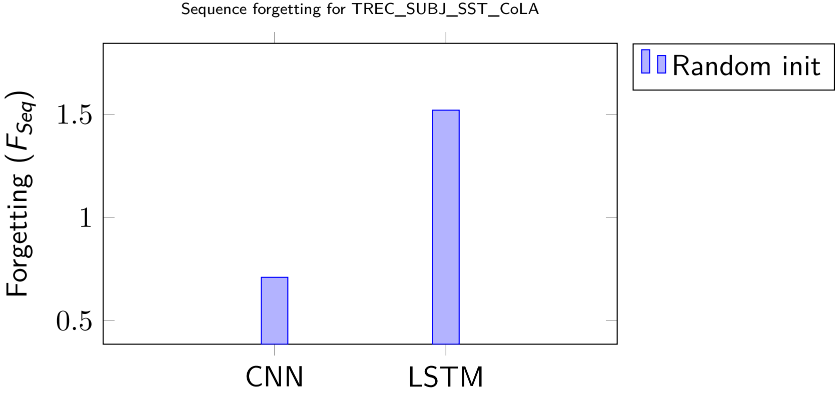
Experiments

Research Question 2

Should we fine-tune pre-trained embedding in continual learning setup?

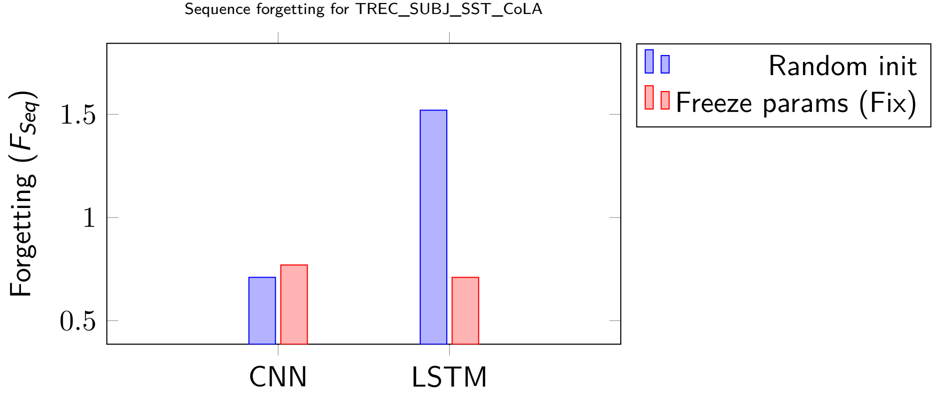
Experiments: ELMo Embeddings

- Using ELMo embeddings as feature extractor (Fix) vs fine-tunning (FT).



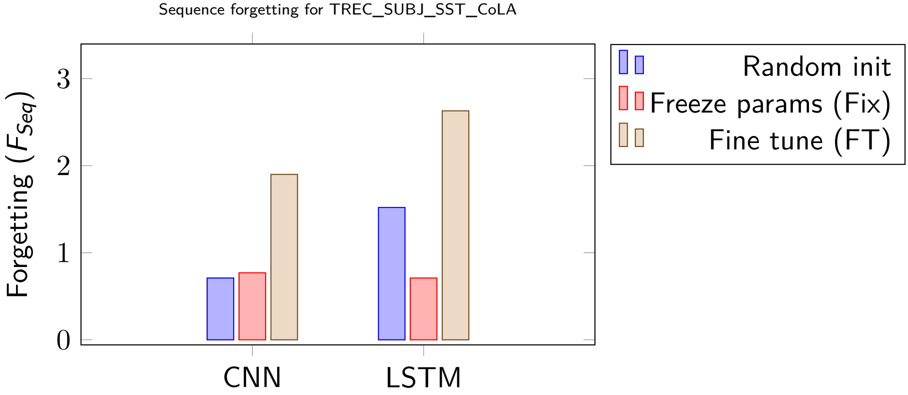
Experiments: ELMo Embeddings

- Using ELMo embeddings as feature extractor (Fix) vs fine-tunning (FT).



Experiments: ELMo Embeddings

- Using ELMo embeddings as feature extractor (Fix) vs fine-tunning (FT).



Freezing params is better in continual learning setup.

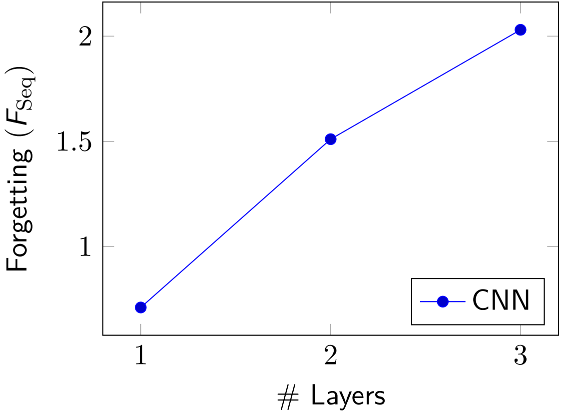
Experiments

Research Question 3

Do networks with more capacity forget less?

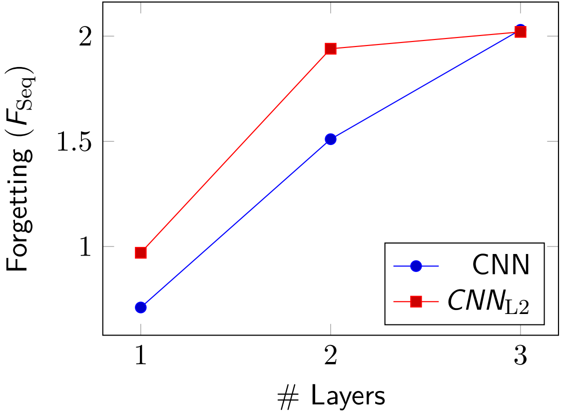
Layers

Does increasing the number of layers decrease forgetting?

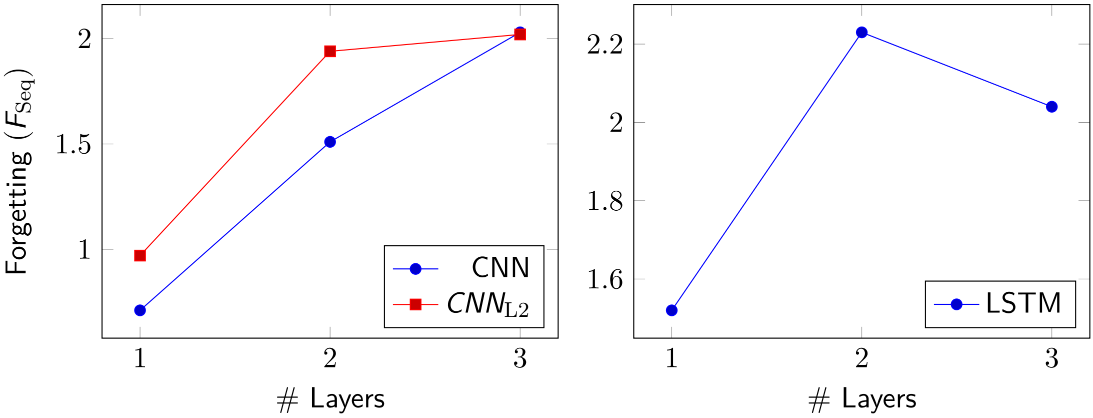


Layers

Does increasing the number of layers decrease forgetting?

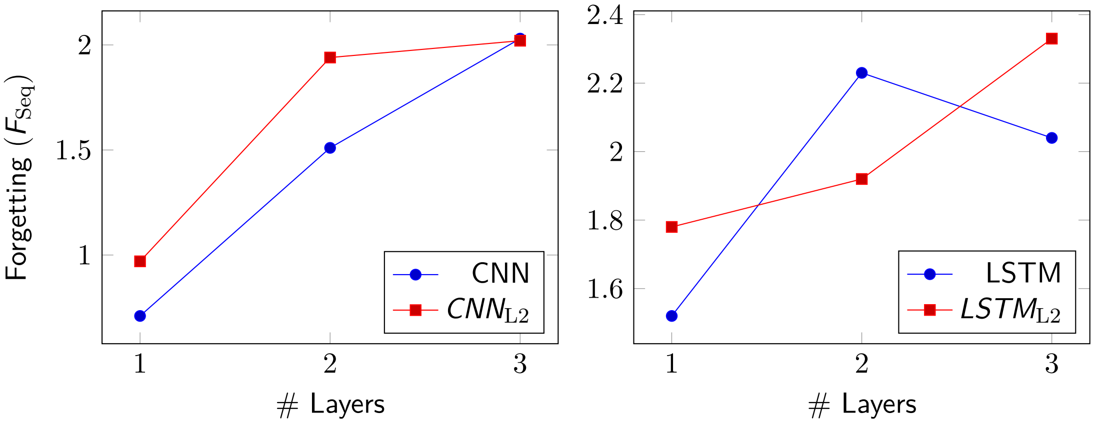


Does increasing the number of layers decrease forgetting?



Layers

Does increasing the number of layers decrease forgetting?



Experiments

Research Question 4

Do networks forget more during training over a difficult task?

Task Sequencing

Curriculum Learning, placing hard tasks at the end of a task sequence [Bengio et al., 2009] reduces forgetting.

Task Sequence	F_{Seq}
TREC_SUBJ_CoLA_SST	0.63
TREC_SUBJ_SST_CoLA	0.78
SST_TREC_SUBJ_CoLA	0.81

Task Sequencing

Curriculum Learning, placing hard tasks at the end of a task sequence [Bengio et al., 2009] reduces forgetting.

Task Sequence	F_{Seq}
TREC_SUBJ_CoLA_SST	0.63
TREC_SUBJ_SST_CoLA	0.78
SST_TREC_SUBJ_CoLA	0.81
CoLA_SUBJ_SST_TREC	1.3
SST_CoLA_SUBJ_TREC	1.4
CoLA_SST_SUBJ_TREC	1.4

Table: Top three (Green) and bottom three (Red) tasks sequence with F_{Seq} for Layer $= 1$.

Findings

Hard Tasks

Training hard task later in the sequence is beneficial.

Findings

Hard Tasks

Training hard task later in the sequence is beneficial.

Embeddings

Fine-tuning embeddings perform worse than using them as a feature extractor.

Findings

Hard Tasks

Training hard task later in the sequence is beneficial.

Embeddings

Fine-tuning embeddings perform worse than using them as a feature extractor.

Architecture

CNN forgets less due to max-pooling.

Findings

Hard Tasks

Training hard task later in the sequence is beneficial.

Embeddings

Fine-tuning embeddings perform worse than using them as a feature extractor.

Architecture

CNN forgets less due to max-pooling.

Thanks!

: gauravaror/catastrophic_forgetting

References

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009).
Curriculum learning.
In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

Goodfellow, I. J., Mirza, M., Da, X., Courville, A. C., and Bengio, Y. (2014).
An empirical investigation of catastrophic forgeting in gradient-based neural networks.
In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Nguyen, C. V., Achille, A., Lam, M., Hassner, T., Mahadevan, V., and Soatto, S. (2019).
Toward understanding catastrophic forgetting in continual learning.
CoRR, abs/1908.01091.

Pang, B. and Lee, L. (2004).
A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.

CNN vs LSTM

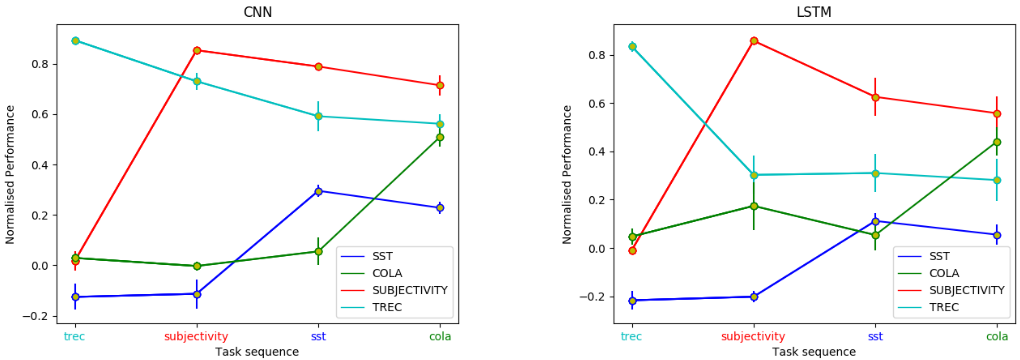


Figure: Performance of LSTM and CNN on task sequence TREC_SUBJ_SST_COLA, with one layer and hidden dimension 100.

More Network Capacity: Hidden Dimension

- We couldn't find any conclusive answer for how forgetting changes with hidden dimension.
- It seems to be dependent on the task sequence.

Regularisation

Does L2 regularisation help in decreasing forgetting?