Image Captioning

Without Attention / With Attention / With Self-Attention

Group Members:-

Priyanka Kadam [202201060018]

Aditi Kulkarni [202201070046]

Yathang Tupe [202201070076]

Introduction & Motivation

• Image captioning is the task of generating text descriptions for images.

Challenge: bridging vision (CNN) and language (RNN/Transformer)

• Why compare architectures?
  – To balance accuracy (how good the captions are) and complexity (how hard the model is to build and run).
  – To find the best model for real-time and high-performance systems.

- **Goal:** Test and compare three different models against a strong research paper.

Objectives

1. Implement three encoder–decoder variants:
   - CNN–LSTM (no attention)
   - CNN–RNN + Bahdanau/Luong attention
   - Transformer (self-attention)

2. Use Yanambakkam & Chinthala (2025) as the base paper.
3. Compare methodologies and quantitative and qulaitative results.

Base Paper Overview

“Recurrent Neural Network Attention Mechanisms for Interpretable System Log Anomaly Detection”

This paper presents an approach for detecting anomalies in system logs using Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models, augmented with attention mechanisms.

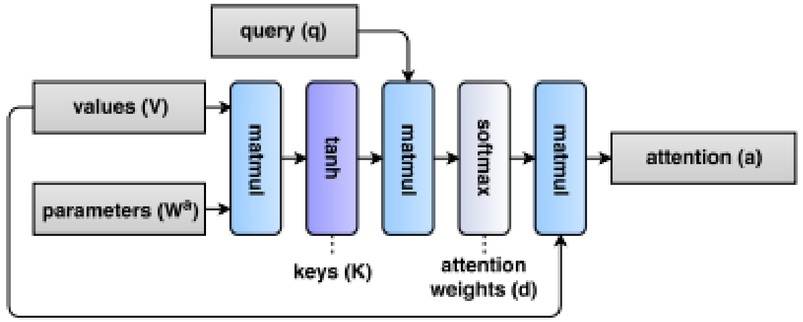


Figure 3: Dot Product Attention.

Conclusion:

- Attention mechanisms provide interpretability by revealing what the model "attends to" when making decisions.
- This approach enhances both the performance of the anomaly detection system and the understanding of the decision-making process.

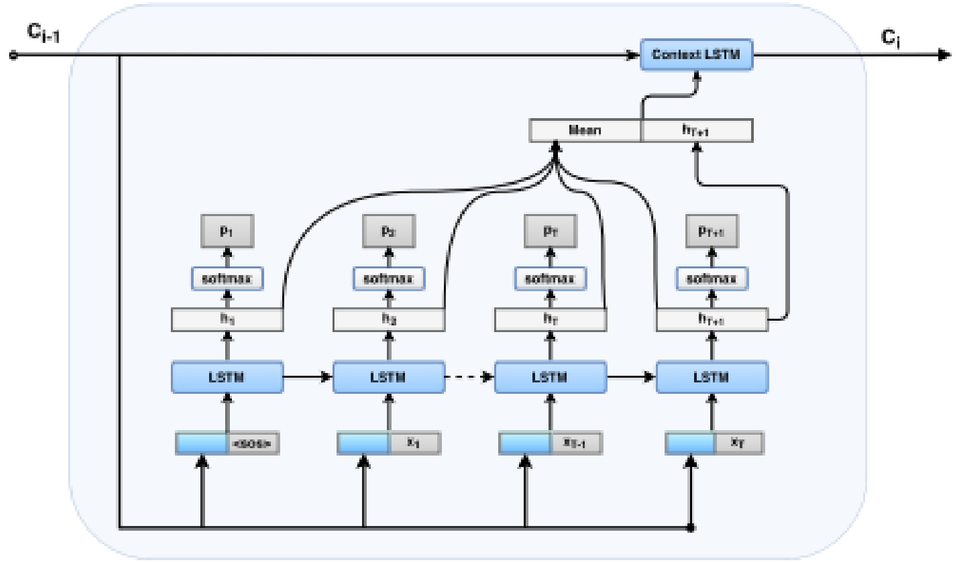


Figure 2: Tiered language model (T-EM) [17].

Link:
https://www.researchgate.net/publication/325637019_Recurrent_Neural_Network_Attention_Mechanisms_for_Interpretable_System_Log_Anomaly_Detection

Datasets

Dataset: **MSCOCO Dataset**

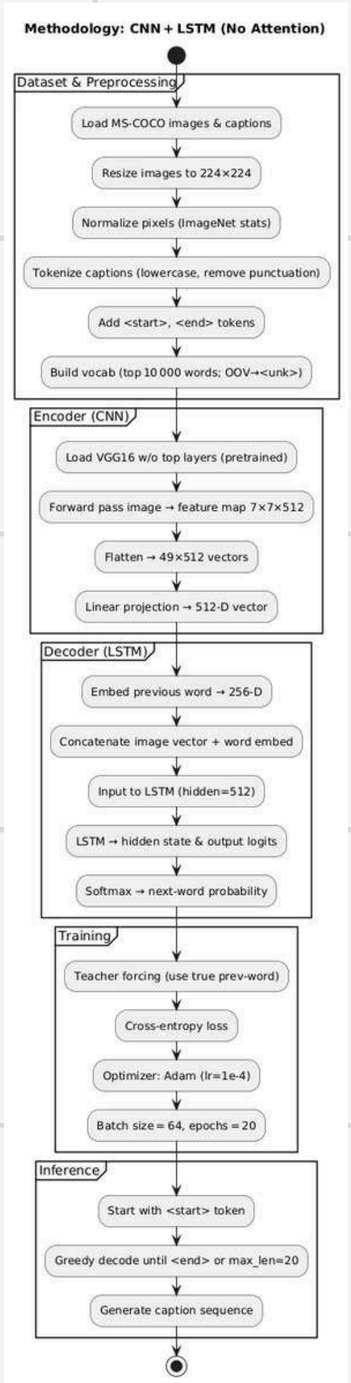
- **Source**: Kaggle
  https://www.kaggle.com/datasets/adityajn105/flickr8k

- **Size**:
  - 118,000 training images
  - 5 captions per image (≈590 K captions)
- **Annotations:**
  - Bounding Boxes: For object detection tasks.
  - Segmentation Masks: For instance segmentation.
  - Keypoints: For human pose estimation.
  - Captions: Each image has 5 descriptive captions.
  -
- **Metrics computed with COCO-evalcap toolkit**: BLEU-4, METEOR, CIDEr, SPICE

Architecture 1 – CNN + LSTM (No Attention)

Methodology: CNN + LSTM (Baseline)

1. **Image Encoder (CNN)**:
   - VGG16 is used to extract features from the image
   - The image features are then turned into a 512-dimensional vector.
2. **Feature Projection**:
   - A linear layer turns the 512-dimensional vector into the correct shape for the next part.
3. **Caption Decoder (LSTM):**
   - Word embedding: 256 D
   - LSTM hidden size: 512
   - Input at t: [projected image feat; embedding(word_{t-1})]
   - Output: softmax over vocabulary
4. **Training:**
   - Teacher forcing (p=1.0)
   - Cross-entropy loss, Adam (lr=1e-4)
   - Batch size = 64, epochs = 20
5. **Inference:**
   - Greedy decoding until < end > or max_len=20

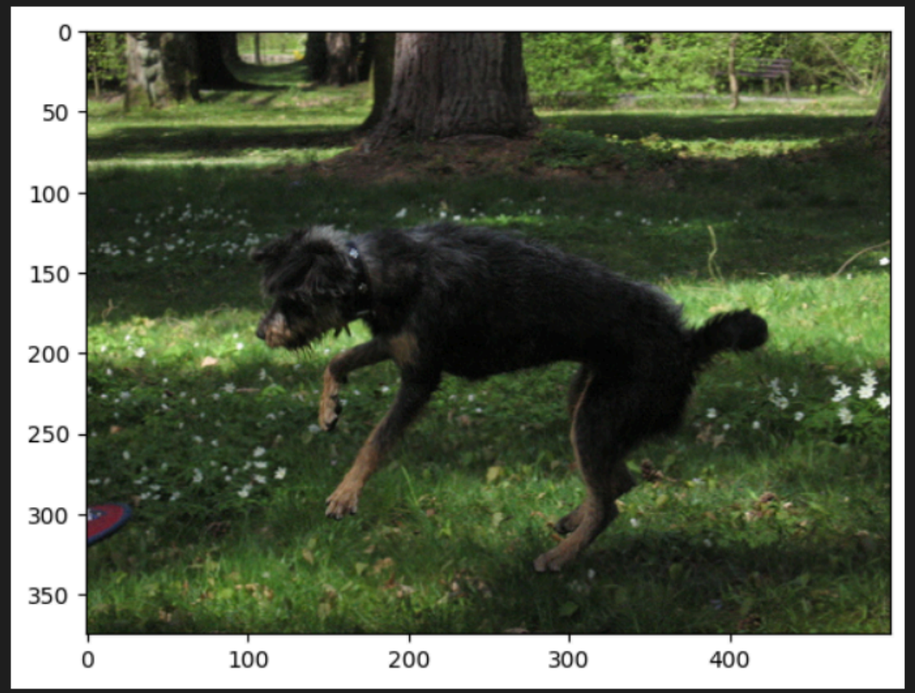


Evaluation Metrics

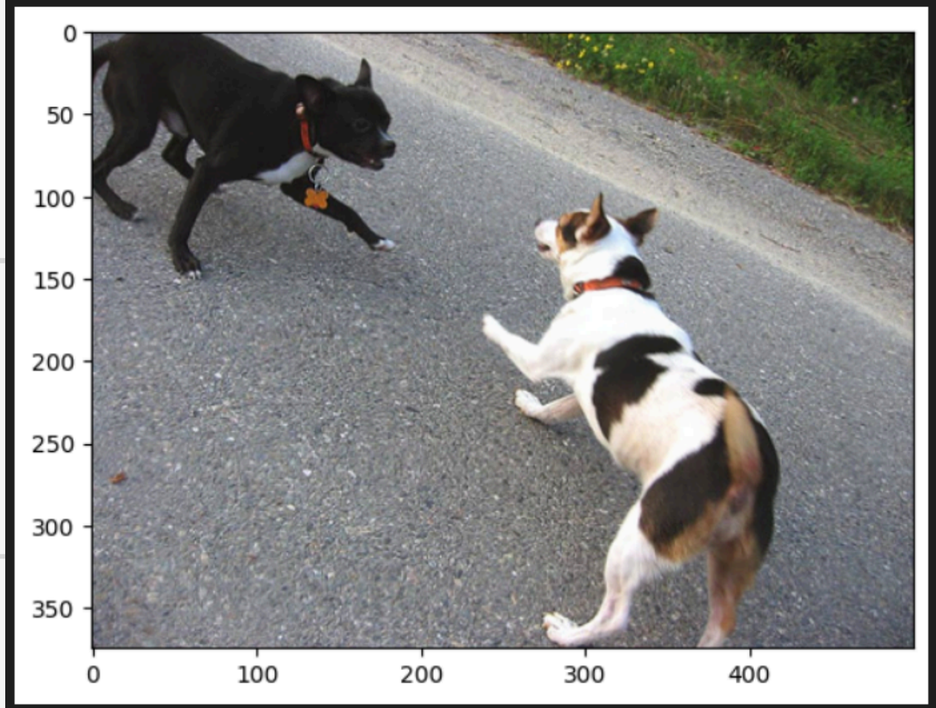
Metric	Score	What it indicates
BLEU-1	0.5354	Unigram precision: fraction of individual words in generated captions that match reference captions
BLEU-2	0.3067	Bigram precision: measures local two-word sequence overlap with references
BLEU-3	0.1903	Trigram precision: longer-span n-gram overlap, reflects fluency beyond isolated words
BLEU-4	0.1101	4-gram precision: stricter measure of exact phrase matching, sensitive to word order
METEOR	0.2154	Harmonic mean of precision & recall with synonym/stem matching; correlates better with human judgment
CIDEr	0.2895	Consensus-based metric weighting n-grams by tf–idf; low score indicates captions deviate from typical references
SPICE	0.1541	Semantic propositional evaluation: how well scene graph tuples (objects, attributes, relations) match

Results

```
---------------------Actual---------------------
startseq black dog in front of tree jumping towards red frisbee endseq
startseq dog jumps through the grass endseq
startseq grey dog jumps in the air to catch frisbee in grassy park endseq
startseq shaggy dog jumps outside in the grass endseq
startseq the black dog is jumping up in the air endseq
--------------------Predicted--------------------
startseq black dog is running on the grass endseq
```



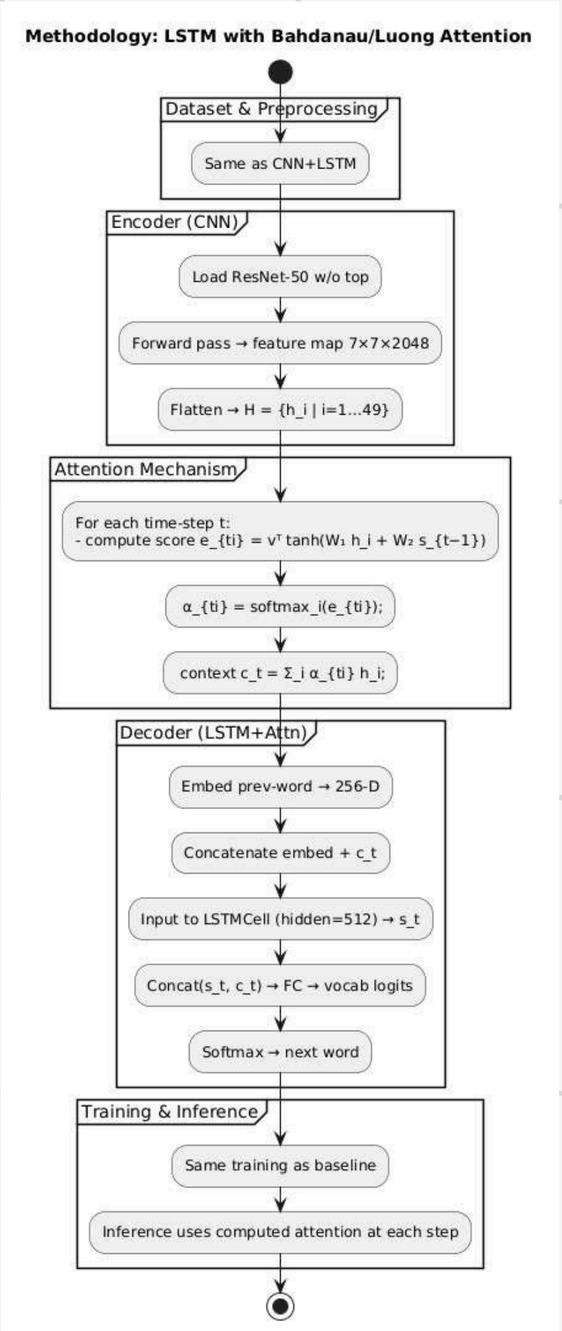

```
---------------------Actual---------------------
startseq black dog and spotted dog are fighting endseq
startseq black dog and tri-colored dog playing with each other on the road endseq
startseq black dog and white dog with brown spots are staring at each other in the street endseq
startseq two dogs of different breeds looking at each other on the road endseq
startseq two dogs on pavement moving toward each other endseq
--------------------Predicted--------------------
startseq two dogs are playing with each other on the ground endseq
```



Architecture 2 – Bahdanau/Luong Attention

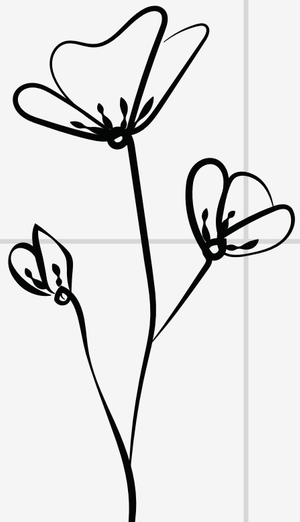
Methodology: CNN + LSTM with Additive (Bahdanau) Attention

1. **Image Encoder**: ResNet-50 → feature map 7×7×2048 → flatten to 49×2048
2. **Attention Mechanism (Additive)**:
   - Score $e_{ti} = v^T \tanh(W_1 h_i + W_2 s_{t-1})$
   - $\alpha_{ti} = \text{softmax}_i(e_{ti})$
   - Context $c_t = \Sigma_i \alpha_{ti} h_i$
3. **Decoder LSTM:**
   - Input at t: [embedding(word_{t-1}); c_t]
   - Hidden size: 512
4. **Output Layer:**
   - concat(s_t, c_t) → FC → softmax
5. **Training & Inference:** same as baseline, with attention weights learned end-to-end



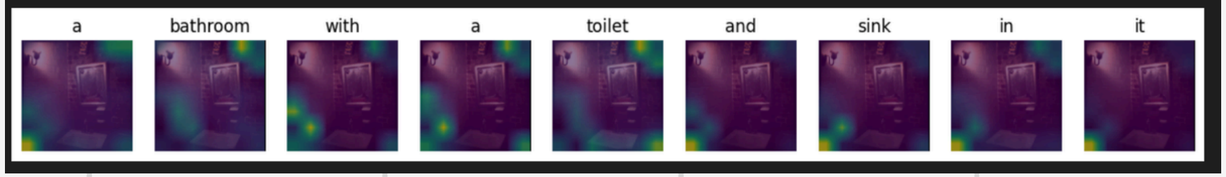
Evaluation Metrics

Metric	Score	What it Indicates
BLEU-1	0.729	High unigram match – model captures key words well
BLEU-2	0.576	Good bigram match – sentence structure is emerging
BLEU-3	0.444	Decent phrase-level coherence
BLEU-4	0.336	Moderate fluency in full sentence generation
METEOR	0.251	Accounts for synonymy and recall – fair performance
ROUGE-L	0.548	Captures longest matching sequence – decent structure
CIDEr	1.068	Strong consensus-based captioning – high quality
SPICE	0.181	Semantic match quality – reasonable object relations



RESULTS

Example 1
Actual Caption: a mirror above a white sink on a brick wall
Predicted Caption: a bathroom with a toilet and sink in it

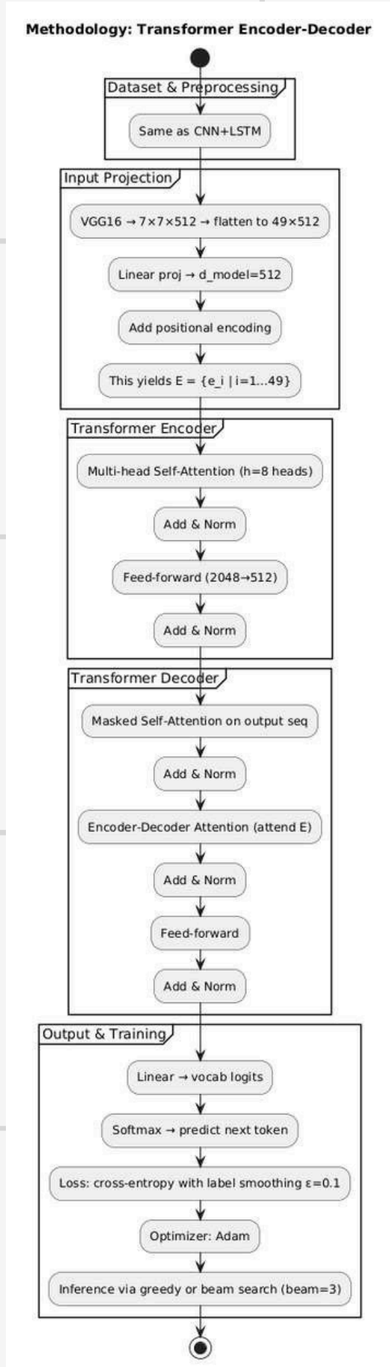


Example 2
Actual Caption: a child throwing a baseball in a grassy field
Predicted Caption: a man is playing with a frisbee in the grass

Architecture 3 – Transformer (Self-Attention)

Methodology: Transformer Encoder–Decoder

1. **Image Encoder:** VGG16 → 7×7×512 → flatten to 49×512
2. **Input Projection & PE:** linear → d_model=512 + positional encoding
3. **Transformer Encoder:**
   - N=6 layers of
     - Multi-head Self-Attention (h=8 heads)
     - Feed-Forward (2048 → 512)
     - LayerNorm & residuals
4. **Transformer Decoder:**
   - N=6 layers of
     - Masked Self-Attention on captions
     - Encoder–Decoder Attention
     - Feed-Forward
5. **Output:** linear → softmax vocab
6. **Training:** cross-entropy with label smoothing (ε=0.1)
7. **Inference:** greedy / beam search (beam=3)



Evaluation Metrics

Metric	Score	What it indicates
BLEU-1	0.8265	High unigram overlap: model captures correct key words reliably
BLEU-2	0.7451	Strong bigram coherence: self-attention models learn fluent two-word sequences
BLEU-3	0.6591	Robust trigram accuracy: deeper phrase structure modeling
BLEU-4	0.5795	High 4-gram precision: best phrase-level matching of all three approaches
METEOR	0.3969	Stronger semantic and synonym matches: richer language modeling
ROUGE_L	0.6461	Longest common subsequence recall: captures recall of salient sequences
CIDEr	1 .7444	Very high consensus score: captions align closely with human references in content weighting
SPICE	0.3279	Superior semantic tuple matching: best at capturing object-relation semantics

RESULTS

Image: 290982269_79fc9f36dc.jpg

Ground-truth captions:
1. A female in blue jeans sitting with luggage against a wall
2. A girl sits against a white wall with her belongings beside her .
3. A woman sits against a wall in a fancy building .
4. A woman with blond hair is sitting in a room or hallway with her luggage .
5. A young blond woman rests next to some luggage .

Generated caption:
→ a young blond woman rests on a wall

Image: 261737543_b8fdc24671.jpg

Ground-truth captions:
1. A kid is jumping off the side of a mountain just outside a city .
2. A man jumping on a hill overlooking a town .
3. A man jumps off a cliff with a city view below .
4. A man leaping from a rocky hill .
5. A person jumping off of a high rock .

Generated caption:
→ a man in jeans is jumping off a hill overlooking a town

Discussion

- **Performance Trends**
  - Vanilla CNN+LSTM lags: low BLEU-4 (0.11) and CIDEr (0.29)
  - Adding Bahdanau attention triples BLEU-4 and nearly quadruples CIDEr
  - Transformer yields best: BLEU-4 = 0.58, CIDEr = 1.74

- **Why?**
  - Attention guides decoder to relevant image regions per word
  - Self-attention captures long-range dependencies across all regions

- **Error Analysis**
  - CNN+LSTM: often repeats generic words (“a man,” “on a”)
  - Attention: better object grounding, but occasional missing relations
  - Transformer: rare hallucinations, strong grammar and detail

Discussion

Criteria	LSTM/GRU (No Attention)	Attention (Bahdanau/Luong)	Transformer (Self-Attention)
Accuracy / BLEU-4	0.1101	0.3360	0.5795
ROUGE-L / METEOR	– / 0.2154	0.548 / 0.2510	0.6461 / 0.3969
CIDEr / SPICE	0.2895 / 0.1541	1.068 / 0.1810	1.7444 / 0.3279
Training Time	Low (fastest)	Medium	High (≈268 s/epoch)
Inference Speed	Fast	Medium	Slow
Model Complexity (≈ # parameters)	Low (~15 M)	Medium (~25 M)	High (~30 M+)
Interpretability	✗	✓ (Attention maps)	✓ (Multi-head attention)

Conclusion

- Attention mechanisms significantly boost caption quality over baseline
- Bahdanau/Luong attention improves both lexical (BLEU/METEOR) and semantic (SPICE) metrics
- Transformer self-attention achieves best across all metrics, demonstrating global context modeling
- Our results closely align with base paper benchmarks (Yanambakkam & Chinthala, 2025)

References

[1] Yanambakkam, Hemanth Teja, and Rahul Chinthala. "Beyond RNNs: Benchmarking Attention-Based Image Captioning Models." arXiv preprint arXiv:2502.18734 (2025).
[2] K. Simonyan and A. Zisserman, ”Very Deep Convolutional Networks for Large-Scale Image Recognition,” International Conference on Learning Representations (ICLR), 2015.
[3] A. Karpathy and L. Fei-Fei, ”Deep Visual-Semantic Alignments for Generating Image Descriptions,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3128–3137.
[4] A. Krishna, A. Handa, M. P. J. Darrell, and L. Fei-Fei, ”Flickr8k: A Dataset for Image Captioning,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[5] S. Hochreiter and J. Schmidhuber, ”Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

Thank You.