Redundant Complexity in Deep Learning:
An Efficacy Analysis of NeXtVLAD in NLP

Sina Mahdipour Saravani

Committee: Dr. Indrakshi Ray (Advisor), Dr. Ritwik Banerjee (Co-Advisor), Dr. Steve Simske

June 23, 2022

Colorado State University

Outline

- Introduction
- NeXtVLAD, the Studied Architecture, and Prerequisites
- Sarcasm Detection
- Deepfake Text Detection
- Conclusion

Introduction

Introduction

- Extensive expressive power of deep learning
- Less human expert intervention and effort

→ The ubiquity and success of deep learning

→ Inherent promotion of a naïve and negatively simplistic employment

Introduction

- Success of overparameterization
- Trends in model size

- Correlation between increased complexity and increased accuracy

  → Trivial enlarging of models and hoping for better accuracy

- Extreme computation demands

  → Costs and concerns in sustainability

Introduction

- Literature in efficiency and redundancy studies in deep learning
  - → Focused on compression
    - Vector quantization, weight pruning, etc.

- Are all the components necessary? Are they serving a determined and effective purpose?
  - – Redundancy in the form of redundant components with the objective of distilling effective components from complex models

“plurality should not be posited without necessity”

William of Ockham’s principle of parsimony (Occam’s Razor)

Introduction

- In any design, including in deep learning
  - → Expected necessity and justification and reasoning behind the components
- Beneficial to make contribution of components clear and explain their role

- Enriching deep learning with theories and intuitions
  - → Real-world applications
  - → Further improvements of accuracy in downstream tasks

Motivational Factors

- Reproducibility and reliability
- Interpretability for intuitive and scientific design
- Sustainability
  - Environmental
  - Financial and Technical

Reproducibility

- Increase in popularity of deep learning
  - → Decrease in reliability and reproducibility of findings
- Acuteness of these concerns in deep learning compared to general computer science
  - – Hyperparameter settings, hardware, framework version, random seeds
- A relatively high focus in experimental findings and benchmark improvements in NLP
  - → Eminently prone to these concerns (amplified importance)
- Providing reproducibility details in research publications
  - → Preventing redundancy by channeling others' efforts in the correct directions

Interpretability for Design

- Use of augmented pipelines
  - Gluing multiple sophisticated components together

  → The interpretability issue beyond reliability:
    - Which component is responsible for improved outcomes?
    - How effective is a component of the pipeline for the downstream task?

- Such details are sometimes omitted from research papers

Interpretability for Design

- Increasing interpretability and suppressing the non-interpretable, black-box view to deep learning architectures
  - → Improvements in interaction, control, and trustworthiness
  - → Enabling the use of mathematical tools in a more intuitive and insightful way
    - → Both scoreboard improvements and societally-aware practices
- Interpretability-aware perspectives
  - → Promotion of scientific or (at least) intuitive design of the architectures
    - → Inherent redundancy avoidance

Interpretability for Design

- The intuitive rule in practicing interpretability in statistical learning

  → William of Ockham's principle of parsimony (Occam's Razor)

- For the purpose of interpretability

  → A smaller subset of predictors with the strongest effect is preferred

Sustainability

- Environmental: carbon emission and non-renewable energy consumption
  - Training a big Transformer model with hyperparameter search > 5 * A car's lifetime
  - Climate change

- Financial and technical: cost of acquiring hardware or using cloud computing
  - Inequity in access
  - Proprietary hardware (Google TPU)

Motivations

- Difficulty of NLP
  - “These tasks are so hard that Turing could rightly make fluent conversation in natural language the centerpiece of his test for intelligence.”
    - Page 248, Mathematical Linguistics, 2010
  - Unsuitability of shallow syntactic and semantic features for advanced language understanding tasks

Motivations

Reliability and interpretability concerns

Difficulty of natural language processing

It is reasonable NOT to expect outright success in NLP only based on the use of a neural network layer

Motivations

- Importance of publication of negative results about deep learning components
  - Controversial and not easy
  - Saves the time, efforts, and resources of others

- Ozan İrsoy, Adrian Benton, and Karl Stratos. 2021. Corrected CBOW Performs as well as Skip-gram. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 1–8, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Problem Statement

- Investigate the effect of a neural component called NeXtVLAD on predictive accuracy in two downstream NLP tasks
  - Context-dependent sarcasm detection
    - State-of-the-art sarcasm detection pipeline by Lee et al. (2020)
    - **F_1 score = 93.1%** (14% higher than the next best results; reported to FigLang2020 workshop)
  - Deepfake text detection
- Perform ablation experiments by removing NeXtVLAD from the architecture
  - Introduce a custom CNN architecture to extract *helping* features for NeXtVLAD and analyze its performance based on them

Publications

- Sina Mahdipour Saravani, Ritwik Banerjee, and Indrakshi Ray. 2021. **An Investigation into the Contribution of Locally Aggregated Descriptors to Figurative Language Identification**. In *Proceedings of the EMNLP Workshop on Insights from Negative Results in NLP*. ACL.

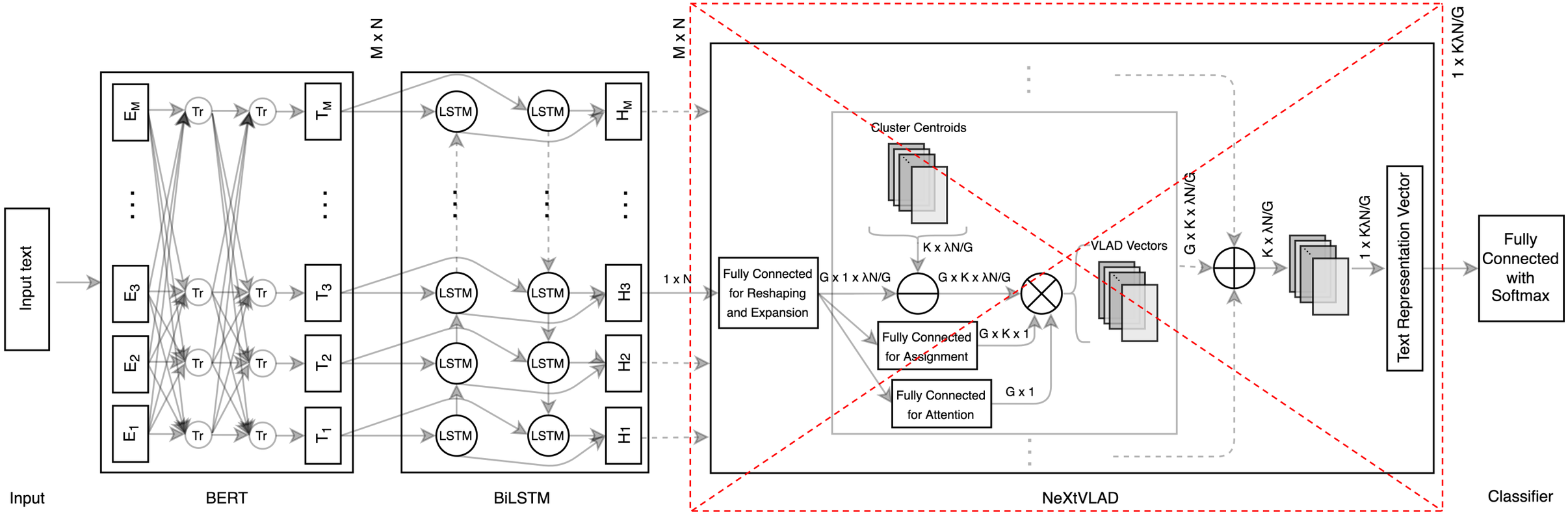
- Sina Mahdipour Saravani, Indrajit Ray, and Indrakshi Ray. 2021. **Automated Identification of Social Media Bots using Deepfake Text Detection**. In *Proceedings of the International Conference on Information Systems Security (ICISS)*. Springer.

NeXtVLAD, the Studied Architecture, and Prerequisites

Text Classification with Deep Learning

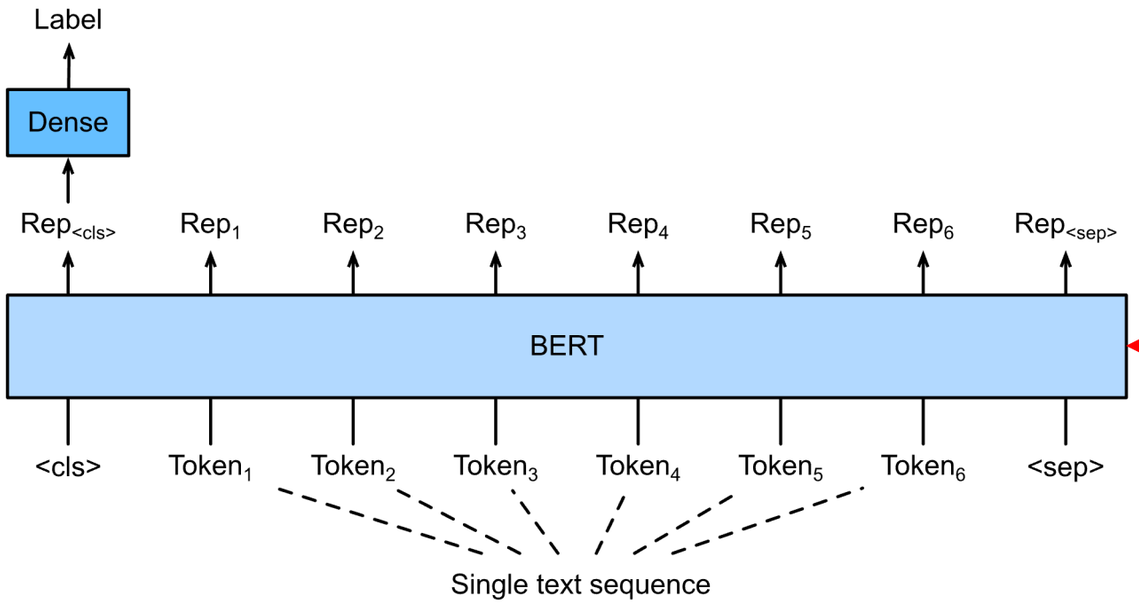
- Preprocessing
  - Tokenization, lemmatization, stop-word removal, noise removal, etc.
- Language Representation (Embedding), Feature Extraction, and Learning
  - BoW, word2vec, BERT
  - RNN, BiLSTM, CNN
- Pooling
  - Fully-connected, Max/Avg-pooling (summarizing, reducing dimension, removing variance)
- Classification
  - Fully-connected NN with softmax (assigning the final class probabilities)

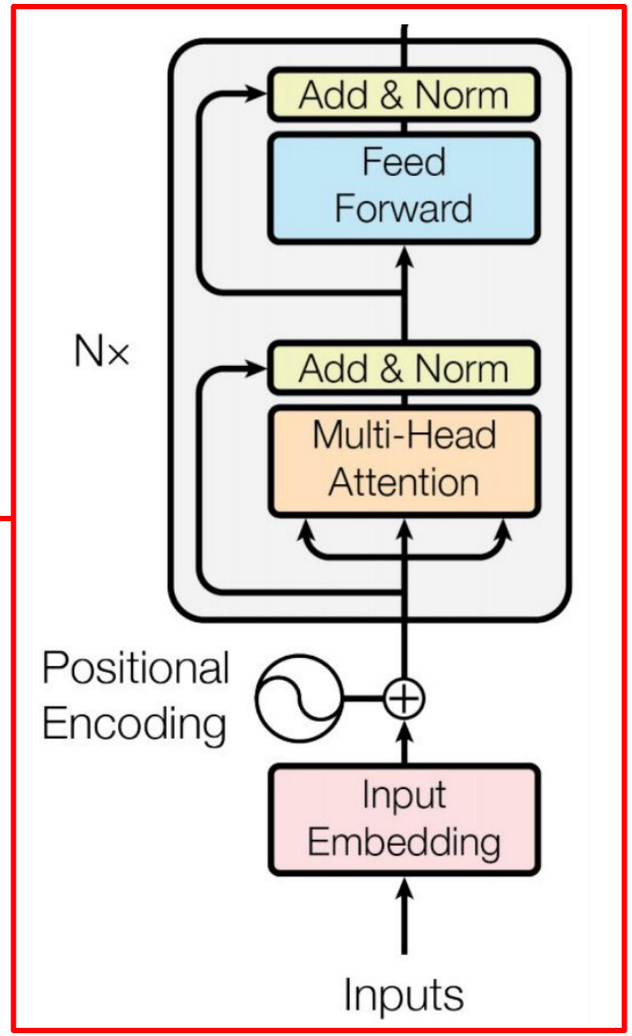
Architecture



BERT

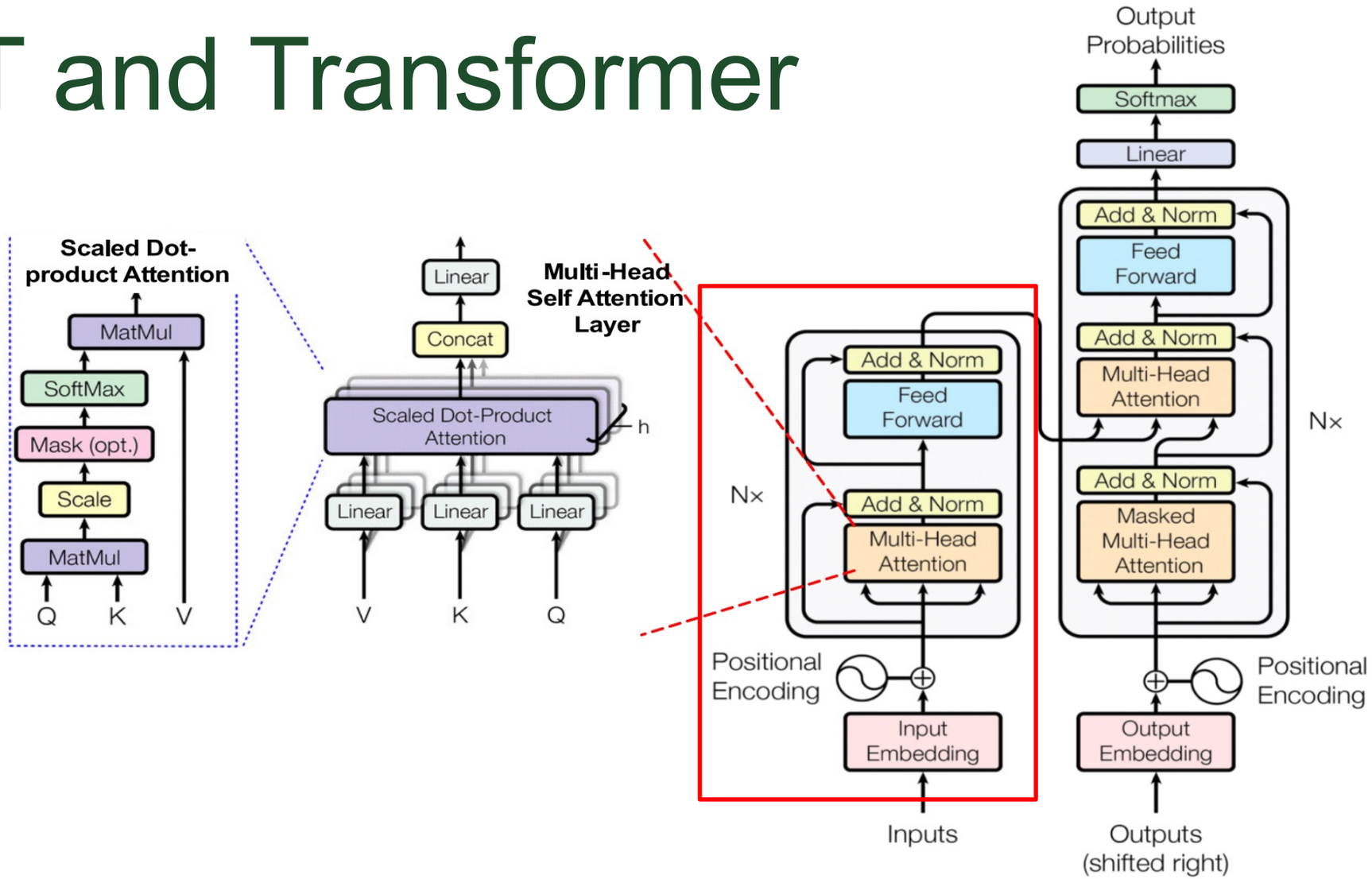
- Train for MLM and NSP
- Use multiple parallel self-attentions





https://d2l.ai/chapter_natural-language-processing-applications/finetuning-bert.html

BERT and Transformer



VP Singh, Harsh, and Qusay H. Mahmoud. 2020. NLP-Based Approach for Predicting HMI State Sequences Towards Monitoring Operator Situational Awareness. Sensors. no. 20(11).

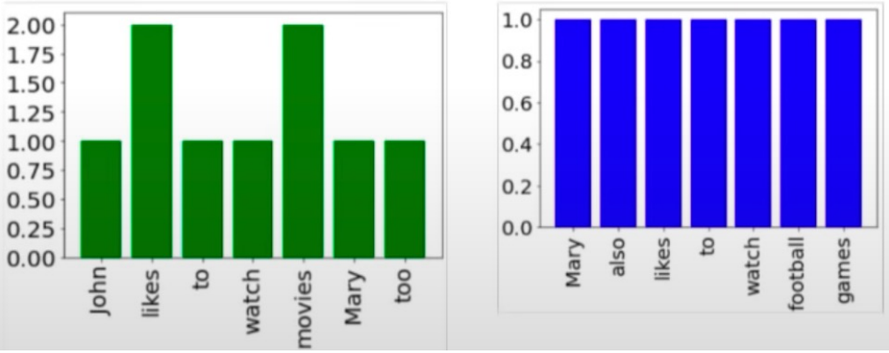
Bag of Visual Words

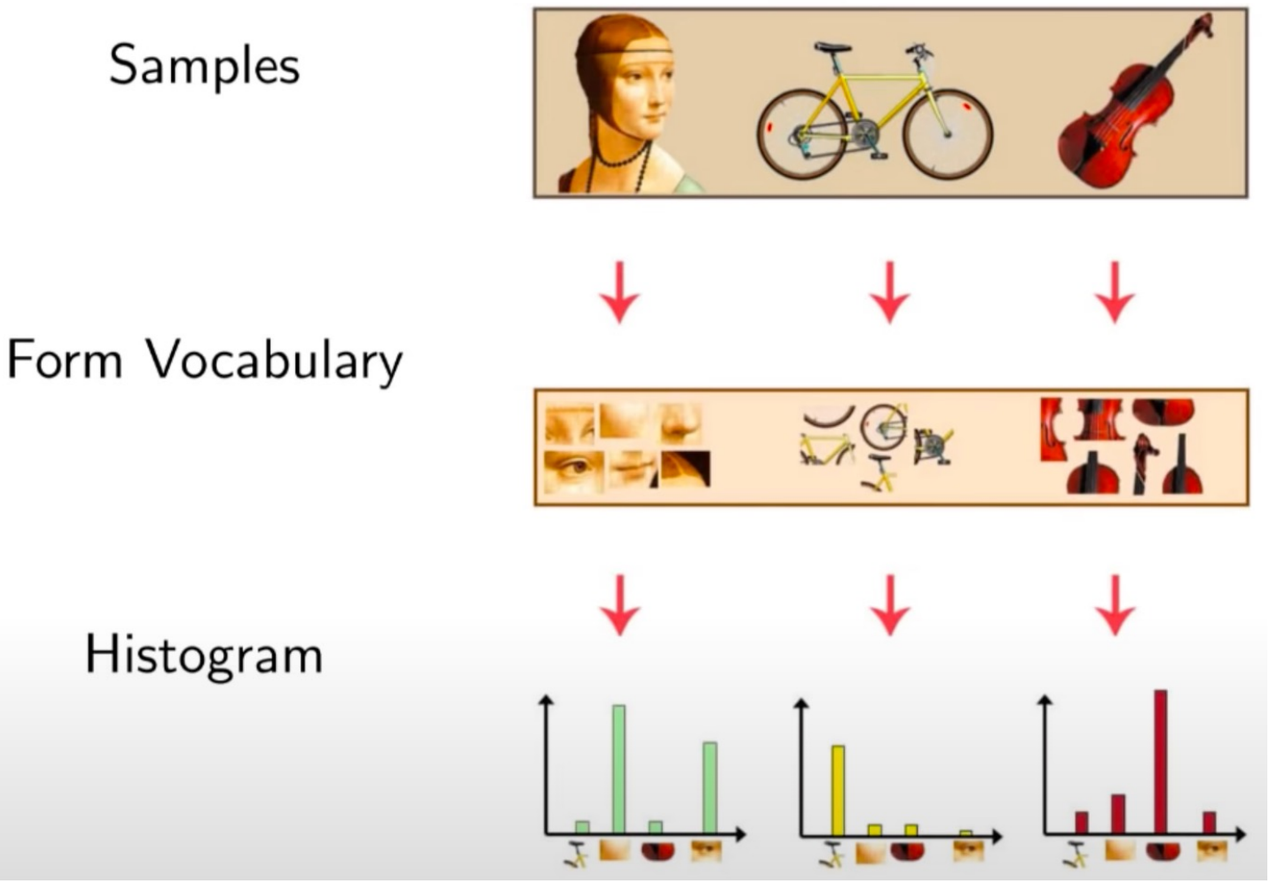
1) John likes to watch movies. Mary likes movies too.

2) Mary also likes to watch football games..

"John","likes","to","watch","movies","Mary","likes","movies","too"

"Mary","also","likes","to","watch","football","games"

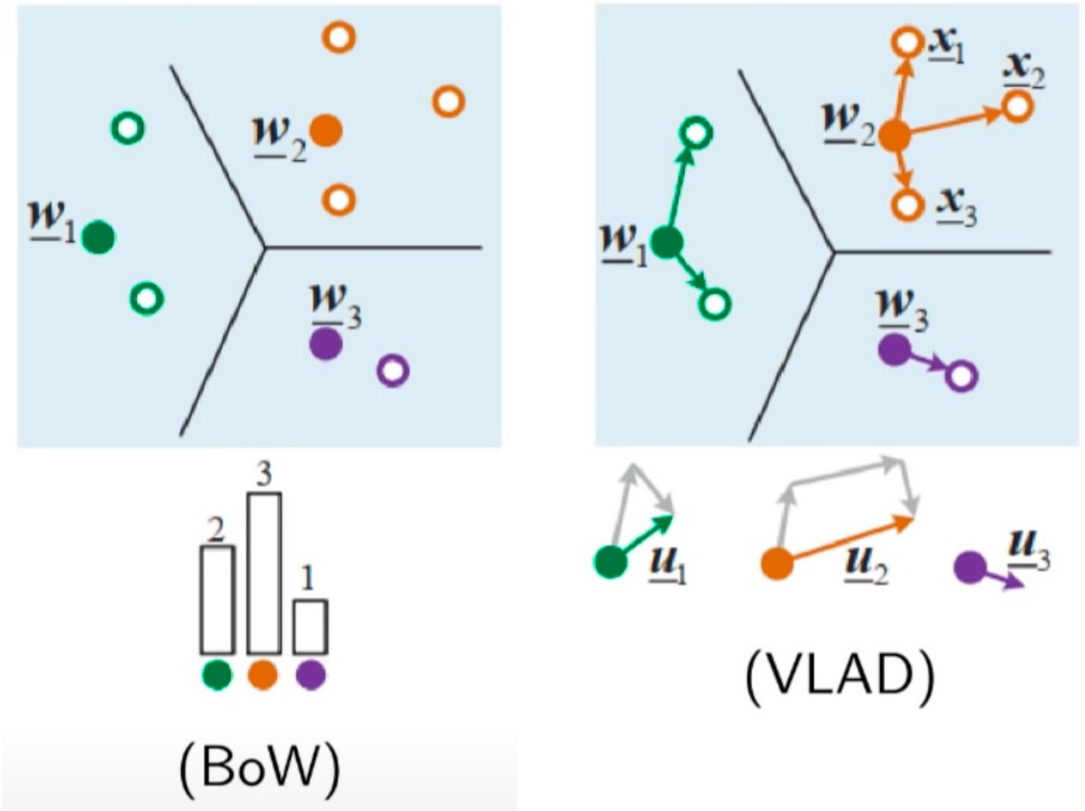




Vineeth N Balasubramanian, https://www.youtube.com/watch?v=CwJPEMcuAxY

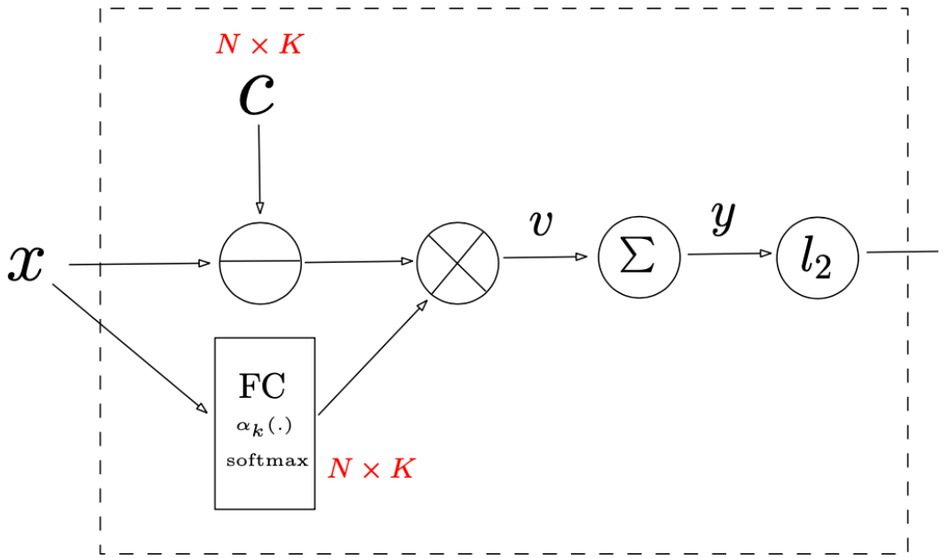
VLAD

- Vector of Locally Aggregated Descriptors
- Built on top of Bag of Visual Words
- Difference vector instead of presence frequency
  - Considering K clusters of all features

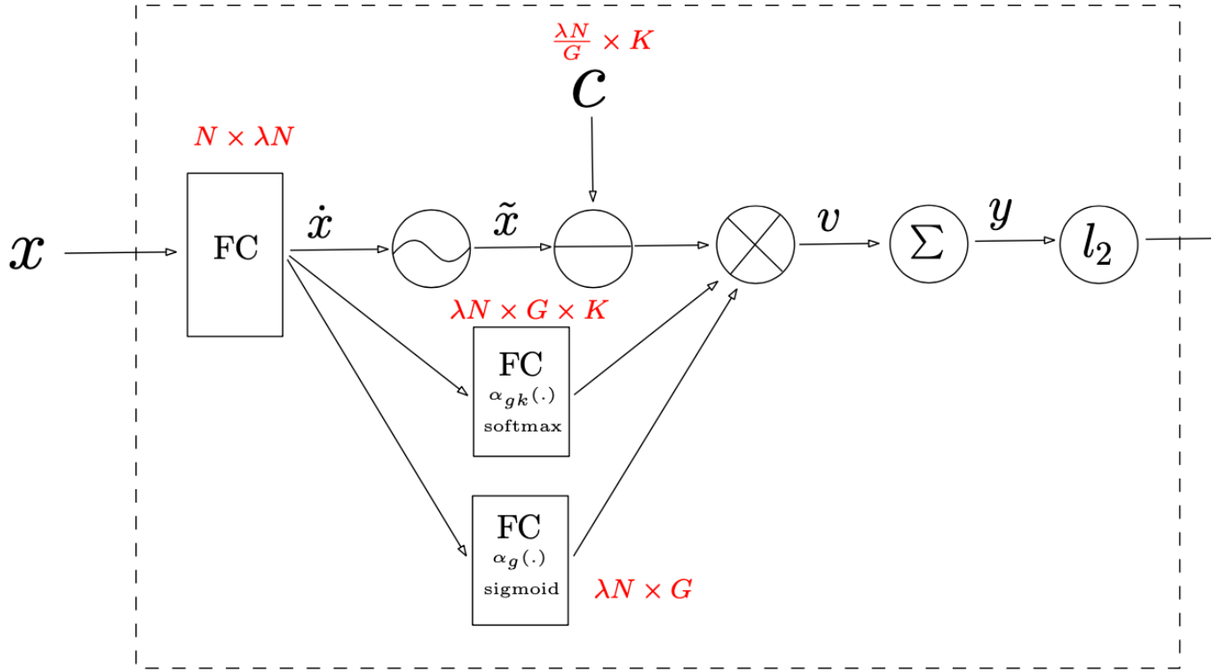


Vineeth N Balasubramanian, https://www.youtube.com/watch?v=CwJPEMcuAxY

NetVLAD and NeXtVLAD



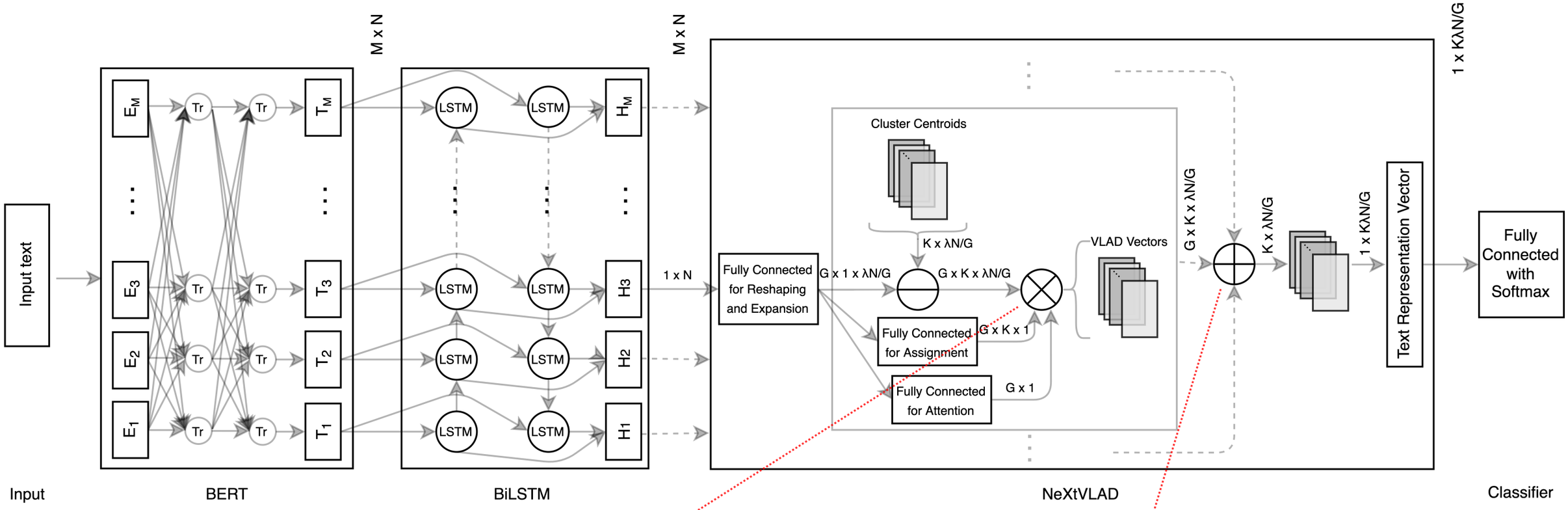
NetVLAD



NeXtVLAD

Lin, Rongcheng, Jing Xiao, and Jianping Fan. 2018. NeXtVLAD: An Efficient Neural Network to Aggregate Frame-level Features for Large-scale Video Classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 206-218. Springer.

NeXtVLAD in the Architecture



$$v_{tki}^{g} = \alpha_g(\dot{x}_t)\alpha_{gk}(\dot{x}_t)(\tilde{x}_{ti}^{g} - c_{ki})$$

$$r_{ki} = \sum_{t,g} v_{tki}^{g}$$

Sarcasm Detection

Task and Dataset

Sarcasm Detection

- Determine if the final response in a thread of tweets is sarcastic
- FigLang 2020 sarcasm detection dataset (4000 training &1000 validation & 1800 testing samples)

Turn	Tweet	Label
Context-1	The [govt] just confiscated a $180 million boat shipment of cocaine from drug traffickers.	Sarcastic
Context-2	People think 5 tonnes is not a load of cocaine.	
Response	Man! I've seen more than that on a Friday night.	

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A Report on the 2020 Sarcasm Detection Shared Task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 1—11. Association for Computational Linguistics.

Experiments

Sarcasm Detection

- Reproduction of the experiments by Lee et al. (2020):
  - Use of unpublished additional training data, hyperparameters, and validation set by Lee et al. (2020)

  → Exact reproduction is impossible

- To analyze the contribution of the NeXtVLAD component independently

  → Perform a comprehensive set of experiments with modifications in architecture, hyperparameters, and data

Experiments

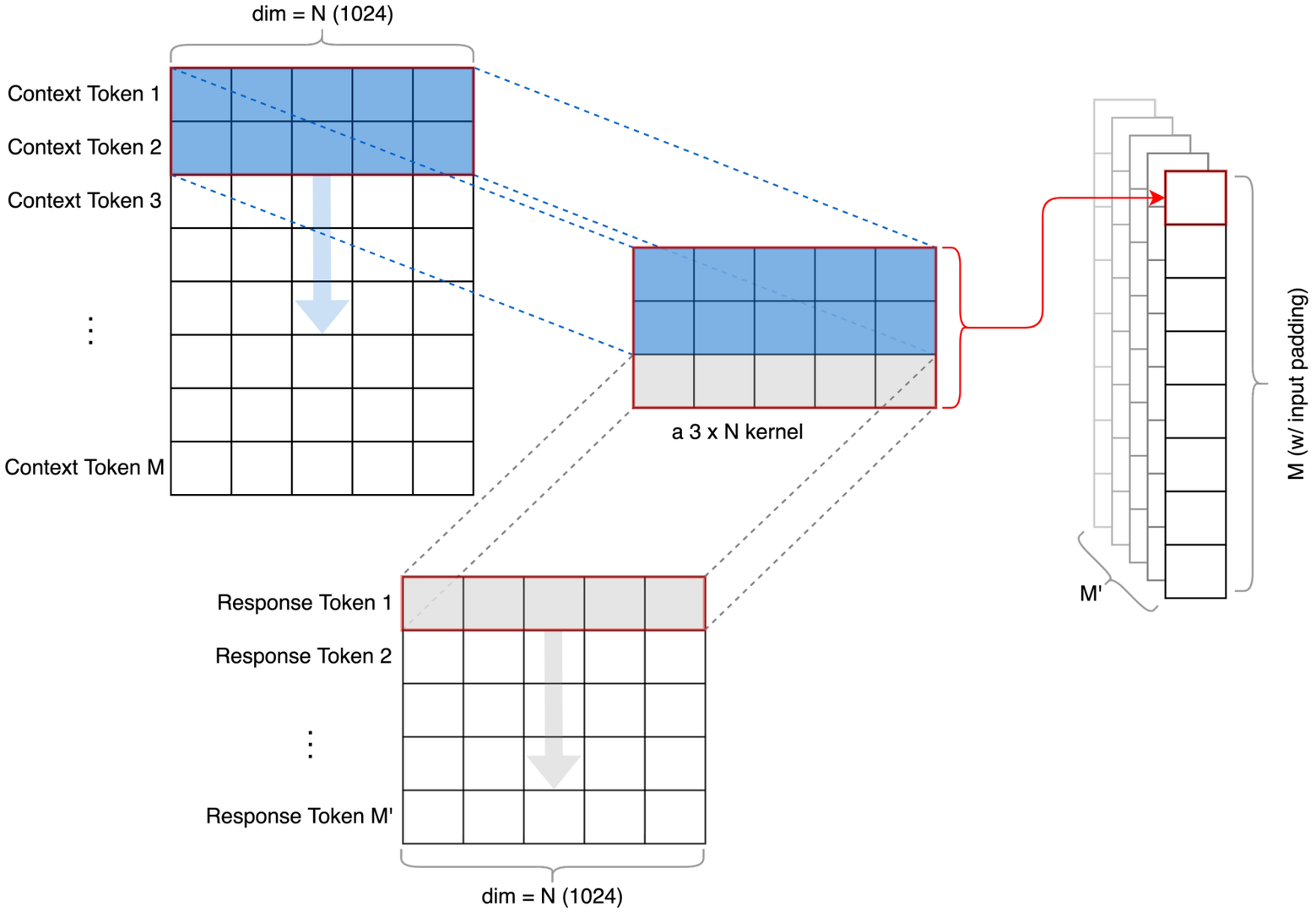
Sarcasm Detection

- Architecture modifications:
  - BERT (Base & Large), CTBERT(v1 & v2)
  - BiLSTM
  - NeXtVLAD
  - KimCNN
  - Our custom CNN
- Data modifications:
  - Labeled augmentation
  - Data expansion
- Hyperparameter modifications:
  - Num. training epochs
  - Linear and cyclic LR schedulers with various values
  - Batch size

Experiments

Sarcasm Detection

- To reduce the differences in shape and quantity of features fed to NeXtVLAD in Computer Vision and NLP
  - → Designed our custom CNN



Results

Sarcasm Detection

Model	Validation set results				Test set results			
	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1	Accuracy
$BERT_{Large\text{-}Cased}$	0.75	0.84	0.80	0.79	0.71	0.78	0.74	0.73
$BERT_{Large\text{-}Cased}$ + BiLSTM + NeXtVLAD	***0.74***	***0.84***	***0.79***	***0.78***	***0.71***	***0.77***	***0.74***	***0.72***
$BERT_{Large\text{-}Cased}$ + NeXtVLAD	0.71	0.82	0.76	0.74	0.69	0.77	0.73	0.71
$BERT_{Large\text{-}Cased}$ + BiLSTM	0.76	0.82	0.79	0.79	0.71	0.74	0.72	0.72
$BERT_{Large\text{-}Cased}$ + KimCNN + NeXtVLAD	0.74	0.84	0.79	0.78	0.72	0.82	0.77	0.75
$BERT_{Large\text{-}Cased}$ + OurCNN + NeXtVLAD	0.77	0.71	0.74	0.76	0.69	0.79	0.74	0.72
CTBERTv2	0.76	0.83	0.80	0.79	0.72	0.76	0.74	0.73
CTBERTv2 + BiLSTM + NeXtVLAD	0.72	0.85	0.78	0.77	0.71	0.79	0.75	0.73
$BERT_{Large\text{-}Cased}$ (DE)	0.81	0.85	0.83	0.82	0.72	0.73	0.73	0.72
$BERT_{Large\text{-}Cased}$ + BiLSTM + NeXtVLAD (DE)	0.79	0.84	0.82	0.81	0.73	0.74	0.74	0.73
$BERT_{Large\text{-}Uncased}$ (DE)	0.79	0.83	0.81	0.81	0.73	0.73	0.73	0.73
$BERT_{Large\text{-}Uncased}$ + BiLSTM + NeXtVLAD (DE)	0.79	0.87	0.82	0.82	0.73	0.79	0.76	0.75
CTBERTv2 (DE)	0.78	0.83	0.80	0.80	0.75	0.77	0.76	0.75
CTBERTv2 + BiLSTM + NeXtVLAD (DE)	0.81	0.83	0.82	0.82	0.77	0.77	0.77	0.77
$BERT_{Large\text{-}Cased}$ (DE, LA)	0.79	0.84	0.81	0.87	0.73	0.75	0.74	0.82
$BERT_{Large\text{-}Cased}$ + BiLSTM + NeXtVLAD (DE, LA)	0.67	0.60	0.63	0.77	0.63	0.52	0.57	0.74
Ensemble of 3 [CTBERTv2 + BiLSTM + NeXtVLAD] (DE)	0.62	0.61	0.61	0.62	0.60	0.54	0.57	0.59

Results

Sarcasm Detection

- Validation set results with and without NeXtVLAD in each epoch for the first model configuration (1st and 2nd rows):

Model	Accuracy for each epoch							
	1	2	3	4	5	6	7	8
w/o NeXtVLAD	0.69	0.73	0.77	0.78	0.78	0.78	0.78	0.79
w NeXtVLAD	0.51	0.51	0.49	0.49	0.76	0.77	0.77	0.78

Results

Sarcasm Detection

- No significant improvement accomplished by incorporation of NeXtVLAD in this sarcasm detection task
- The excellent F1 score of Lee et al. (2020) is
  - due to the natural language augmentation techniques
  - and NOT the novel architecture

- Based on all experiments:

  → Ablated version of the model w/o NeXtVLAD performs largely equally well

Deepfake Text Detection

Task and Dataset

Deepfake Text Detection

- Determine if a given Tweet is generated by a machine or human
- TweepFake deepfake Tweets dataset (20712 training & 2302 validation & 2558 testing samples)

Tweet text	Label
the world needs more whale stories. I would love to know what whalefacts are hiding in them.	GPT-2 Bot
I will make [FOLLOWERS OF A RELIGION] victims. They come into the United States but should have been crippled so I flourish. I can do it. @USERNAME #debate	RNN Bot
it literally what time of gucci shorts or not tolerate Libra slander on my face	Other Bot
I think if i put my mind to it, I could put a tree in my house like they do at the Cherry hill mall	Human

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. "TweepFake: About detecting deepfake tweets." Plos one 16, no. 5 (2021): e0251415.

Experiments

Deepfake Text Detection

- Architecture modifications:
  - BERT (Base & Large), CTBERT (v2)
  - XLNET (Base)
  - BERTweet
  - BiLSTM
  - NeXtVLAD
  - Average Pooling
  - Maximum Pooling
- Hyperparameter modifications:
  - Num. training epochs
  - Num. NeXtVLAD clusters
  - Learning rate

Results

Deepfake Text Detection

Configuration (Accuracy)	Model	Pre-Training	Pooling	num. of NeXtVLAD clusters	post-BiLSTM Operation
Cfg 1 (0.92)	T+Bi+NV+Cl	CTBERT-v2	NeXtVLAD	128	Addition
Cfg 2 (0.91)	T+Bi+NV+Cl	CTBERT-v2	NeXtVLAD	2	Addition
Cfg 3 (0.92)	T+Cl	CTBERT-v2	—	—	—
Cfg 4 (0.88)	T+Bi+NV+Cl	$\text{BERT}_{\text{Large-Cased}}$	NeXtVLAD	2	Addition
Cfg 5 (0.91)	T+Bi+AP+Cl	CTBERT-v2	Avg Pooling	—	Addition
Cfg 6 (0.91)	T+Bi+MP+Cl	CTBERT-v2	Max Pooling	—	Addition
Cfg 7 (0.91)	T+Bi+NV+Cl	CTBERT-v2	NeXtVLAD	128	Concatenation
Cfg 8 (0.87)	T+Bi+NV+Cl	$\text{XLNET}_{\text{Base-Cased}}$	NeXtVLAD	128	Addition
Cfg 9 (0.91)	T+Cl	BERTweet	—	—	—
Cfg 10 (0.91)	T+Bi+NV+Cl	BERTweet	NeXtVLAD	128	Addition

Results

Deepfake Text Detection

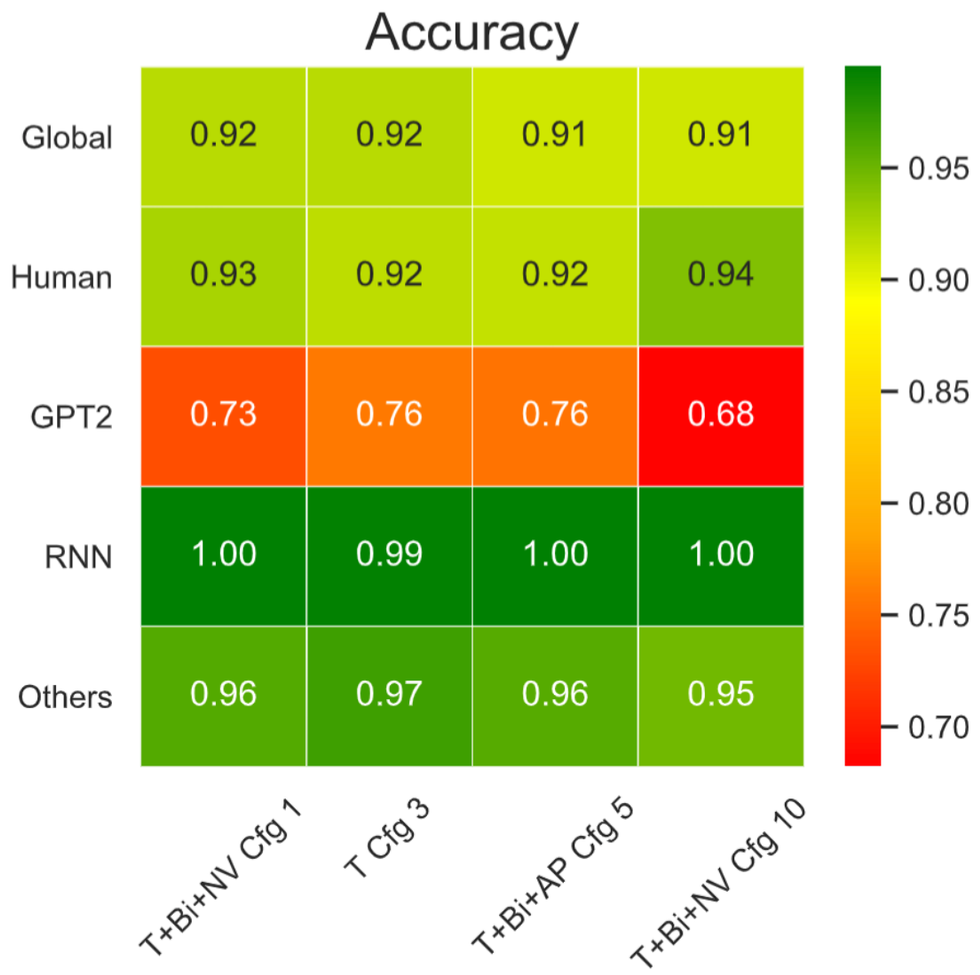
Model	Human			Bot			All
	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy
BERT (General-FT) [104]	0.90	0.88	0.89	0.88	0.90	0.89	0.89
RoBERTa (General-FT) [104]	0.90	0.89	0.90	0.89	0.90	0.90	0.90
LSTM on GloVe (twitter-glove-200)	0.84	0.81	0.82	0.81	0.85	0.83	0.83
BERT+BiLSTM+NeXtVLAD (Domain-FT) *Cfg 1*	0.92	0.91	0.92	0.92	0.92	0.92	0.92
BERT+BiLSTM+NeXtVLAD (Domain-FT) *Cfg 2*	0.92	0.90	0.91	0.91	0.92	0.91	0.91
BERT (Domain-FT) *Cfg 3*	0.91	0.92	0.92	0.92	0.91	0.92	0.92
BERT+BiLSTM+NeXtVLAD (General-FT) *Cfg 4*	0.90	0.87	0.88	0.87	0.90	0.88	0.88
BERT+BiLSTM+AvgPooling (Domain-FT) *Cfg 5*	0.91	0.92	0.91	0.92	0.91	0.91	0.91
BERT+BiLSTM+MaxPooling (Domain-FT) *Cfg 6*	0.91	0.91	0.91	0.91	0.91	0.91	0.91
BERT+BiLSTM+NeXtVLAD (Domain-FT) *Cfg 7*	0.92	0.91	0.91	0.91	0.92	0.91	0.91
XLNET+BiLSTM+NeXtVLAD (General-FT) *Cfg 8*	0.86	0.88	0.87	0.88	0.85	0.87	0.87
RoBERTa (Domain-FT) *Cfg 9*	0.90	0.94	0.92	0.93	0.89	0.91	0.91
RoBERTa+BiLSTM+NeXtVLAD (Domain-FT) *Cfg 10*	0.89	0.94	0.92	0.94	0.88	0.91	0.91
FastText's Supervised Classifier [57]	0.83	0.81	0.82	0.82	0.83	0.82	0.82
GROVER Discriminator (BS=1, MaxSeqLength=1024)	0.92	0.89	0.90	0.89	0.92	0.91	0.90
GROVER Discriminator (BS=32, MaxSeqLength=256)	0.92	0.90	0.91	0.91	0.92	0.91	0.91
GROVER Discriminator (BS=32, MaxSeqLength=256, DM)	0.91	0.90	0.90	0.90	0.91	0.91	0.91

Results

Deepfake Text Detection

- Comparing the detection accuracy over the type of the generator model

  → GPT2 is the most difficult to detect



Results

Deepfake Text Detection

- No significant improvement accomplished by incorporation of NeXtVLAD in this deepfake text detection task
- The 2 percent improvement is
  - due to the domain-specific pre-training
  - almost no difference between NeXtVLAD, MaxPooling, AvgPooling, and FC

- Based on all experiments:

  → Ablated version of the model w/o NeXtVLAD performs largely equally well

Conclusions

Conclusions

- Investigated the extent to which NeXtVLAD contributes to improved results

  → No improvement by incorporation of NeXtVLAD across the tasks

- NeXtVLAD's redundancy in NLP

  → No justification for its computational costs

Conclusions

- Local aggregators like NeXtVLAD
  → Unlikely to offer significant benefits to natural language processing when using Transformers

- Reinforcing the papers
  – “Attention is all you need” by Vaswani et al. (2017)
  – “Attention Is Indeed All You Need” by Juraska et al. (2021)
  – and many other papers solely based on Transformers

Discussions

- NeXtVLAD's success in Computer Vision vs. its failure in NLP
  - Difference in sub-vector representations
    - Low-dimensional split forming non-meaningful units in NLP
  - Interpretability of CV features compared to word embeddings

- This type of redundancy analysis

  → Provides *attribution* to specific components

  → Enables building comparable systems that are less resource-intensive

Recommendations

To mitigate the concerns and enhance the quality of research findings

1. In the design phase, provide explanations of intuitions and reasons for the design decisions
2. During the work, incrementally add components to a method or perform ablation studies
3. When reporting results, attribute the success or failure to the correct component

Recommendations

- Architectural studies, scientific design of methods, and attention the mentioned concerns
  - → Channeling researchers' efforts into aspects of a system with tangible and attributable benefits
  - → Mitigating deep learning's hunger for computation by redundancy reduction or novel interpretability-aware algorithms

References

- Hankyol Lee, Youngjae Yu, and Gunhee Kim. Augmenting Data for Sarcasm Detection with Unlabeled Conversation Context. In Proceedings of the Workshop on Figurative Language Processing, pages 12–17. Association for Computational Linguistics, 2020.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, pages 5998–6008, 2017.
- Martin Müller, Marcel Salathé, and Per E Kummervold. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. arXiv preprint arXiv:2005.07503, 2020.
- iziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. TweepFake: About detecting deepfake tweets. Plos one, 16(5):e0251415, 2021.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 13693–13696, 2020.
- Pierre Stock. Efficiency and redundancy in deep learning models: Theoretical considerations and practical applications. PhD thesis, École Normale Supérieure de Lyon, 2021.
- Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. The computational limits of deep learning. arXiv preprint arXiv:2007.05558, 2020.

Thank you

Colorado State University