Report

Vaibhav Gupta
M.Tech, CSE

April 1, 2024

1 Introduction

ELMo (Embeddings from Language Models) is a state-of-the-art model for generating contextualized word representations. Unlike static embeddings, ELMo leverages bidirectional LSTM (Bi-LSTM) layers to capture nuanced word meanings based on their context. This report evaluates ELMo's performance against traditional methods like SVD and Skip-gram on a news classification task.

2 Dataset Overview

The dataset consists of:

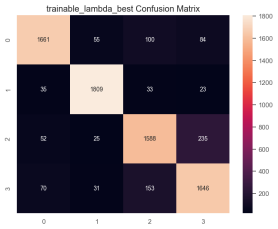
- **Training data**: 57340 sentences (30,000 per class) with a maximum length of 50 words.

3 Hyperparameter Tuning

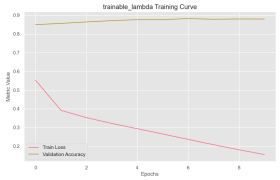
3.1 Trainable λs

- Batch size: 128
- Embedding dimension: 300
- Learning rate: 0.001
- Epochs: 10

Results: **Train Accuracy: 96.38%**, **Test Accuracy: 88.21%**.



(a) Confusion Matrix



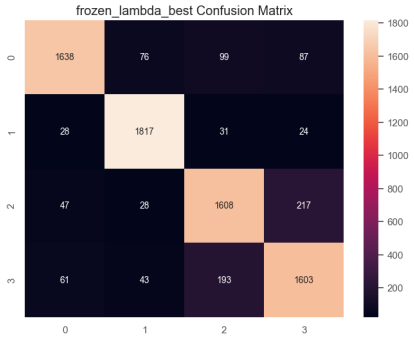
(b) Training Curve

Figure 1: Trainable λs: Results and Training Dynamics

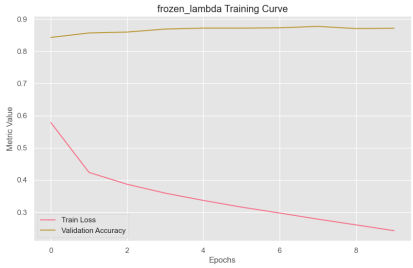
3.2 Frozen λs

- λs initialized randomly between 0.5 and 1 and frozen.
- Other parameters identical to trainable λs setup.

Results: **Train Accuracy: 93.18%**, **Test Accuracy: 87.71%**.



(a) Confusion Matrix



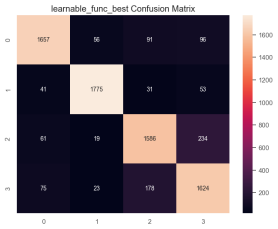
(b) Training Curve

Figure 2: Frozen λs: Results and Training Dynamics

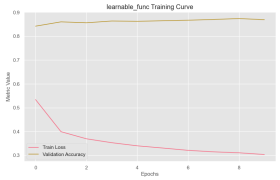
3.3 Learnable Function

- Neural network-based function to combine embeddings adaptively.
- Architecture: Two-layer MLP with ReLU activation.

Results: **Train Accuracy: 92.01%**, **Test Accuracy: 87.39%**.



(a) Confusion Matrix

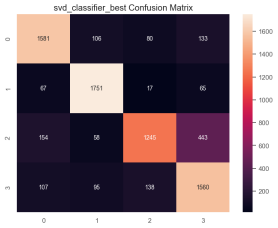


(b) Training Curve

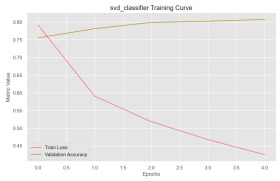
Figure 3: Learnable Function: Results and Training Dynamics

4 Base Line Models

4.1 SVD



(a) Confusion Matrix

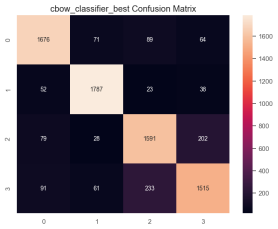


(b) Training Curve

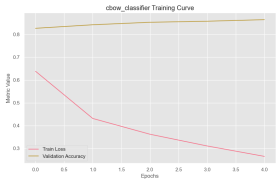
Figure 4: SVD: Results and Training Dynamics

Results: **Train Accuracy: 86.0%**, **Test Accuracy: 80.75%**.

4.2 CBOW



(a) Confusion Matrix

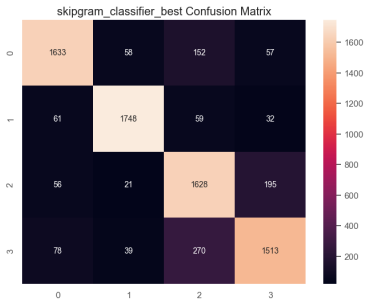


(b) Training Curve

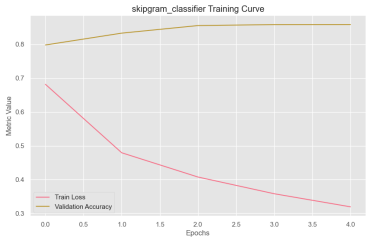
Figure 5: CBOW: Results and Training Dynamics

Results: **Train Accuracy: 93.70%**, **Test Accuracy: 86.43%**.

4.3 Skipgram



(a) Confusion Matrix



(b) Training Curve

Figure 6: Skipgram: Results and Training Dynamics

Results: **Train Accuracy: 90.10%**, **Test Accuracy: 85.82%**.

5 Model Comparison

Table 1: Performance of Embedding Models on Downstream Classification

Model	**Train Accuracy**	**Test Accuracy**
ELMo (Trainable λs)	96.38%	88.21%
ELMo (Frozen λs)	93.18%	87.71%
ELMo (Learnable Function)	92.01%	87.39%
CBOW	93.70%	86.43%
Skip-gram	90.10%	85.82%
SVD	86.00%	80.75%

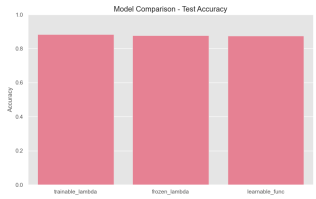


Figure 7: Test Accuracy Comparison Across Models

6 Analysis

6.1 Why ELMo Outperforms others?

- **Contextualized Representations**: Adapts embeddings based on sentence context.
- **Bidirectional LSTM**: Captures dependencies in both forward and backward directions.
- **Subword Handling**: Effective for rare/unseen words.
- **Higher Dimensionality**: 300-dimensional embeddings retain richer semantics.

Train Classification Report:

	precision	recall	f1-score	support
World	0.88	0.87	0.87	30000
Sports	0.92	0.96	0.94	30000
Business	0.90	0.73	0.81	30000
Sci/Tech	0.77	0.88	0.82	30000
accuracy			0.86	120000
macro avg	0.86	0.86	0.86	120000
weighted avg	0.86	0.86	0.86	120000

Test Classification Report:

	precision	recall	f1-score	support
World	0.83	0.83	0.83	1900
Sports	0.87	0.92	0.90	1900
Business	0.84	0.66	0.74	1900
Sci/Tech	0.71	0.82	0.76	1900
accuracy			0.81	7600
macro avg	0.81	0.81	0.81	7600
weighted avg	0.81	0.81	0.81	7600

(a) SVD Classifier Report

Train Classification Report:

	precision	recall	f1-score	support
World	0.93	0.89	0.91	30000
Sports	0.96	0.96	0.96	30000
Business	0.84	0.91	0.87	30000
Sci/Tech	0.89	0.85	0.87	30000
accuracy			0.90	120000
macro avg	0.90	0.90	0.90	120000
weighted avg	0.90	0.90	0.90	120000

Test Classification Report:

	precision	recall	f1-score	support
World	0.89	0.86	0.88	1900
Sports	0.94	0.92	0.93	1900
Business	0.77	0.86	0.81	1900
Sci/Tech	0.84	0.80	0.82	1900
accuracy			0.86	7600
macro avg	0.86	0.86	0.86	7600
weighted avg	0.86	0.86	0.86	7600

(b) Skip-gram Classifier Report

Train Classification Report:

	precision	recall	f1-score	support
World	0.94	0.93	0.94	30000
Sports	0.96	0.99	0.97	30000
Business	0.90	0.91	0.91	30000
Sci/Tech	0.91	0.89	0.90	30000
accuracy			0.93	120000
macro avg	0.93	0.93	0.93	120000
weighted avg	0.93	0.93	0.93	120000

Test Classification Report:

	precision	recall	f1-score	support
World	0.88	0.88	0.88	1900
Sports	0.92	0.94	0.93	1900
Business	0.82	0.84	0.83	1900
Sci/Tech	0.83	0.80	0.81	1900
accuracy			0.86	7600
macro avg	0.86	0.86	0.86	7600
weighted avg	0.86	0.86	0.86	7600

(c) CBOW Classifier Report

Figure 8: Classification Reports for Baseline Models

6.2 Learnable Function Superiority

- Dynamically optimizes layer combination weights.

- Captures non-linear relationships between embeddings.
- Task-specific optimization improves generalization.

Train Classification Report for trainable_lambda:

	precision	recall	f1-score	support
World	0.98	0.97	0.97	30000
Sports	0.98	0.99	0.98	30000
Business	0.92	0.95	0.94	30000
Sci/Tech	0.95	0.91	0.93	30000
accuracy			0.96	120000
macro avg	0.96	0.96	0.96	120000
weighted avg	0.96	0.96	0.96	120000

(a) Train classification report of trainable λs

Classification Report for trainable_lambda:

	precision	recall	f1-score	support
World	0.89	0.88	0.89	1900
Sports	0.94	0.95	0.94	1900
Business	0.85	0.84	0.84	1900
Sci/Tech	0.84	0.84	0.84	1900
accuracy			0.88	7600
macro avg	0.88	0.88	0.88	7600
weighted avg	0.88	0.88	0.88	7600

trainable_lambda Best Accuracy: 0.8821

(b) Test classification report ofTrainable λs

Train Classification Report for frozen_lambda:

	precision	recall	f1-score	support
World	0.96	0.91	0.93	30000
Sports	0.94	0.99	0.97	30000
Business	0.91	0.89	0.90	30000
Sci/Tech	0.89	0.91	0.90	30000
accuracy			0.93	120000
macro avg	0.93	0.93	0.93	120000
weighted avg	0.93	0.93	0.93	120000

(c) Train classification report of frozen λs

Classification Report for frozen_lambda:

	precision	recall	f1-score	support
World	0.89	0.88	0.88	1900
Sports	0.95	0.93	0.94	1900
Business	0.85	0.81	0.83	1900
Sci/Tech	0.81	0.86	0.83	1900
accuracy			0.87	7600
macro avg	0.87	0.87	0.87	7600
weighted avg	0.87	0.87	0.87	7600

frozen_lambda Best Accuracy: 0.8771

(d) Test classification report of frozen λs

Train Classification Report for learnable_func:

	precision	recall	f1-score	support
	precision	recall	f1-score	support
World	0.91	0.90	0.91	30000
World	0.91	0.90	0.91	30000
Sports	0.92	0.98	0.95	30000
Business	0.90	0.83	0.86	30000
Business	0.90	0.83	0.86	30000
Sci/Tech	0.85	0.87	0.86	30000
Sci/Tech	0.85	0.87	0.86	30000

(e) Train classification report of learnable λs

Classification Report for learnable_func:

	precision	recall	f1-score	support
World	0.91	0.86	0.89	1900
Sports	0.91	0.96	0.94	1900
Business	0.84	0.82	0.83	1900
Sci/Tech	0.82	0.84	0.83	1900
accuracy			0.87	7600
macro avg	0.87	0.87	0.87	7600
weighted avg	0.87	0.87	0.87	7600

learnable_func Best Accuracy: 0.8739

(f) Test classification report of learnable λs

Figure 9: Classification Reports for ELMo Variants

7 Conclusion

ELMo with trainable λs achieves the highest test accuracy (88.21%), surpassing both static embeddings and fixed-weight ELMo variants. The ability to dynamically adjust layer contributions allows for better contextual representation, making it well-suited for downstream NLP tasks. The increased train accuracy suggests stronger feature learning, though potential overfitting should be

monitored. Training curves and confusion matrices further validate the model's robustness and effectiveness.