Soft Computing Project

Correlational Neural Networks

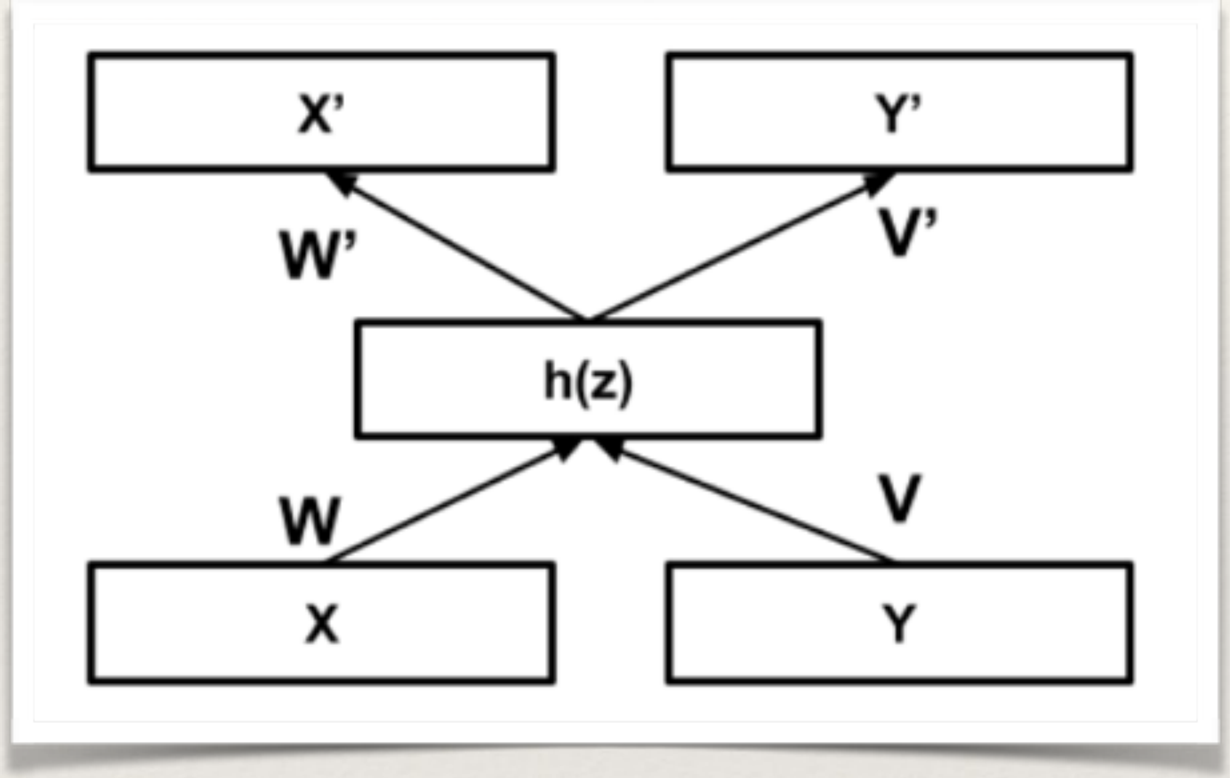
Akshita Mittel
CS13B1040

Project Guide:
Debaditya Roy

Motivation

- *"Detecting complex video events based on audio and visual modalities is still a largely unresolved issue."*
- The paper: Video Event Detection via Multi-modality Deep Learning had used a single layer neural network (CNN) to represent video for event detection.
- The main contribution from their end was enhancing the regularisation term, from the error function derived from RICA.
- The main objective was to come up with an alternate architecture that could be applied to more general cases.
- There were several such architectures, but the one that I chose for my project was Correlational Neural Network (CorrNet).

Correlational Neural Networks



- Common Representation Learning (CRL), wherein different views of the data are embedded in a common subspace.
- The applications described in the paper include:
  - Reconstruction of images from one half
  - Translations between languages such as Eng/French, Eng/German

Correlational Neural Network

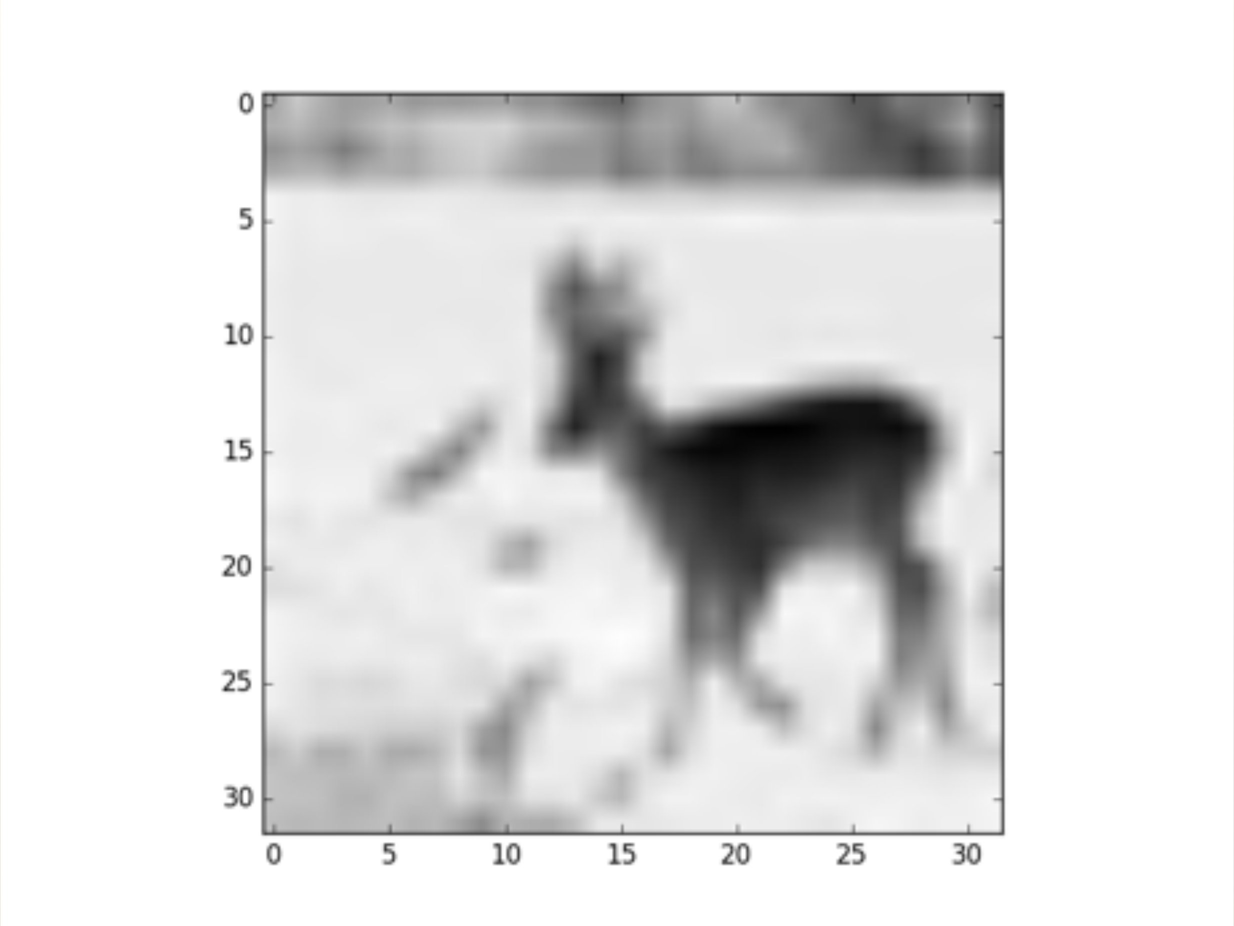
$$\mathcal{J}_{\mathcal{Z}}(\theta) = \sum_{i=1}^{N} \left(L(\mathbf{z}_i, g(h(\mathbf{z}_i))) + L(\mathbf{z}_i, g(h(\mathbf{x}_i))) + L(\mathbf{z}_i, g(h(\mathbf{y}_i))) \right) - \lambda \operatorname{corr}(h(X), h(Y))$$

$$\operatorname{corr}(h(X), h(Y)) = \frac{\sum_{i=1}^{N} (h(\mathbf{x}_i) - \overline{h(X)})(h(\mathbf{y}_i) - \overline{h(Y)})}{\sqrt{\sum_{i=1}^{N} (h(\mathbf{x}_i) - \overline{h(X)})^2 \sum_{i=1}^{N} (h(\mathbf{y}_i) - \overline{h(Y)})^2}}$$

- The fundamental equation behind correlational neural network has 4 main parts:
  - The reconstruction error for each modality individually
  - The reconstruction error with both the modalities together
  - The advantage gained from the correlation between the two modalities

Objective

- To expand the horizon of the network by including dataSets which aren't as simple as the MNIST dataSet.
- For this objective we use the CIFAR10 image dataSet.
- This has approximately 60,000 tiny 32x32 pixel images, with appropriate labels.
- The main goal as the motivation suggests is to apply this to a multi modal domain, which is not restricted to just images.



Example CIFAR 10 images after basic preprocessing.

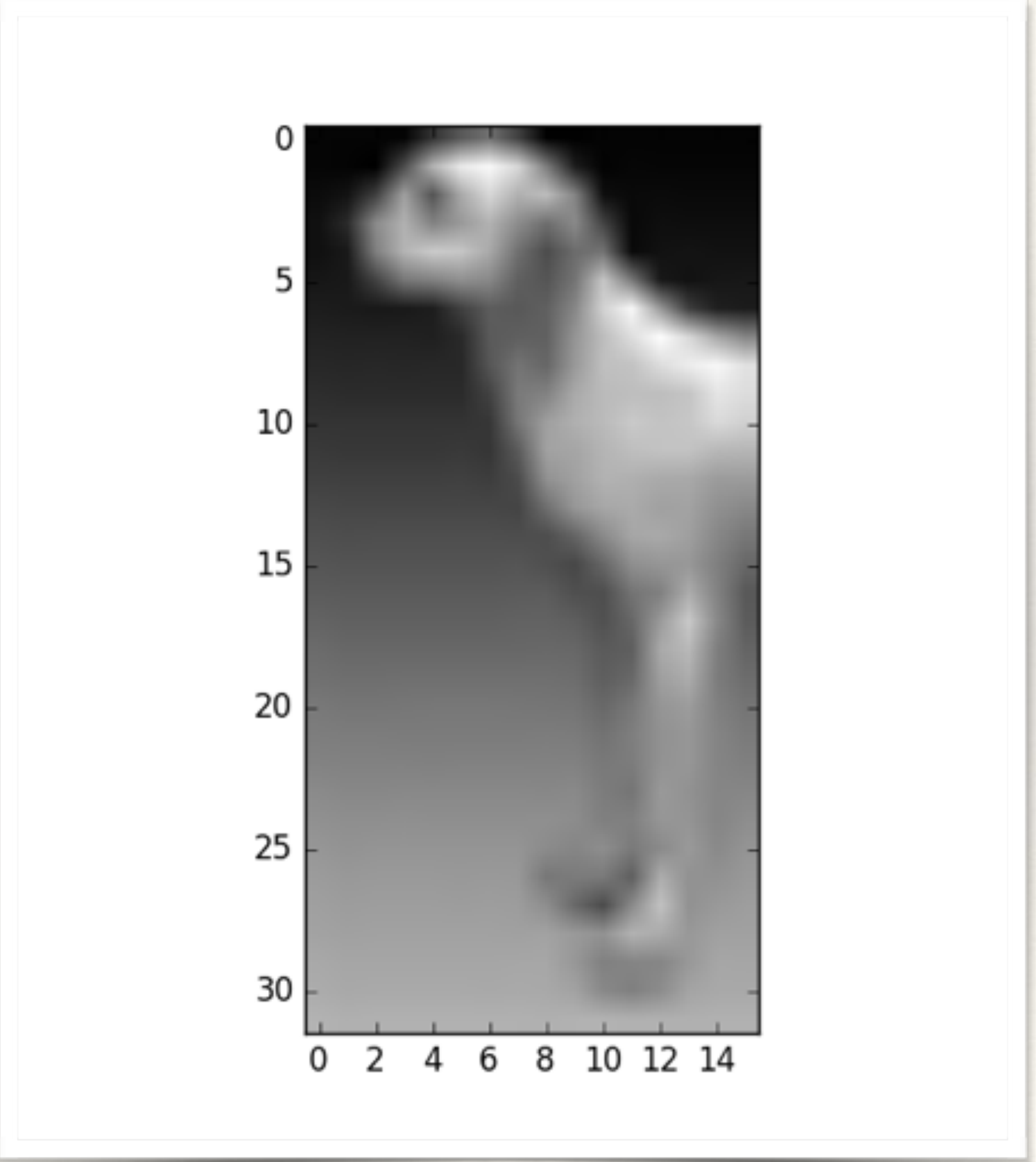
MNIST dataSet

- **Accuracy:**
  - The transfer learning for left-right and right-left are coming as 77.05% and 78.81%
  - The correlation is 42.57% for the MNIST data set.
- **Architecture:**
  - These results required only a 3 layer deep network.
  - The activation function is sigmoidal
  - Weights are determined using SGD.

Modules for CIFAR10

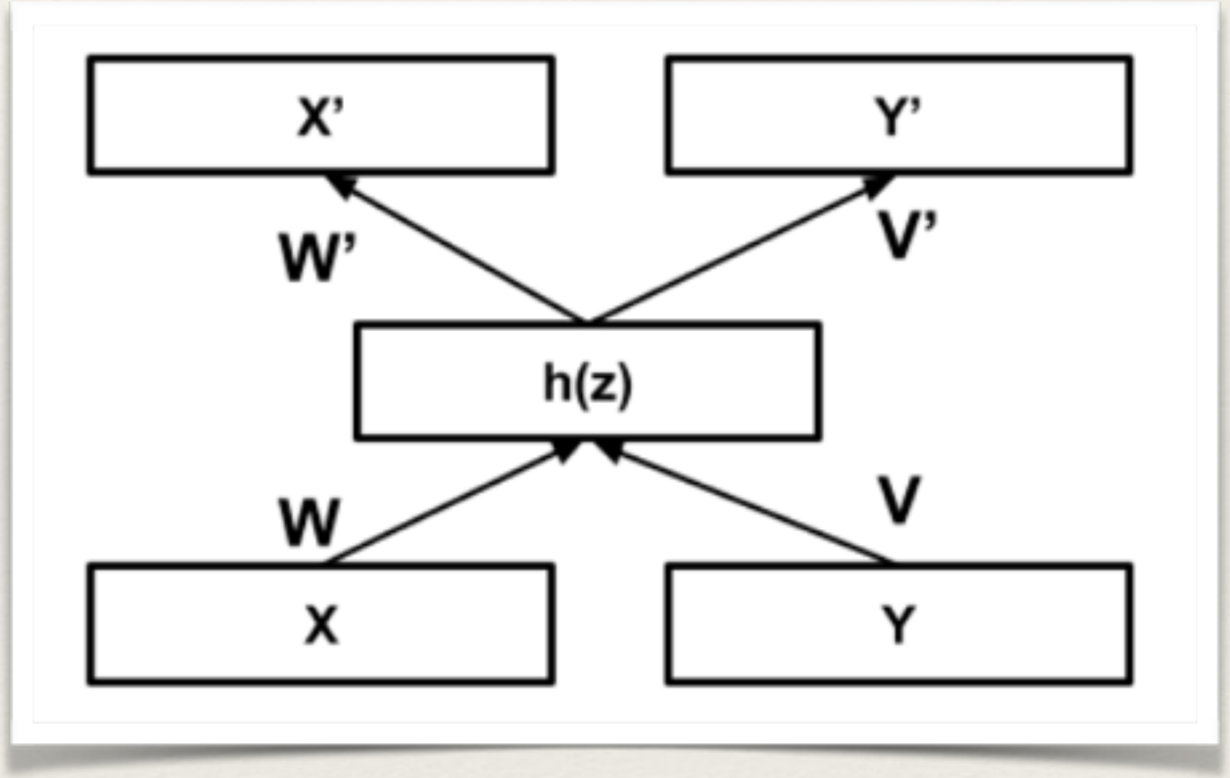
Modification to DataSet

- Extract the DataSet using Pickle.
- Reshape the dataSet from a vector of size 3072 to 3 instances of the image 32*32 (Red, Blue, and Green).
- Convert all the images to grey scale and divide them into left and right views.
- Convert each half of the image into a vector form (each of size 512). This will be fed into our CorrNet.
- Divide the images for Test, Train, and 2 Validation sets. Out of the 5 batches, the data is split as follows:
    - Test :10000x512 (batch 1)
    - Train: 20000x512 (batch 2, 3, & 4)
    - Validation 1: 8000x512 (batch 5)
    - Validation 2: 2000x512 (batch 5)



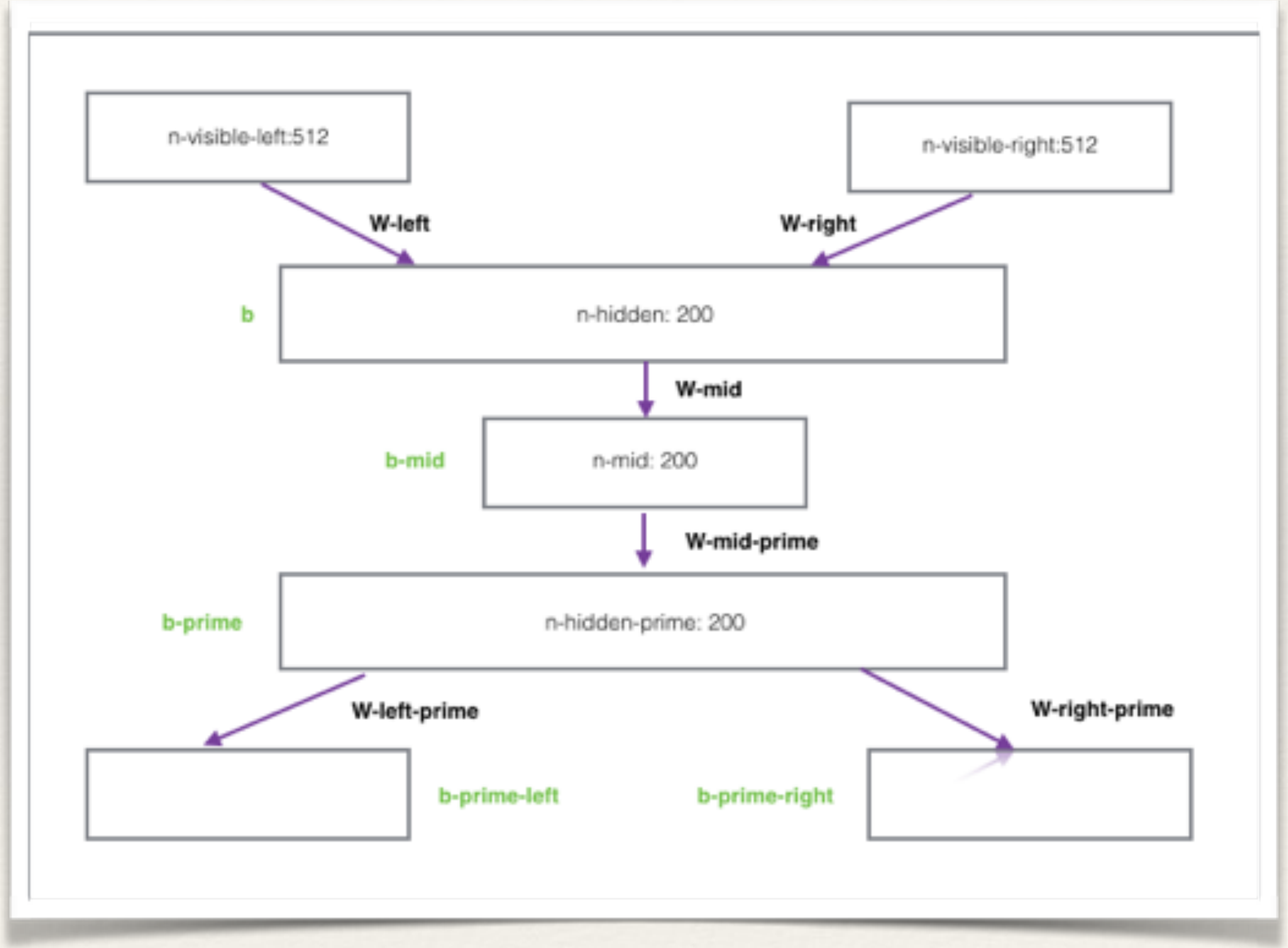
Basic Model

- The basic model had 3 layers:
  - **The input layer:** 1024 units, separated into 2 for each half of the image.
  - **The correlational layer:** This middle layer had 500 units.
  - **The output layer** was identical to the input layer.
- The accuracy:
  - The **transfer learning** for left-right and right-left are coming as 23.87% and 23.19%
  - The **correlation** is 36.40%

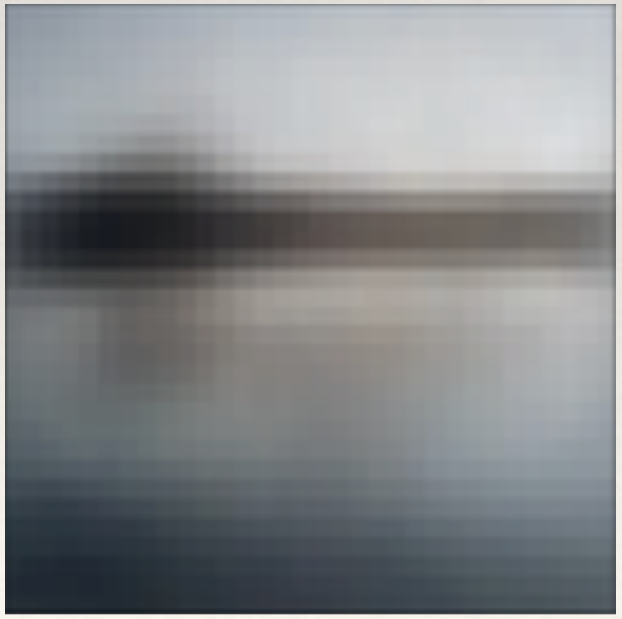


A deeper approach

- The model was created as shown
- The correlational accuracy was around 23%
- The transfer accuracy was around 11%
- It can be seen from the diagram that the reconstruction is just some abstract features that were drawn from the images.
- The reconstructed figures are not close to the original.
- A more robust model is required.



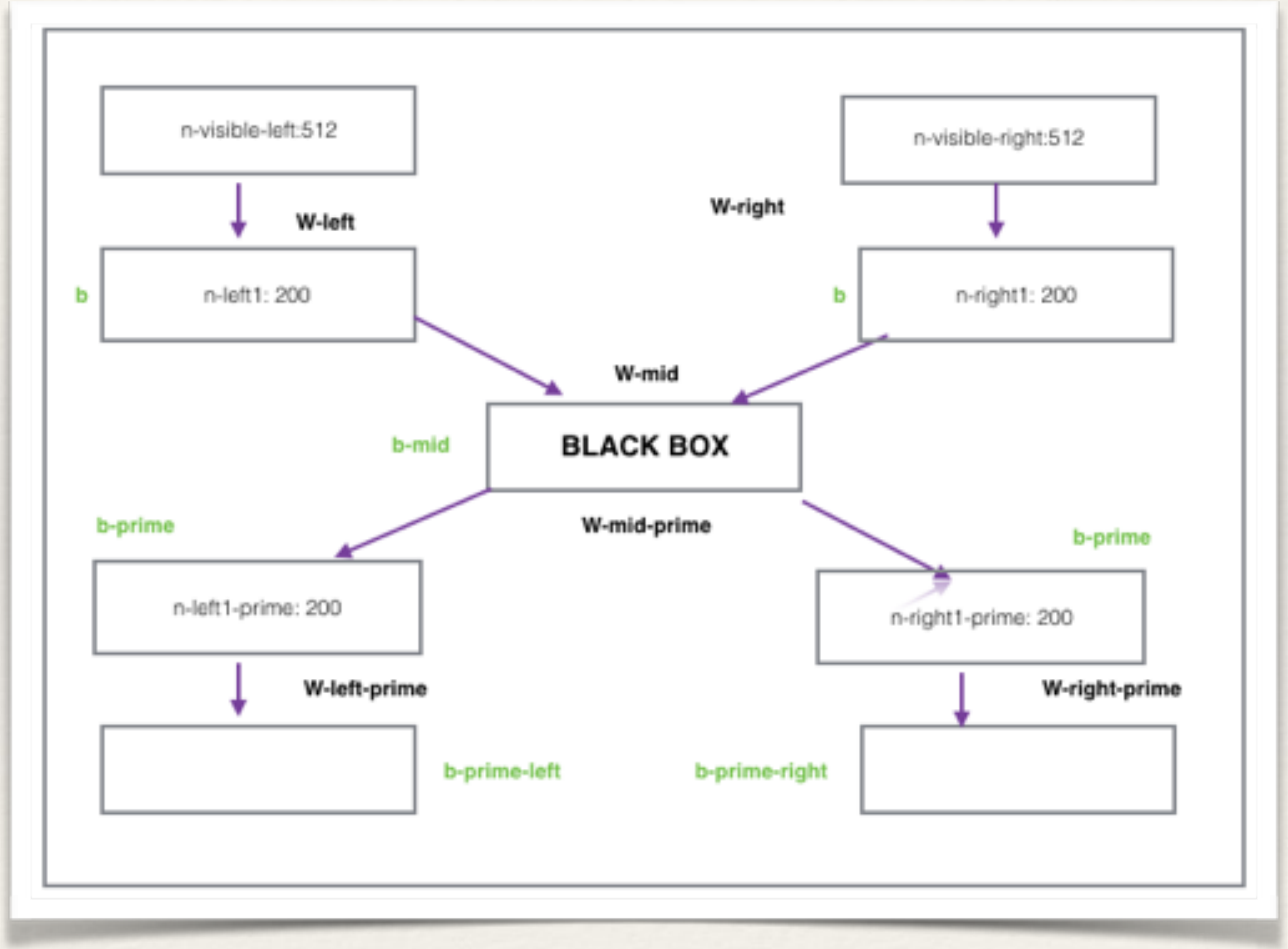
Very deep auto encoders



- Alex Krizhevsky and Geoffrey E. Hinton
- They used “very deep auto-encoders” to get a binary representation in the middle layer.
- Image retrieval was efficient for various dataSets, using this binary representation.
- An example of image retrieval for the CIFAR10 data set is shown on the left.
- The main points in consideration are:
  - They initialised the network using Deep belief networks.
  - Mean Square root error was used
  - The layers in the actual auto-encoder started from an arbitrary value associated with the image (336), the next layer was expanded to 1024 units.
  - From there on each layer was reduced to half the size of the previous layer, till we reached 32.
  - Instead of the innermost layer being 32, they used 28 neurons.

Application of Hinton's paper

- The middle section of the CorrNet14 is constructed using the architecture described in Hinton's paper.
- The current architecture is: 512*, 200*, 1024, 512, 256, 128, 64, 28.
- Instead of using CNN's as a preprocessor, the model has used Relu as an activation function.



Current work

- There were certain issue’s while training the CorrNet, I am currently working on resolving these. The idea’s that I am currently working on are as follows:
  - Increasing the size of the DataSet, by simple techniques such as pivoting and so on.
  - Using various Keras modules to train my network greedily layer by layer.

Extension of project

- Expand the range of modalities:
  - Instead of restraining to image correlations, aim to correlate between different modalities using CorrNet
- Possibility of employing dropout mechanism to enhance performance.

References

- **Video Event Detection via Multi-modality Deep Learning**
  - I-Hong Jhuo[1] and D.T. Lee[1,2]
    [1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
    [2]Dept. of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan ihjhuo@gmail.com, dtlee@ieee.org
- **Correlational Neural Networks**

  Sarath Chandar[1], Mitesh M Khapra[2], Hugo Larochelle[3], Balaraman Ravindran[4]

  [1]University of Montreal. apsarathchandar@gmail.com

  [2]IBM Research India. mikhapra@in.ibm.com

  [3]University of Sherbrooke. hugo.larochelle@usherbrooke.ca [4]Indian Institute of Technology Madras. ravi@cse.iitm.ac.in
- **Using Very Deep Autoencoders for Content-Based Image Retrieval**
  - Alex Krizhevsky and Geoffrey E. Hinton
  - University of Toronto - Department of Computer Science
  - 6 King's College Road, Toronto, M5S 3H5 - Canada
- The basic Code:
  - Sarath Chandar, Mitesh M Khapra, Hugo Larochelle, Balaraman Ravindran. [Correlational Neural Networks](http://arxiv.org/abs/1504.07225)

Thank you!