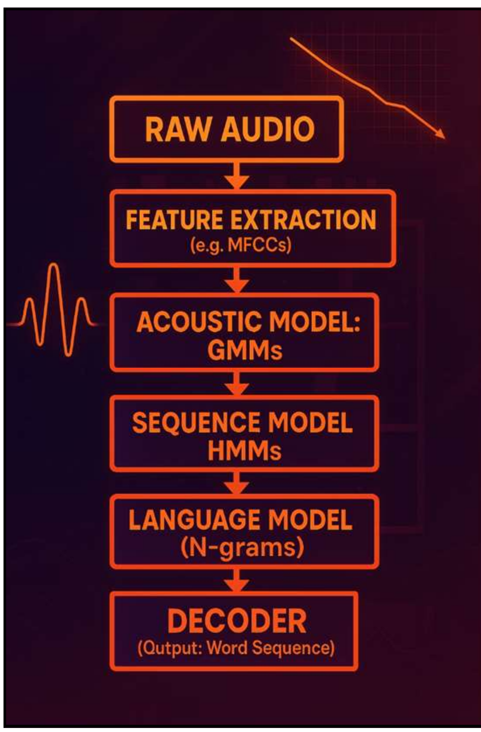
Advanced Speech Recognition: Leveraging Deep Learning with Segmental CRFs

Fusing Powerful Features for Accurate ASR

by Harsh Bajpai

1



The Old Guard: GMM-HMM Systems

Acoustic Modeling

Gaussian Mixture Models (GMMs) [Sound Likelihoods per State]

Sequence Modeling

Hidden Markov Models (HMMs) [State Sequence / modal temporal sequence]

Limitations

Performance ceiling, limited adaptability

Example System

CU-HTK Broadcast News recognizer (Gales et al., 2006)

2

Deep Acoustics: DNNs Learning Rich Sound Features

Hierarchical Learning

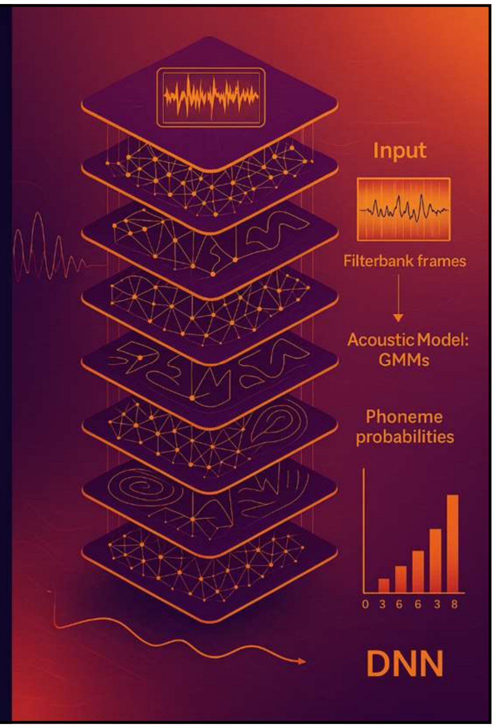
Features learned directly from from data

Performance

Outperforms traditional GMM GMM acoustic models

Reference

Hinton et al. (2012)



3

Deep Language: NNLMs Capturing Context

RNNs

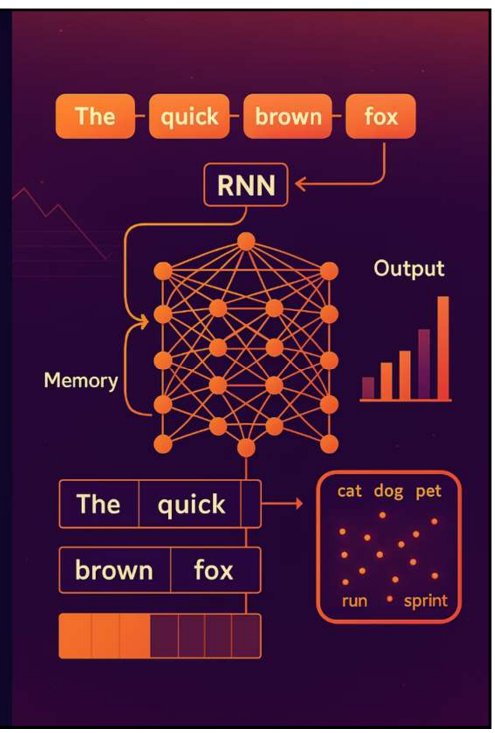
Memory for long word word dependencies

Word Embeddings

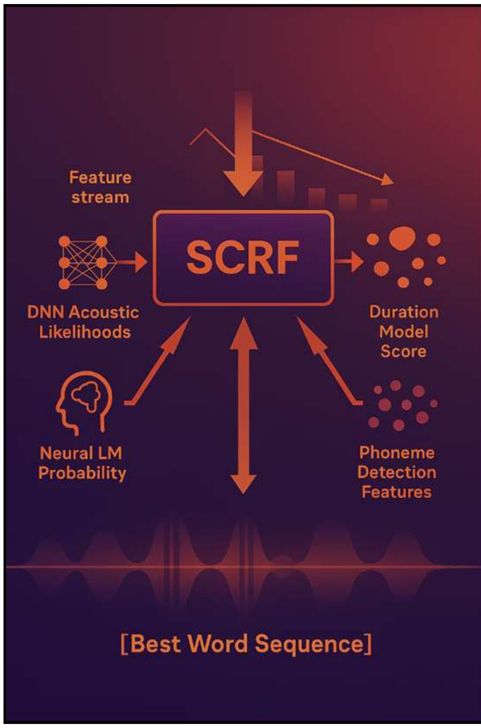
Semantic relationships relationships in vector vector space

Reference

Mikolov et al. (2011)



4



Segmental CRFs: Principled Feature Fusion

Feature Fusion

Combines diverse evidence at at word-segment level

Discriminative Learning

Lears Optimal Weights Discriminatively

Reference

SCARF Toolkit (Zweig & Nguyen, 2010)

5

JHU Workshop: SCRFs + DL Deliver SOTA Results

Feature Integration

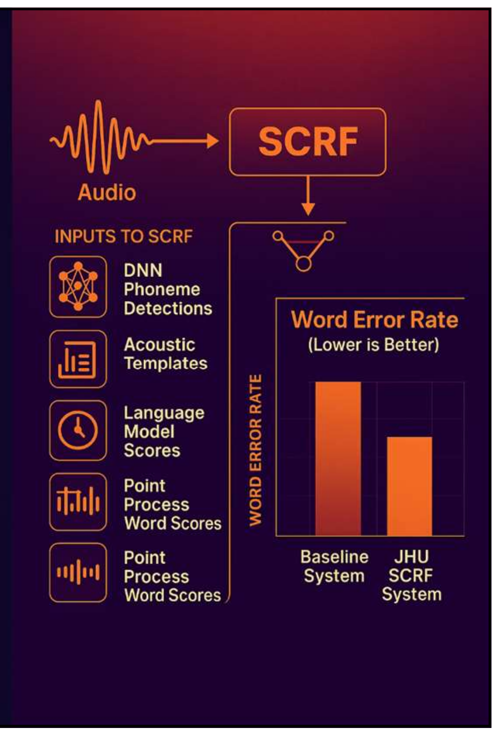
DNN Phoneme Features, Templates, LM, Duration

Results

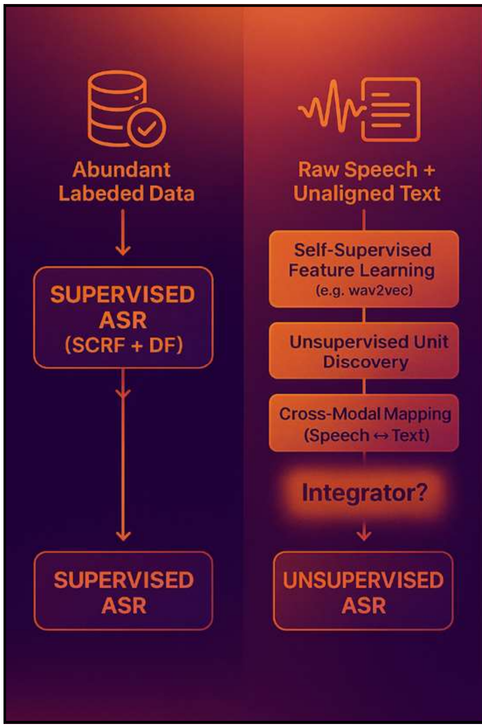
Significant ASR accuracy gains on on BN & WSJ

Reference

Zweig et al. (2011), Jansen & Niyogi (2009)



6



Beyond Labeled Data: Unsupervised ASR Challenges

Unsupervised ASR

Discovering Acoustic Acoustic Units & Word Word Segments from from Raw Speech

SCRF Role

SCRF Principles (Feature Integration) Remain Relevant

Reference

Aldarmaki et al. (2022) (2022) Review

7

The Winning Combination: Deep Learning Power + SCRF Integration

Deep Learning

Rich, learned features for acoustics & language

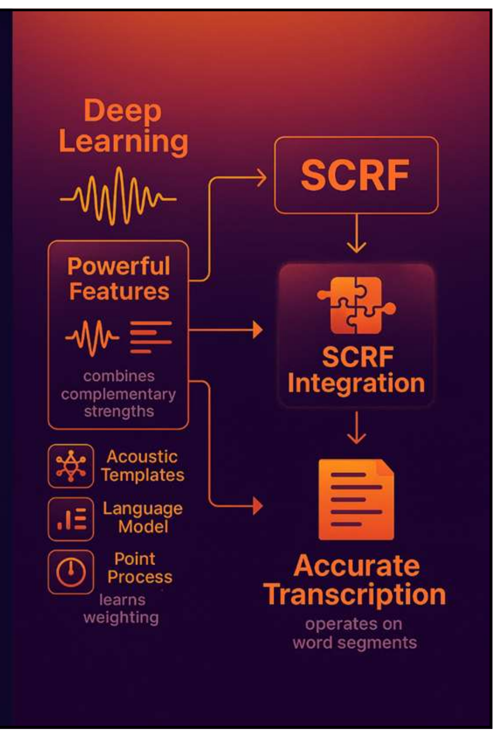
SCRFs

Flexible, principled framework framework for fusing these diverse, segment-level features. features.

Result

More accurate and robust speech recognition

Thank You!



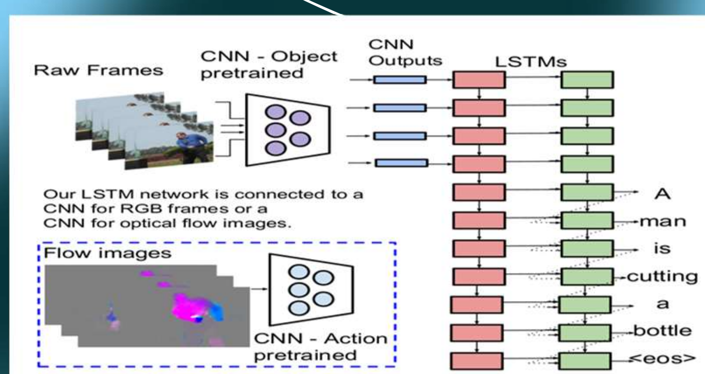
8

Video Captioning

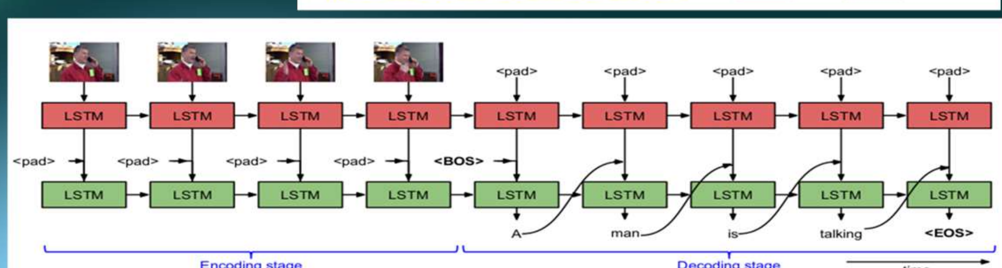
9

9

S2VT

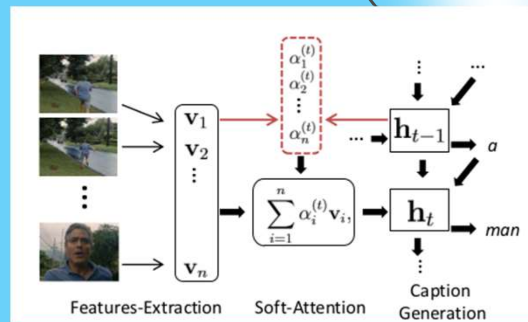


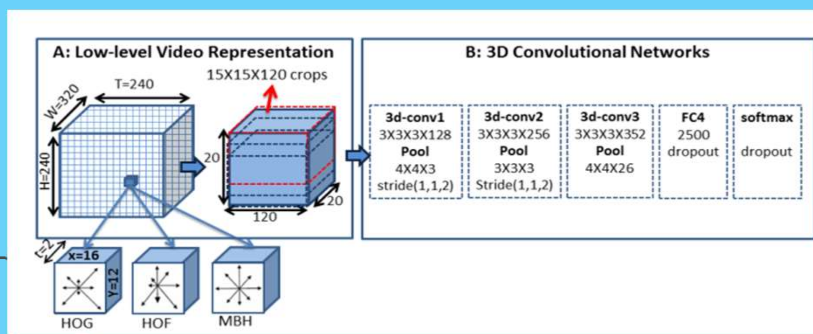
$$\alpha \cdot p_{rgb}(y_t = y') + (1-\alpha) \cdot p_{flow}(y_t = y')$$



10

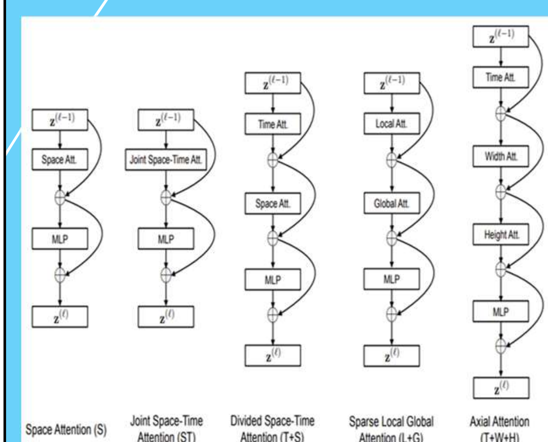
Temporal Attention

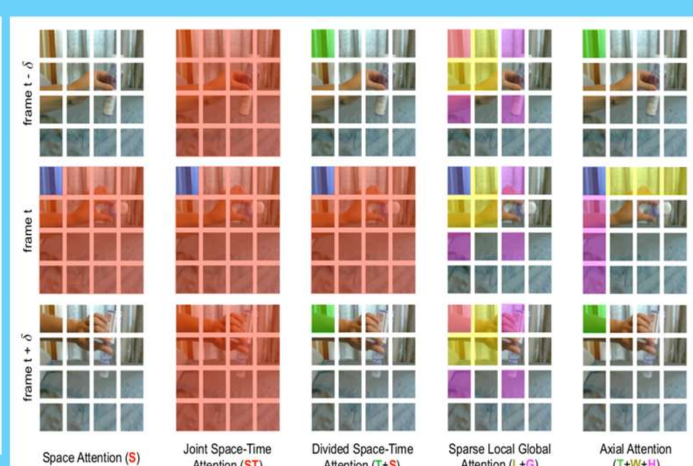




11

TimeSformer



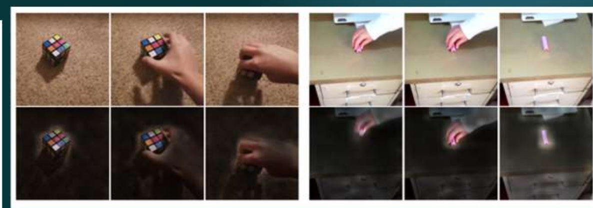


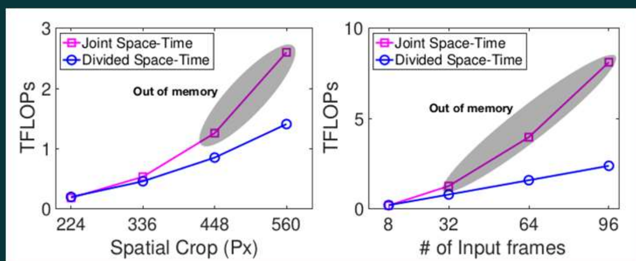
12

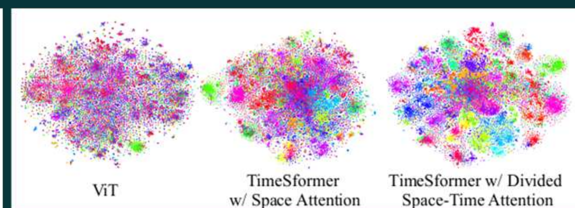
12

Interesting Ablations

Attention	Params	K400	SSv2
Space	85.9M	76.9	36.6
Joint Space-Time	85.9M	77.4	58.5
Divided Space-Time	121.4M	**78.0**	**59.5**
Sparse Local Global	121.4M	75.9	56.3
Axial	156.8M	73.5	56.2



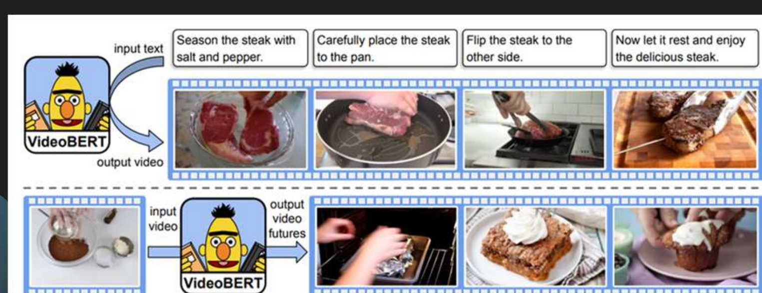


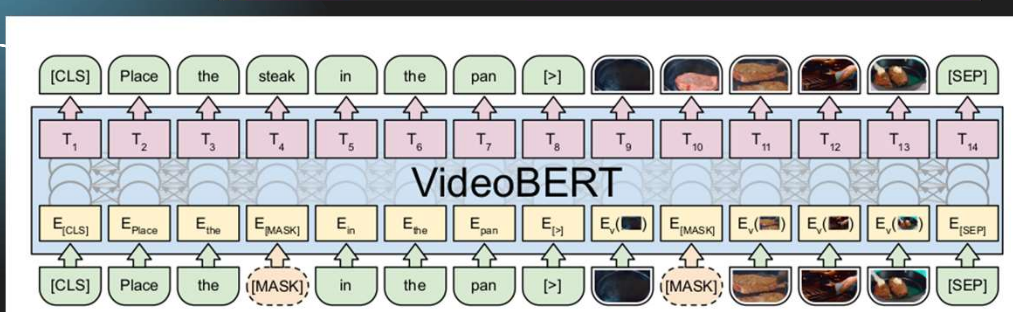


13

13

VideoBert

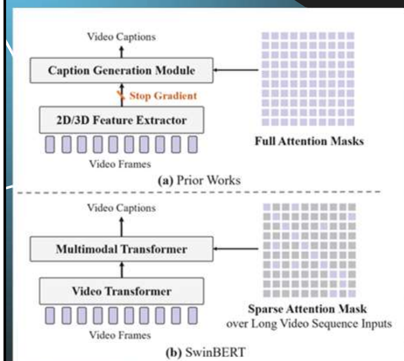


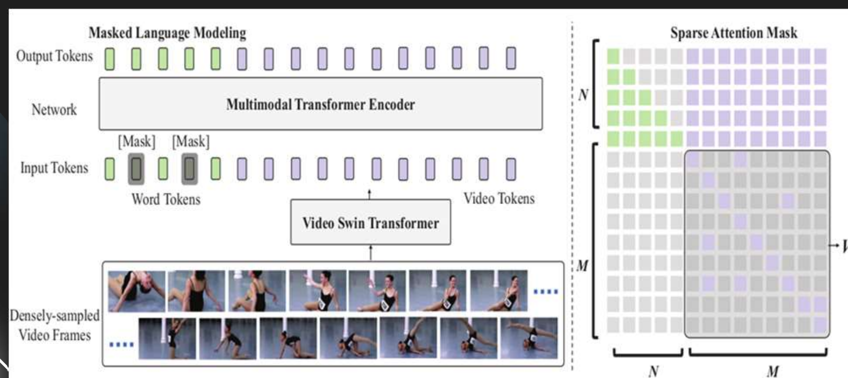


14

14

SwinBert



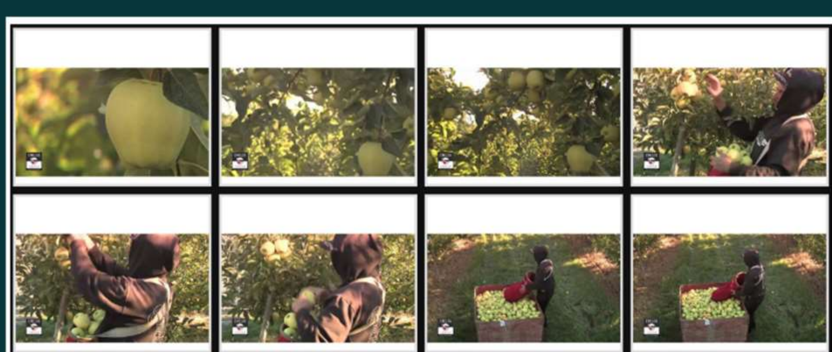


15

15

Future Directions

- **Audio**
- **Better evaluation metrics**
- **Reasoning**
- **Quality data**



Close-up of a single yellow apple on a tree, followed by a broader view of several apples on branches. A worker in a black hoodie picks apples, placing them into a red basket and later empties the basket into a large wooden crate.

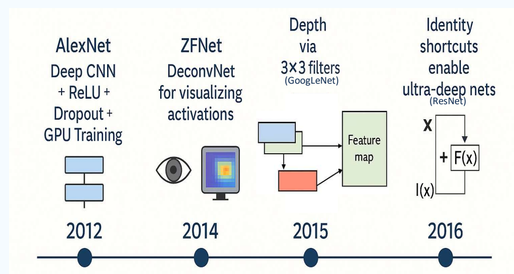
16

16

The Evolution of Convolutional Neural Networks for Image Classification:

Architectural Innovations and Performance Breakthroughs (2012–2016)

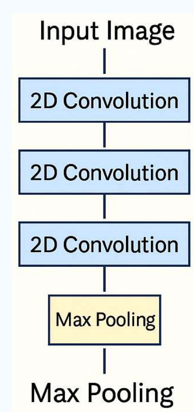
Sai Saketh Bavisetti



17

01 AlexNet: The First Breakthrough CNN

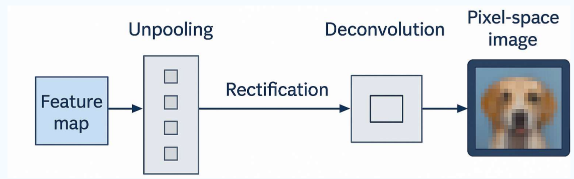
- Architecture: (5 Convolutional + 3 Fully Connected layers)
  – AlexNet was 8 layers deep, which was unprecedented at the time.
  – Each convolutional layer was followed by max pooling to reduce spatial size and extract features.
  – Fully connected layers processed these features and made predictions across 1,000 categories.
- ReLU: $f(x)$=max(0,x) → 6× faster convergence compared to sigmoid or tan h activations.
- Dropout: ρ= 0.5 in FC layers; during training 50% of neurons were randomly turned off(overfitted)
- Dual-GPU training: 90 epochs on 1.2 M images
  – The model was trained across two *NVIDIA GTX 580* GPUs in parallel.
  – Each GPU handled a separate branch of the model;
  – Which made it feasible to train such a large network on the available hardware.
- The training images augmented using translations and mirrored copies. This helped reduce overfitting and improved the model's ability to generalize to new data.
- Local Response Normalization: normalizing each pixel in relation to its neighbors. This technique was later abandoned by future models but contributed slightly to performance at the time.
- Top-5 error dropped down from 26.2% (previous record) → 15.3% (ILSVRC)
- This result stunned the computer vision community and shifted focus toward deep learning.
- Every CNN afterward built on its design principles: depth, ReLU, dropout, GPU training.



18

02 ZFNet: Understanding What CNN See

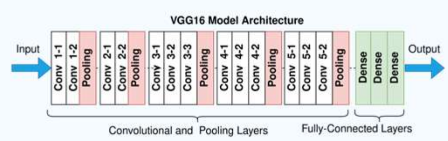
- Deconvolution Network (DeconvNet): ZFNet introduced a way to project activation maps from deeper layers back to input pixel space using:
  – Un-pooling (reverse max pooling),
  – Rectification (ReLU), and
  – Deconvolution (transpose convolution)



- Layer-wise Visualization of Filters:
  – Early layers were shown to detect edges and colors.
  – Middle layers responded to textures and motifs.
  – Higher learned to detect complex object parts like dog faces or bicycle wheels.
- Selectively zeroed out individual layers during testing and measured the accuracy drop.
- Guided refinement:
  – By analyzing visualizations, ZFNet redesigned some convolution filter sizes (e.g., reducing first layer strides from 4 to 2).
  – These changes led to a better-performing model with fewer artifacts in learned representations.
  – Isolated and reconstructed the parts of the image that maximally activated a specific feature.
  – This analysis revealed that many filters in earlier networks were redundant or noisy.
- Improvement: +3% top-1 accuracy vs. AlexNet

19

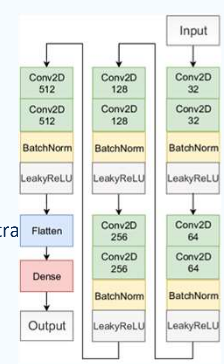
03 VGGNet: For Large Image Recognition



- Uniform Use of 3x3 Convolutional Filters Across All Layers:
  – Larger filters like 5×5 or 7×7 replaced with stacks of 3×3 filters.
  – Two stacked 3×3 filters approximate a 5×5 receptive field.
  – Three stacked 3×3 filters approximate a 7×7 receptive field.
- Following mathematical configuration gives same special coverage as with fewer parameters.

$$2 \times 9C^2 = 18C^2 < 25C^2$$

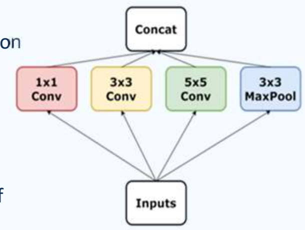
- VGG-16 consists of 13 convolutional layers followed by 3 fully connected layers.
- VGG-19 extends this to 16 convolutional layers.
- ReLU after every convolution ensures non-linear transformation among layers.
- Every few convolutional layers are followed by a 2×2 max pooling layer. This reduces the feature map size by a factor of 2.
- Consistent Channel Expansion: As the network goes deeper, the number of filters increase: 64 → 128 → 256 → 512.
- VGG-16 has 138 million parameters, mostly in the fully connected layers. Despite its size, the architecture tra reliably and delivered excellent generalization.
- VGG-16 achieved ≈ 7.3% top-5 error, better than AlexNet and ZFNet
- Key takeaway: depth via small filters boosts expressiveness



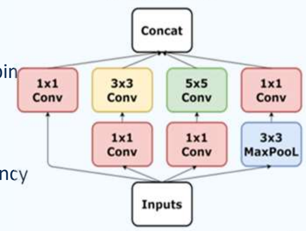
20

04 GoogLeNet: A Multi-Scale Design

- Inception Module with Parallel Paths - Instead of applying a single filter size per layer, the Inception module applies four operations in parallel:
  – 1×1 convolution
  – 3×3 convolution
  – 5×5 convolution
  – 3×3 max pooling
- Before applying expensive 3×3 and 5×5 filters, a 1×1 convolution is used to reduce the number of input channels.
  Example: Going from 256 input channels to 64 using 1×1 conv → then apply 3×3.
- GoogLeNet stacked 9 Inception modules, achieving 22 trainable layers deep and added auxiliary classifiers halfway through to help train such a deep model.
- Two small classifiers were added mid-way through the network. They acted like regularizes, helpin gradients flow and improving convergence.
- Similar to VGG, the model uses ReLU activation throughout.
- Top-5 error: ≈ 6.7 % with efficient FLOPs
- Inception demonstrated that careful factorization of convolutions can drastically improve efficiency without sacrificing accuracy.



Inception Module

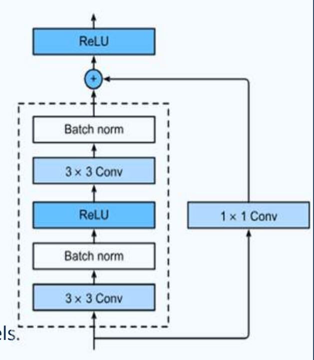


Inception Module with Dimension Reduction

21

05 ResNet: Residual Learning

- Residual Learning Framework:– Instead of computing $H(x)$H(x) (the desired mapping), the network learns $F(x) = H(x) - x$. The final output becomes: $y = F(x, W) + x$
  Where x is the input, $F(x, W)$ is residual function, and y is the output.
- Vanishing gradient solved: identity shortcuts let $\partial L / \partial x$ bypass layers.
- Bottleneck Blocks for Scalability: 1×1 → 3×3 → 1×1 for depth with efficiency
  – The 1×1 convolutions reduce and then restore dimensionality.
- Unlike VGG, where deeper networks sometimes performed worse, ResNet showed monotonic improvements with depth.
- Training Extremely Deep Networks:
  – ResNet-50, ResNet-101, and ResNet-152 became standard benchmarks.
  – Even a 152-layer ResNet outperformed all shallower versions.
  – 5x5 convolution
  – 3x3 max pooling
- Performance: ResNet-50 ≈ 5.3 % | ResNet-152 ≈ 4.5 % top-5 error
- Residual connections became standard practice in almost all vision and transformer-based models.
- ResNet is now used in Faster R-CNN, Mask R-CNN, RetinaNet, and more.
- Its modularity, clarity, and gradient flow properties remain unmatched.



22

Core Design Patterns in CNN

- Depth vs. Width: Using many small filters in deep stacks enhances feature richness (VGG).
- Interpretability: Visual diagnostics like DeconvNet guide architecture choices (ZFNet).
- Factorization: Breaking large convolutions into smaller parallel paths improves efficiency (Inception).
- Gradient Stability: Identity shortcuts ensure reliable training of extremely deep models (ResNet).

(2012)
AlexNet

ZFNet
(2014)

(2015)
VGG + GoogLeNet/Inception

ResNet
(2016)

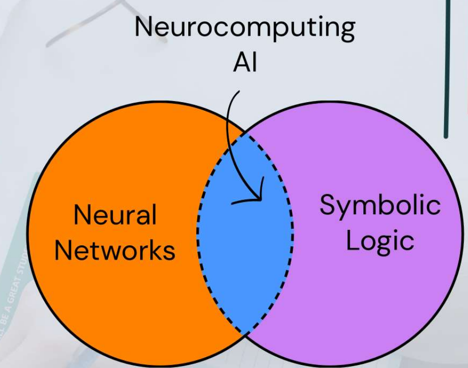
23

Thank you!

24

NEUROSYMBOLIC AI

By Dhvanil Bhagat

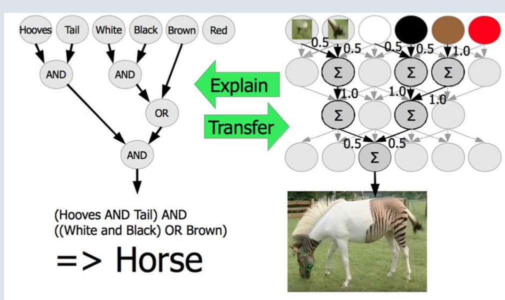


25

GOALS OF THIS RESEARCH

- Develop safe, real-world deployable AI systems
- Incorporate human-like reasoning into neural models
- Enhance interpretability and trust in AI decisions
- Enable faster training and generalization across environments

These goals aim to solve core problems in domains like autonomous driving and mental health care



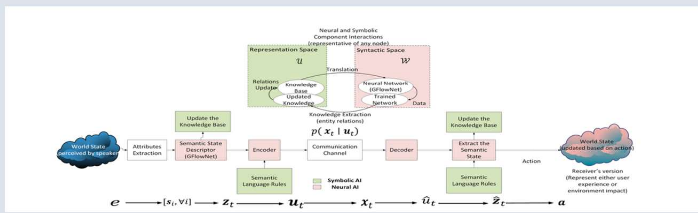
26

MOTIVATION & TECHNICAL CONTEXT

Traditional deep learning systems are powerful but face key challenges:

- Lack of explainability
- Unsafe decision-making in real-time settings
- Poor performance in novel or low-data situations

Neurocomputing AI integrates rules and reasoning to overcome these issues—injecting symbolic knowledge into models and reducing the risk of unsafe or opaque actions.



PROBLEMS AND PROPOSED METHODS

Problems

- Unsafe actions during DRL training in autonomous driving
- Semantic loss in AI-driven communication
- Black-box decisions in cybersecurity threat detection
- AI models lack clinical reasoning and produce unexplainable mental health predictions.
- Deep Learning models struggle with spatial rules and complex safety inspections.

Proposed solutions

- DRLSL filters actions using symbolic logic (Sharifi et al.)
- GFlowNet models causal meaning to preserve intent (Thomas et al.)
- Knowledge graphs enable explainable AI decisions (Piplai et al.)
- Infuse clinical knowledge into AI systems using symbolic rules for explainable, safe predictions. (Roy et al.)
- Combine Symbolic Logic with Computer Vision to interpret spatial relationships and enforce safety rules. (Luo et al.)

ASSUMPTIONS & LIMITATIONS

SHARIFI ET AL.

📌 Safety rules can be formalized in logic
⚠ Hard to scale symbolic rules to real-world complexity

ROY ET AL.

📌 Expert ontologies reflect clinical reality
⚠ Symbolic AI may not generalize across populations

THOMAS ET AL.

📌 Shared knowledge bases exist between nodes
⚠ Symbol grounding is complex and costly

PIPLAI ET AL.

📌 Threats can be structured in knowledge graphs
⚠ Limited performance against novel attacks

LUO ET AL.

📌 Spatial safety rules can be logically encoded
⚠ Symbolic models lack flexibility in dynamic sites

29

EXPERIMENTS & RESULTS

Roy et al.

Design: Used annotated Reddit and clinical datasets with ontology-guided classification
Results:
- Improved detection of mental health symptoms
- Better gender-specific symptom recognition

Piplai et al.

Design: Combined knowledge graphs with reinforcement learning on malware datasets
Results:
- 8% faster convergence in RL
- 4% improvement in detection accuracy

Thomas et al.

Design: Simulated speaker-listener system using GFlowNet for message generation
Results:
- Maintained semantic reliability
- Outperformed traditional systems in intent preservation

30

COMPARISION, PROS & CONS

Pros

- Symbolic logic makes decisions explainable
- Rules prevent unsafe or harmful actions
- Less training data needed with domain knowledge
- Prior knowledge helps adapt to new scenarios

Cons

- Rule creation is effort-intensive
- May not adapt well to noisy data.
- Hard to scale logic in complex environments
- Neural and symbolic components are hard to merge

31

IMPROVEMENTS & MY VIEWS

Improvements

- Automate symbolic rule generation using large language models
- Improve integration between neural and symbolic components
- Enable adaptive learning of new logic from changing environments
- Optimize computational efficiency of reasoning modules
- Expand domain-specific knowledge coverage (especially in healthcare & cybersecurity)

My views

- Neurocomputing AI is powerful but needs to evolve beyond static rules.
- I believe combining deep learning with dynamic, learnable logic will lead to safer and smarter AI.
- Future models should continuously learn from real-world feedback, not just predefined ontologies.
- Human-AI collaboration should be built on transparency, trust, and explainability.

THANK YOU

32

Mapping and Localization with Deep ConvNets

Why is visual perception crucial for autonomous systems?

What are some real-world applications?

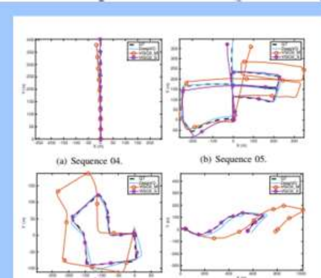
Why are there problems with traditional Mapping and Localization methods?

Matthew Bush

May 13, 2025

33

1. Localization: robot's position using sensors
2. Mapping: building a model of the surrounding environment from those visual cues.
3. SLAM and VO: methods use ConvNets to learn features and improve performance.

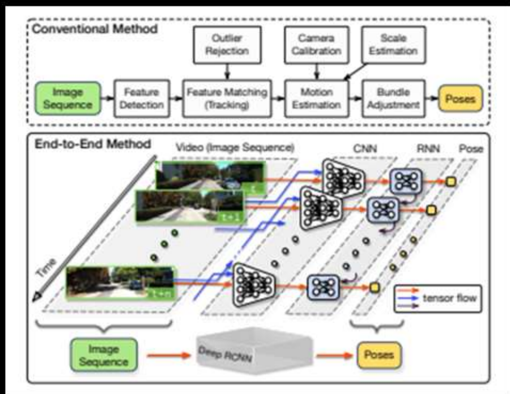


What is Localization and Mapping?

Matthew Bush

Page 2

34



Visual Odometry

1.

Analyzes consecutive image frames. Relying on matching features between frames

2.

Features like corners or textures but this has problems in low light

3.

CNN-Based VO learns motion patterns directly from pixel sequences

4.

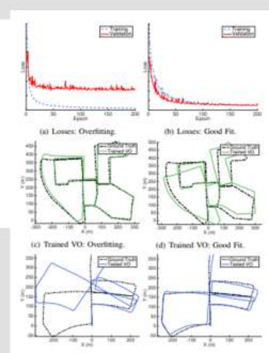
DeepVO and SfMLearner combines CNNs with recurrent layers to learn motion patterns.

35

Input Data

DeepVO

- CNN Find the geometric features to be used in the RNN
- MSE error is scaled for distance and angle errors
- RNN has feedback loops so the current value can be affected by what happened previously
- Long term dependencies might vanish



Model Output

DeepVO (Wang et. 2018)

36

Pose Estimation

Feature Match

Loop
Closure

- Camera
- Radar
- Lidar

- DoG
- SIFT

- Probabilities
- Outlier elimination
- RANSAC
- Checks the changes

- Seen locations
- Feeds back

- Aggregation of errors
- Feeds back

Sensors

Bundle
Adjustment

Understanding SLAM

Matthew Bush

Page 5

37

DeepSeqSLAM (Chancán, et al. 2020)

- Sequence-based place recognition methods perform well under extreme conditions

- single monocular image sequence.

- Tested on Nordland (728 km) and Oxford RobotCar (10 km) datasets.

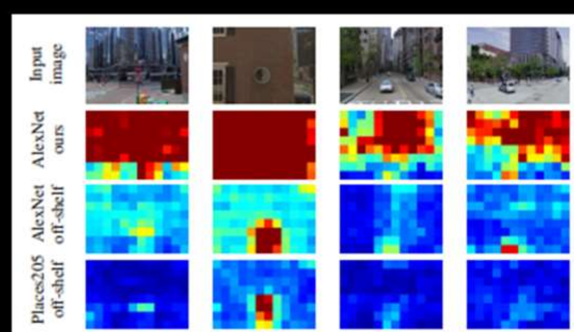
NetVLAD (Arandjelović et al., 2016)

- Combines CNN + VLAD pooling to create compact, trainable place descriptors.

- Trained on large datasets like Google Street View Time Machine examples.

- Widely used in visual place recognition benchmarks.

Place Recognition



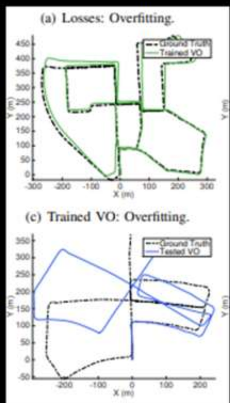
NetVLAD
CNN architecture for weakly supervised place recognition

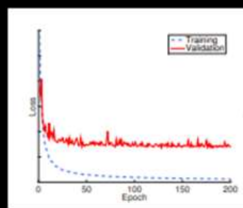
Authors: Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, Josef Sivic

Matthew Bush

Page 6

38





1.

Generalization to New Environments
CNN models often perform well in the environments they're trained on, but struggle in unfamiliar settings.

2.

Data Scarcity and Bias
Most public datasets KITTI, Nordland, Oxford are region-and or season-specific.

3.

Computational Complexity
Deep networks are often too heavy for real-time deployment on resource-constrained devices like drones

4.

Lack of Explainability
CNNs are typically black-box models so it's difficult to understand why a certain pose or place was predicted.

Limitations and Challenges

Matthew Bush

Page 7

39

DeepVO and SfMLearner

Demonstrated CNN-based visual odometry ideal for GPS-denied drone navigation.

CNN-SLAM

Predicted dense depth maps allow drones to fly through 3D environments and maintain real-time awareness.

NetVLAD Place Recognition

Shows how how drones can relocalize over long flights despite lighting or seasonal changes.

Future of Drone Navigation

- CNN + RNN methods like DeepVO and DeepSeqSLAM help maintain localization over time without GPS.

- Drones can update maps and re-localize as scenes change

- Use visual cues for terrain classification, slope estimation, and target detection.

- Drones could share CNN-processed visual maps and collaboratively explore large areas.

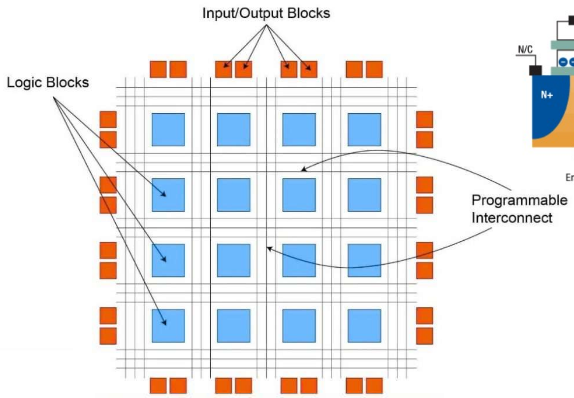
Goals & Outcomes

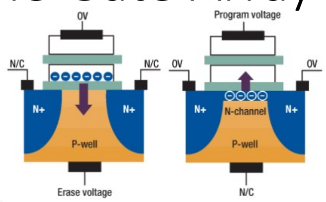
Matthew Bush

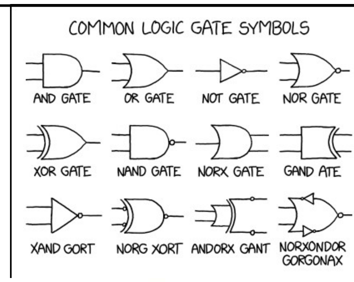
Page 8

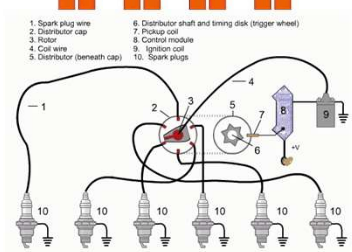
40

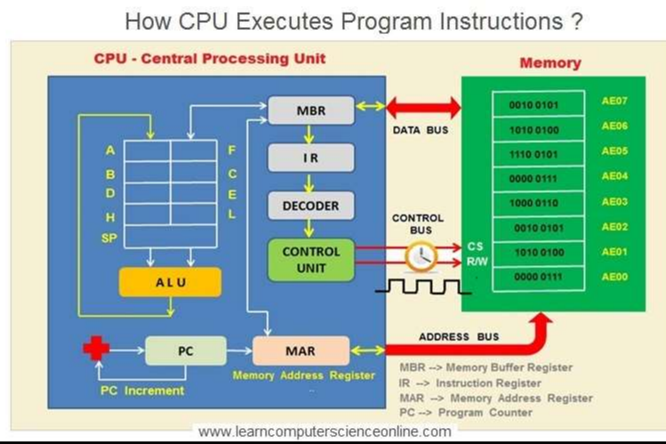
Field Programmable Gate Array



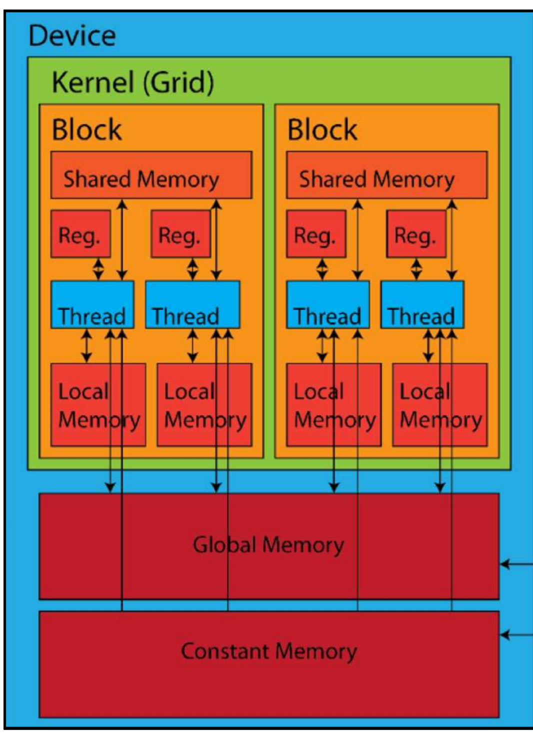








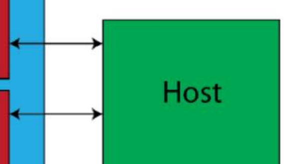
41



Global Memory
- Neural network weights and biases
- Input images
- Output data

Shared Memory
- Feature maps
- Convolutional layer values
- Gradients
- Matrix multiplication or vector operations

Local Memory
- Loop counters
- Backpropagation values
- Activation function outputs
- Partial sums



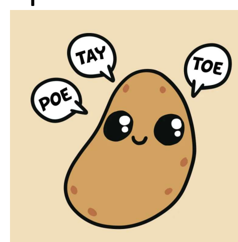
Epochs	CPU Time (seconds)	GPU Time (seconds)
100	5.87	0.28
1,000	295.35	2.68
10,000	3552.43	26.61

All accuracies above 98%

42

Tangent Distance Classifier

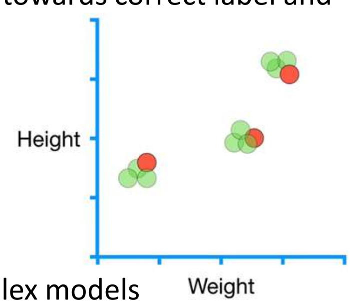
- A nearest-neighbor classifier robust to small transformation or distortion such as translation, rotations, scaling, or skewing
  - Notice that shifting horizontally, vertically, and rotations wrap around
  - Distortions effectively increased size of dataset
- Transformation moves vector in multidimensional space
- Forms a smooth manifold, meaning derivative can be taken infinitely
- A tangent vector is the directional derivative with respect to a specific transformation. Many tangent vectors form a tangent plane.
- Tangent vectors are the columns of a Jacobian matrix
- Unlike raw pixel distance, TDC is insensitive to small variation



43

Modified Machine Learning Algorithms 2003

- Learning Vector Quantization
  - Start with centroids of K-Means clustering now referred to as “prototype”
  - Using unseen data, train our prototype: move correct prototype towards correct label and wrong prototype away. Decrease learning rate over time.
  - Relatively fast and memory efficient but less accurate
    - KNN stores all training data. LVQ only stores prototype data
- Discriminative Learning Quadratic Discriminant Function
  - Start with QDA classifier parameters instead of random weights
  - Loss function is Minimum Classification Error
    - Contains regularization coefficient that penalizes more complex models
  - Update parameters using stochastic gradient descent
    - Stochastic gradient descent works great when your data is clustered, which is what 0-9 digits should be

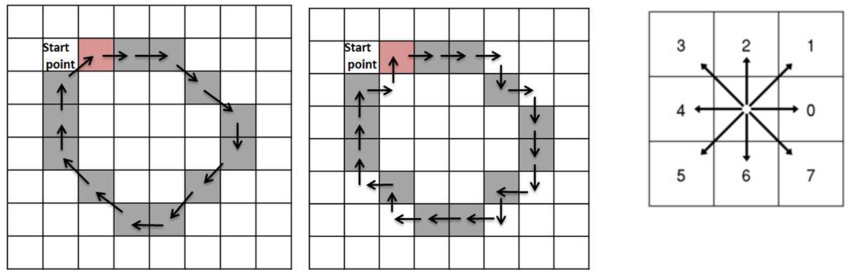


44

Feature Extraction Methods

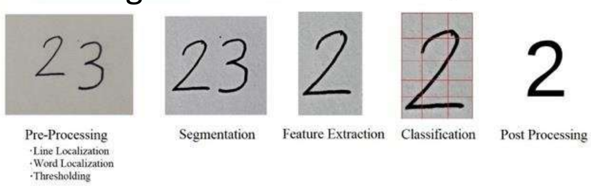
- Chain Code

Unordered count used as input vector



A B

- Crossing Count



- Proportion of black pixels in zone
- White runs: lengths of consecutive white pixels between dark pixels

45

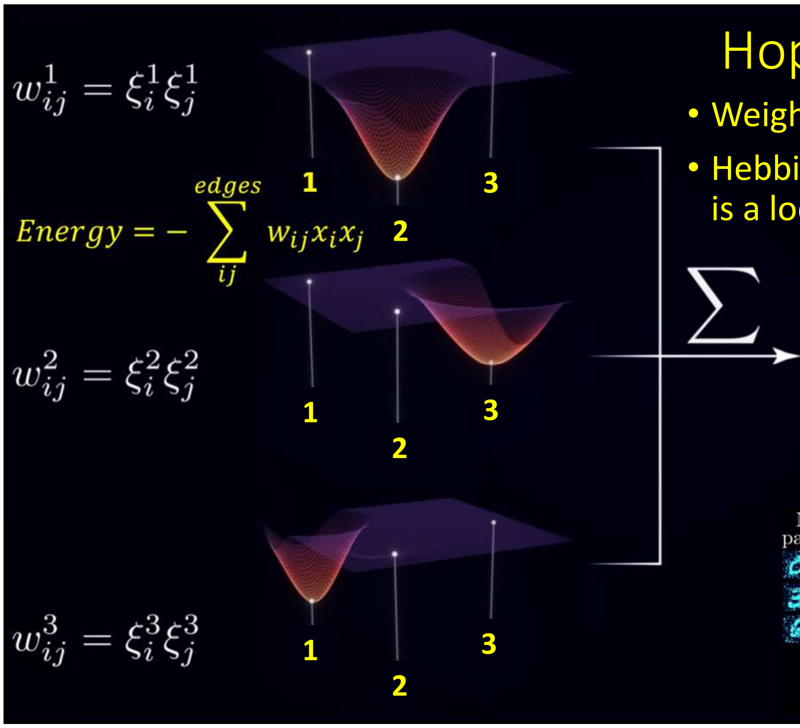
Adaptive Learning Rate

✓Increases learning rate if progress is in the same direction as previous

✓If gradient direction conflicts with previous, learning rate is decreased to reduce oscillation or overcorrection.

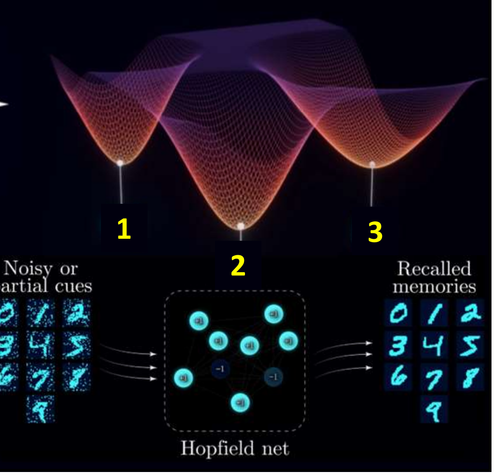
- Adam optimizer (Adaptive Moment Estimation)
  - Gradient descent with velocity term
    - If gradients point in the same direction repeatedly, momemtum accumulates and accelerates learning in that direction
    - If gradient fluctuates, momentum helps dampen the oscillation
- RMSProp (Root Mean Square Propagation) scales the learning rate for each parameter inversely proportional to the recent magnitude of its gradients, thereby adapting the learning rate individually for each parameter over time.

46



Hopfield Network

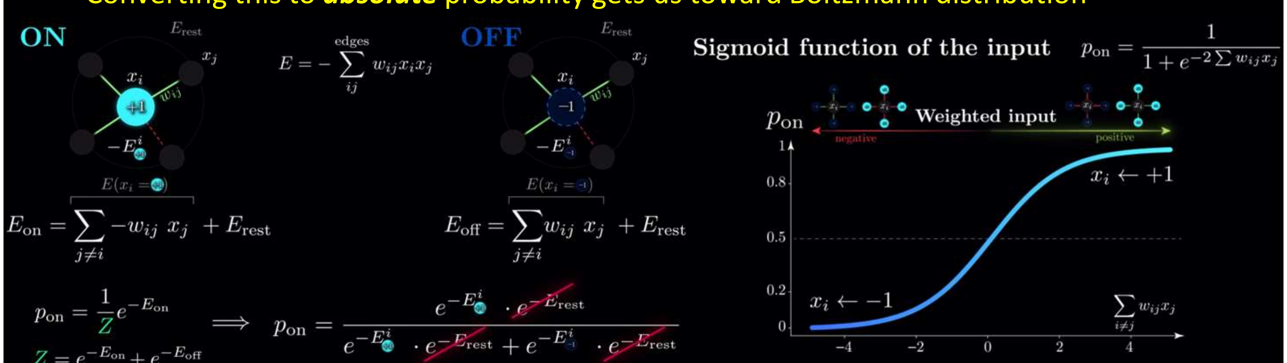
- Weights form energy landscape
- Hebbian weights for trained pattern is a local minima



47

Boltzmann Machine

- Hopfield network will always deterministically move to lower energy state
- Boltzmann's equation from physics states change in energy when moving up or down a state, which can be written as ***relative*** probability
  - Converting this to ***absolute*** probability gets us toward Boltzmann distribution



48

Deep Learning & NLP in Intelligent Machines

by Sai Praneeth Gudala
923832283

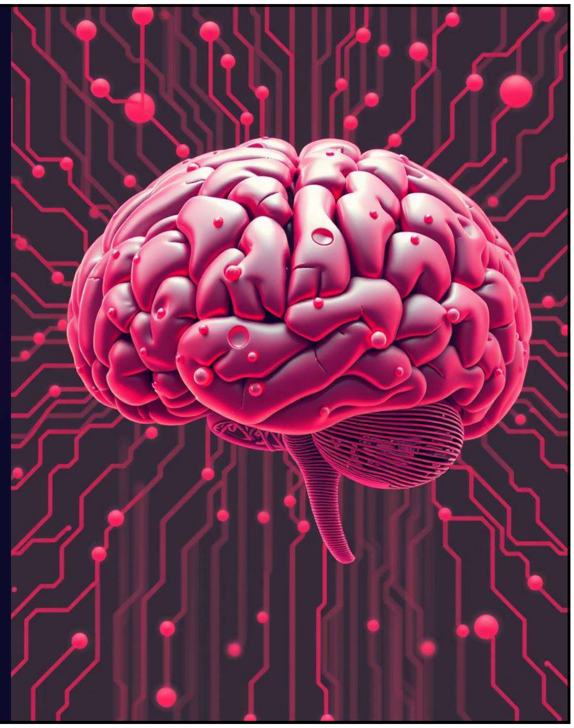
Intelligent machines mimic human intelligence for complex tasks

Deep Learning and NLP are foundational technologies

Machines learn from data, adapt, and improve over time

Applications span healthcare, finance, and customer service

Can machines truly understand human language and emotions?



49

What Are Intelligent Machines?

Cognitive Simulation

Machines simulate learning, decision-making, perception

Adaptive Behavior

React and evolve based on data inputs

Use Cases

- AlphaGo Zero mastering Go without human data
- Moley robot mimicking chefs
- High-risk task automation

50

Role of NLP in Understanding Language

NLU – Natural Language Understanding

- Understands grammar, context, and user intent.
- Analyzes parts of speech, named entities, dependencies.
- Translates human sentences into structured data a machine can act on.
- Powers virtual assistants, spam filters, and recommendation engines.

NLG – Natural Language Generation

- Converts machine-readable data into human-like language.
- Creates readable summaries, responses, or reports.
- Used in chatbots, report generation, and personalized content.
- Text-to-speech systems are a key output of NLG.

51

Power of Deep Learning

Inspired by Human Brain

Multi-layer Networks

CNNs

Extract image and text features

RNNs & LSTMs

Handle sequential data

Enables complex pattern learning for translation, medical imaging

52

Sentiment Analysis – Decoding Emotions

Emotion Detection

Positive, negative, neutral text sentiment

Applications

Social monitoring, brand reputation, politics, finance

Challenges

Sarcasm, idioms, emojis, polysemy

Tools & Methods

VADER, TextBlob, deep learning for context

53

Question Answering with Deep Similarity Models

Embedding Vectors

Represent questions & answers numerically

Deep Similarity Networks

Measure semantic closeness between queries and responses

Applications

- Support chatbots
- Google snippets
- Quora, StackOverflow

Training

Doc2Vec, Word2Vec, Transformer models

54

Decision-Making & Problem Solving

1 **Generate Solutions**
Brainstorm multiple approaches

2 **Assess Plans**
Evaluate outcomes logically

3 **Assemble Strategy**
Choose optimal path using learned patterns

Used in robotics, autonomous vehicles, smart assistants

55

Conclusion & Future Vision

From Automation to Intelligence
Deep Learning and NLP advance machine capabilities

Emotionally-aware AI
Personalizes interactions with sentiment understanding

Challenges Ahead
Bias, transparency, ethical risks

Future Vision
Machines think, feel, assist with empathy and speed

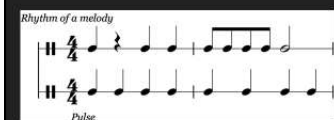
Are we ready for truly intelligent machines?

56

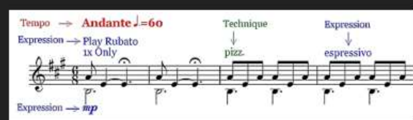
Practice Doesn't Make Perfect: Music Generation with Deep Learning

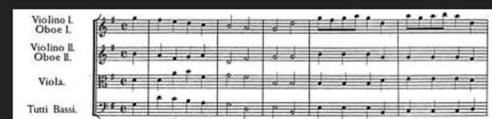
Alexander Kaattari-Lim
CSC 872

Rhythm and Time



Expression and Interpretation



Multiple Voices

57

Early Efforts

- Theoretically can capture past information
- Quantization 8 notes per bar → 96 time steps
- LSTM based model
- 1st exp: model can reproduce a musical chord structure
- 2nd exp: model can learn melody and chords

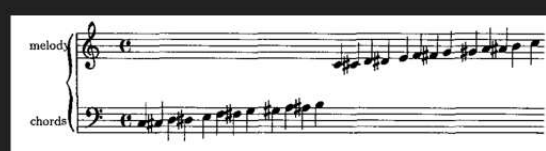
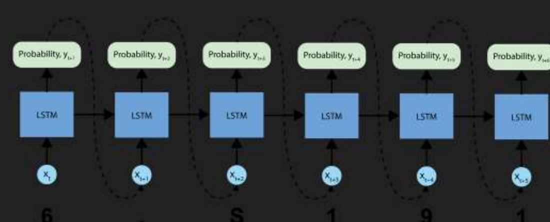


Figure 1: Possible note values for these simulations

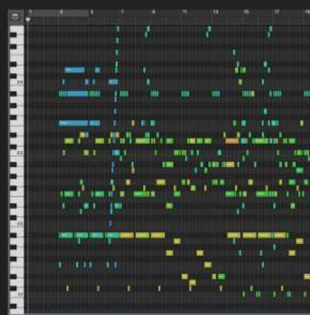
Figure 2: Chords for training data (transposed up one octave).

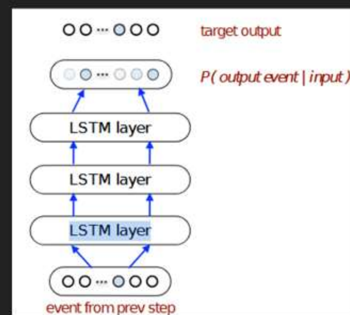


58

Capturing Musical Expression

- Tackling musical expression
- MIDI data representation
- Much deeper LSTM model than before 3 layer 512 cells vs 2 cells per layer
- 125 hz sampling rate for time steps

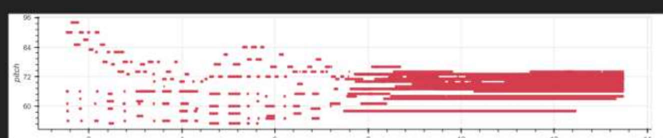




59

Limitations of RNN Based Models

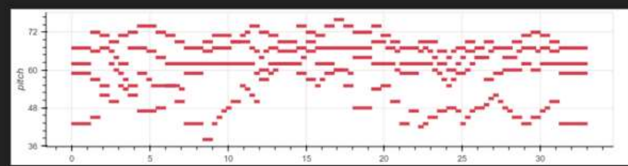
- User studies showed lacking long term coherent structure
- Many clips sound like a mix of classical composers
  - Lacks musical identity
- Computationally intensive for generating longer pieces of music (15sec clips or shorter)

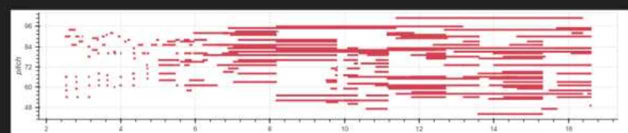


60

Longer Form Musical Structure

- Advent of transformer allows for capturing longer dependencies
- Continues capturing stylistic expression
- Contributes Memory Efficient Relative Position Based Attention
  - $O(L^2D) \rightarrow O(LD)$

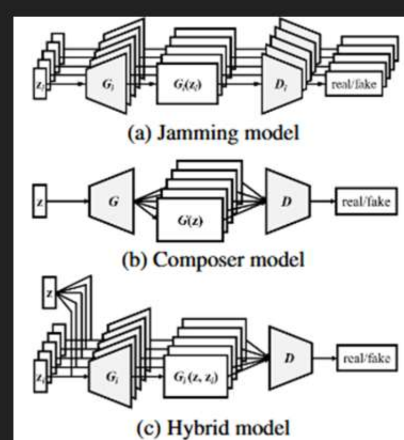




61

Handling Multiple Voices

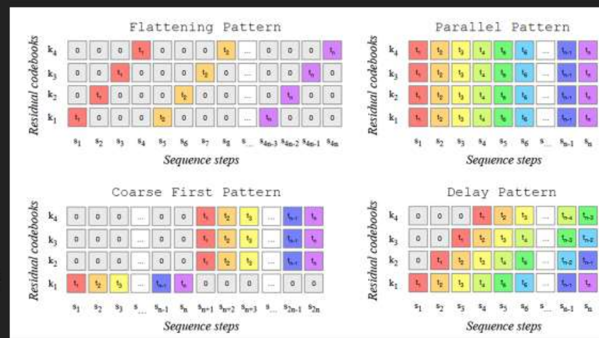
- Convolutional Generative Adversarial Networks (GANs) to synthesize multiple tracks
- Jamming: 1 gan / track
- Composer: 1 gan / piece
- Hybrid: 1 gan / track but with shared inputs



62

Controlling Music Generation

- Generates music with text to music or melody to music
- Single model
- Utilizes codebooks to represent musical patterns, essentially tokenizing
- Language input maps to codebooks which maps to music
- Codebooks are learned



63

Future Work & Insights

- Musical style transfer
  - Can we reimagine happy birthday in the style of Bach or Chopin?
- More fine grained generation as an engine for better expressive and interpretive capabilities
- Long Long form multi instrument composition with structure

Thank you!

64