An investigative study of catastrophic forgetting in NLP

Gaurav Arora

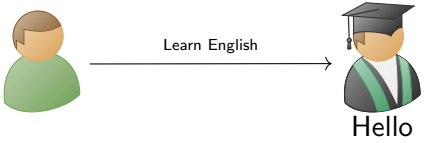
The University of Melbourne

Supervisors: Timothy Baldwin, Afshin Rahimi

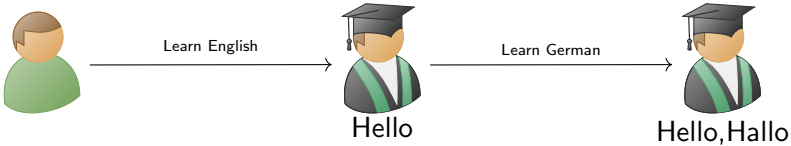
June 5, 2020

Perfect Scenario

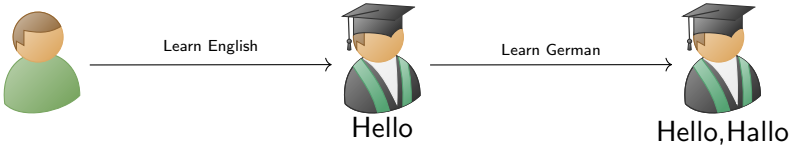
Perfect Scenario



Perfect Scenario

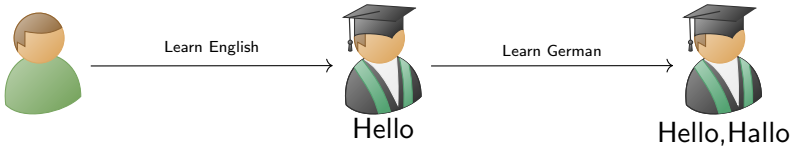


Perfect Scenario



ML models can remember only one language!!

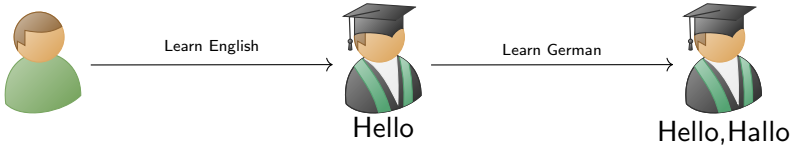
Perfect Scenario



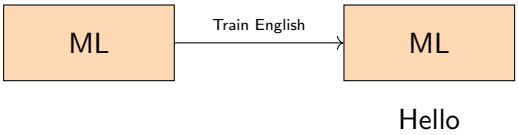
ML models can remember only one language!!

ML

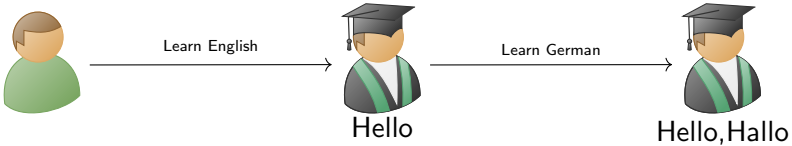
Perfect Scenario



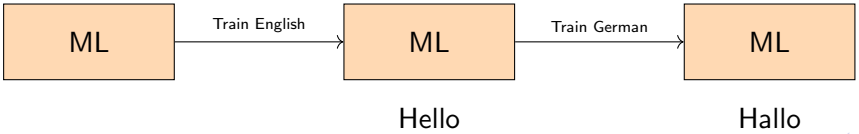
ML models can remember only one language!!



Perfect Scenario



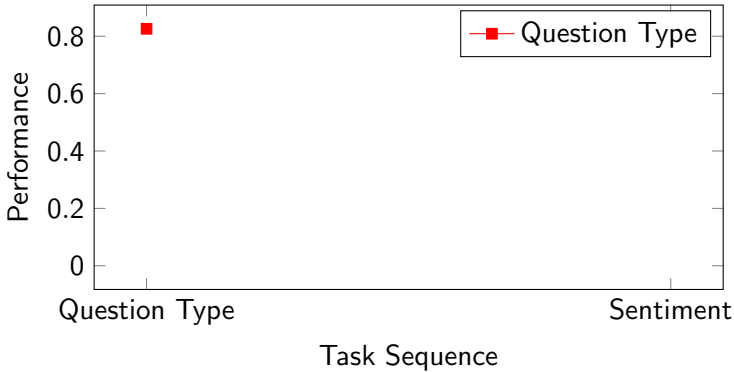
ML models can remember only one language!!



Catastophic Forgetting

Catastrophic Forgetting

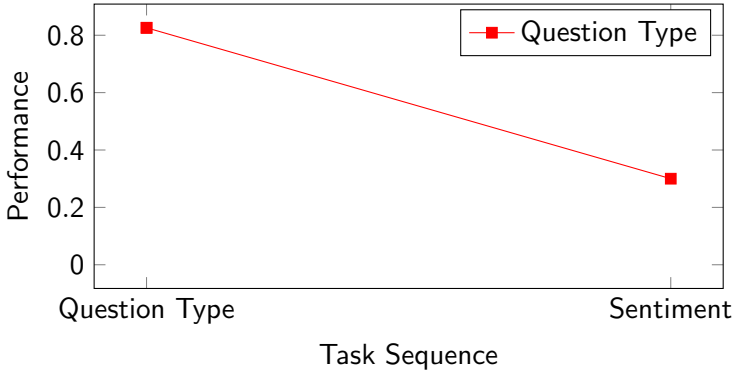
whereby a model trained on one task is fine-tuned on a second, and in doing so, suffers a "catastrophic" drop in performance over the first task.



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

Why is it important?

Motivation

Tesla HydraNet: Need for Sequential Training

Tesla needs to retrain the whole model every time there is a misclassification for any task.

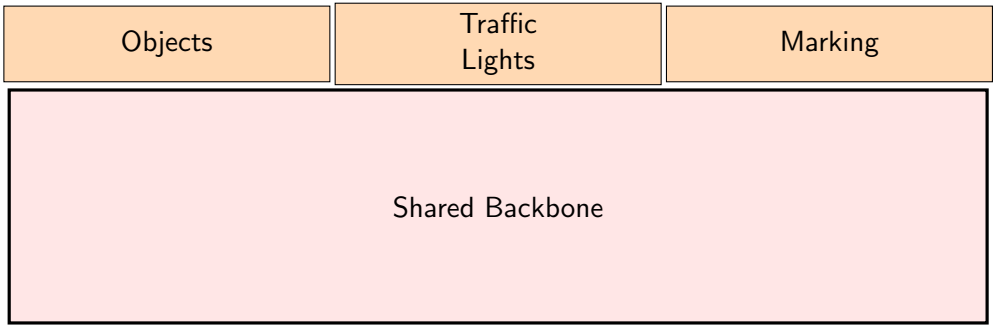


Figure: Tesla HydraNet Model (https://www.youtube.com/watch?v=hx7BXih7zx8)

Research Objectives

RO1: Do current methods to reduce forgetting work for NLP tasks?

Research Objectives

RO1: Do current methods to reduce forgetting work for NLP tasks?

RO2: Comparing forgetting in:

- CNN Vs LSTM
- Networks with different capacity

Research Objectives

RO1: Do current methods to reduce forgetting work for NLP tasks?

RO2: Comparing forgetting in:

- CNN Vs LSTM
- Networks with different capacity

RO3: Proposed methods to reduce catastrophic forgetting:

- Annealing Temperature schedule.
- Adding Task information.

Tasks

TREC Question classification ("TREC"): classify questions into `NUM`, `DESC`, `LOC`, `ENTY`, `HUM`, `ABBR`.

Tasks

TREC Question classification ("TREC"): classify questions into `NUM`, `DESC`, `LOC`, `ENTY`, `HUM`, `ABBR`.

Subjectivity ("SUBJ"): binary classification of Subjectivity vs. Objectivity in IMDB reviews.

Tasks

TREC Question classification ("TREC"): classify questions into `NUM`, `DESC`, `LOC`, `ENTY`, `HUM`, `ABBR`.

Subjectivity ("SUBJ"): binary classification of Subjectivity vs. Objectivity in IMDB reviews.

Corpus of Linguistic Acceptability ("CoLA"): prediction of whether a sentence is grammatical or not.

Tasks

TREC Question classification ("TREC"): classify questions into `NUM`, `DESC`, `LOC`, `ENTY`, `HUM`, `ABBR`.

Subjectivity ("SUBJ"): binary classification of Subjectivity vs. Objectivity in IMDB reviews.

Corpus of Linguistic Acceptability ("CoLA"): prediction of whether a sentence is grammatical or not.

Stanford Sentiment Treebank ("SST"): fine-grained sentiment classification over five ratings (1 lowest).

Continual Learning Setup

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

RO1: Do current methods to reduce forgetting work for NLP tasks?

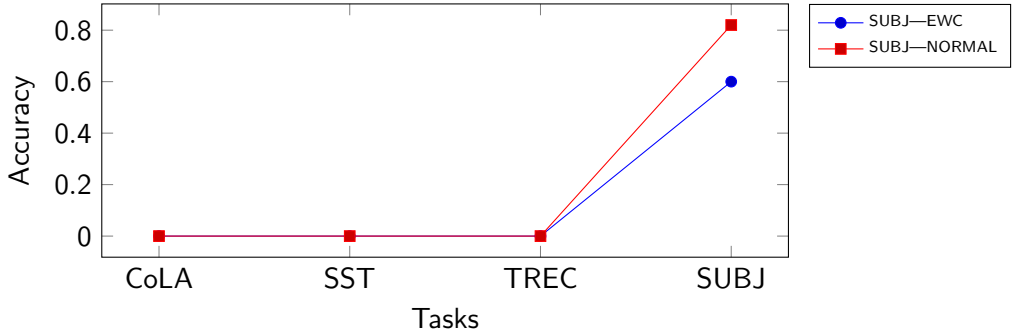
Elastic Weight Consolidation

Observation: Elastic Weight Consolidation reduces forgetting but hinders the learning of future tasks.

RO1: Do current methods to reduce forgetting work for NLP tasks?

Elastic Weight Consolidation

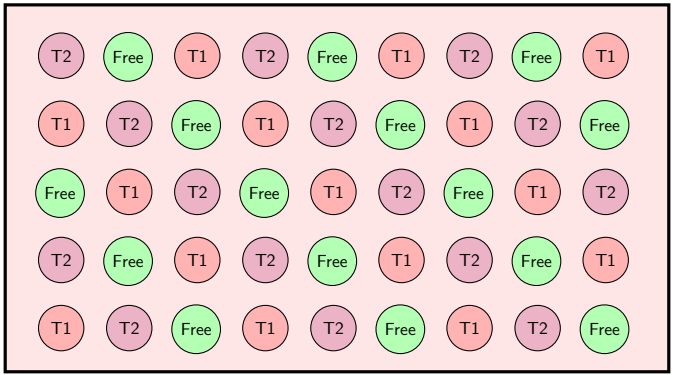
Observation: Elastic Weight Consolidation reduces forgetting but hinders the learning of future tasks.



RO1: Do current methods to reduce forgetting work for NLP tasks?

Elastic Weight Consolidation

Observation: Elastic Weight Consolidation reduces forgetting but hinders the learning of future tasks.



RO2: Comparing Architectures

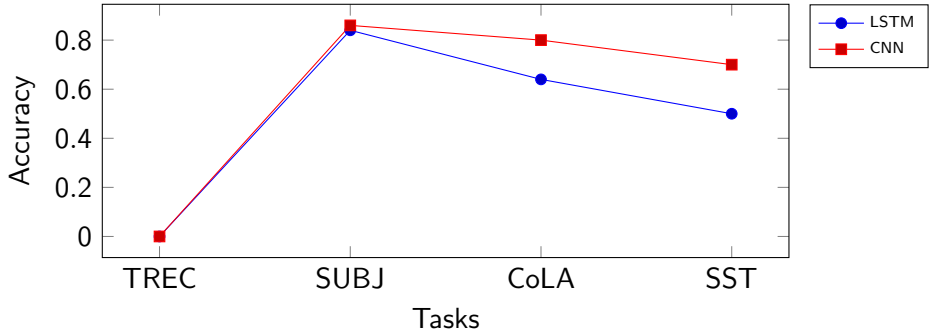
CNNs vs LSTMs

Observation: CNNs forget less than LSTMs due to max-pooling operation.

RO2: Comparing Architectures

CNNs vs LSTMs

Observation: CNNs forget less than LSTMs due to max-pooling operation.



RO2: Findings from Investigative Study

Architecture

CNN forgets less due to max-pooling.

Network Capacity

Deeper network forgets more than a shallow network.

RO3: Adding task information reduces forgetting

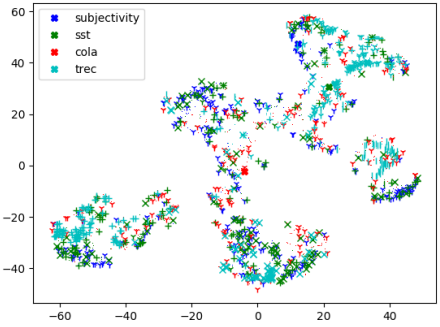
Task Information

Adding task information reduces forgetting.

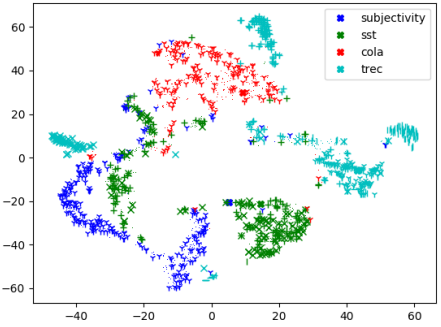
RO3: Adding task information reduces forgetting

Task Information

Adding task information reduces forgetting.



(a) LSTM



(b) LSTM with Task Information

RO3: Decreasing temperature schedule in softmax layer

Softmax Temperature

Temperature annealing reduces forgetting

RO3: Decreasing temperature schedule in softmax layer

Softmax Temperature

Temperature annealing reduces forgetting

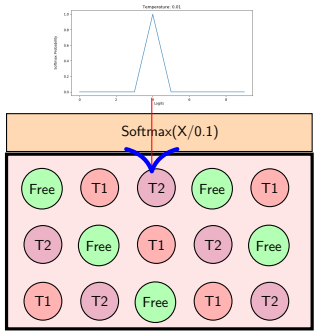


Figure: peaked output distribution

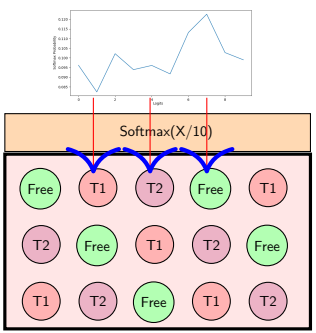


Figure: soft output distribution

Summary

Continual Learning Setup

- Four tasks: `TREC, CoLA, SST, Subjectivity`
- Task trained sequentially without access to previous tasks training data.
- Model Architecture: `Multi-head Setup`

Primary Findings

- CNN forgets less due to max-pooling.
- Training hard task later in the sequence is beneficial.
- Adding task information reduces forgetting.
- Temperature annealing reduces forgetting.

Summary

Continual Learning Setup

- Four tasks: `TREC, CoLA, SST, Subjectivity`
- Task trained sequentially without access to previous tasks training data.
- Model Architecture: `Multi-head Setup`

Primary Findings

- CNN forgets less due to max-pooling.
- Training hard task later in the sequence is beneficial.
- Adding task information reduces forgetting.
- Temperature annealing reduces forgetting.

Thanks!

: gauravaror/catastrophic_forgetting