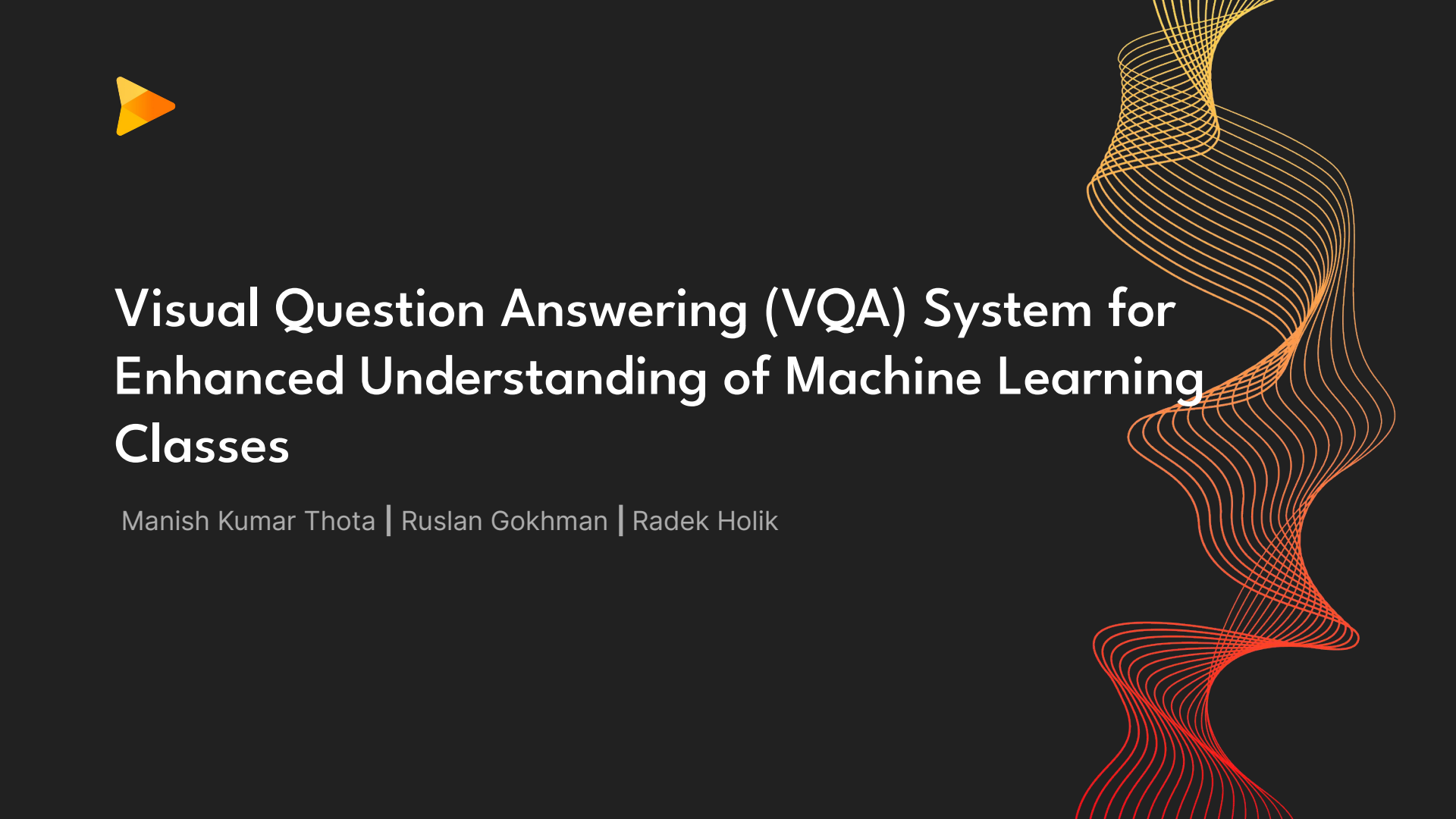


Visual Question Answering (VQA) System for Enhanced Understanding of Machine Learning Classes

Manish Kumar Thota | Ruslan Gokhman | Radek Holik

Introduction

01 **Digital Age Challenge:**

Increasing reliance on visual content in educational environments, particularly in machine learning (ML) lectures.

02 **Unique Aspects of ML Lectures:**

Dense with complex diagrams, mathematical equations, and information, posing comprehension and engagement challenges.

03 **Study Goal:**

Push the boundaries of Visual Question Answering to handle the complexity of educational content.

04 **Approach:**

Tested various models on a dataset from Yeshiva University, focusing on deep analysis and reasoning.

Data Collection

- **Image Collection**
  - Collected 885 slides from ML course PDFs using PyMuPDF.
- **Transcript Collection**
  - Transcribed lecture audio with various tools. Found Deepgram to be most accurate.
  - Manual checking needed due to poor audio quality.
- **Question and Answer Creation**
  - Manually developed 10 (average) QA pairs per slide, focused on the main topic.
  - Organized in a structured JSON format for easy retrieval.
  - Total: 9,416 samples in the data set.

Ref: https://arxiv.org/abs/2304.08485
https://github.com/microsoft/DeepSpeed

LlaVA (model, training)

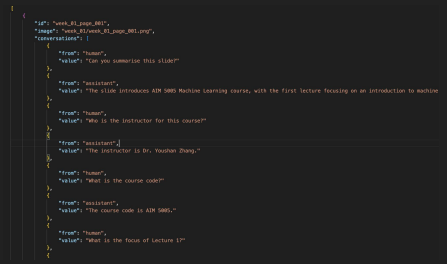
Fine Tuned and Pre Trained the LlaVA model with the following
Vision Encoder: CLIP-ViT-L-336px with an MLP projection
Text Encoder: Llama-2-7B / Llama-2-13B

4 LlaVA models with different combinations were fine tuned and pretrained on the data set. In short the models are labelled in series of F

F1: Fine-Tuned Llava 1.5 7B on the dataset for 50 Epochs.
F2: Fine-Tuned F1 on "Can you explain the slide" question only for 50 epochs.
F3: Pre-Trained on Llama-2-13B for 100 epochs and Fine-Tuned on LlaVA-1.5-13B for 10 epochs
F4: Fine-Tuned Llava-1.5-13B for 50 epochs.

Hardware:

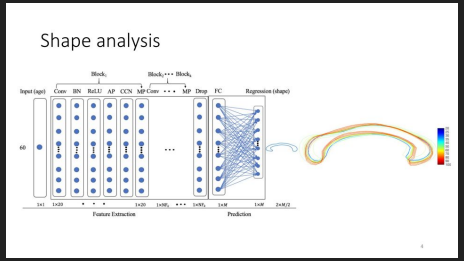
- Trained on RunPod Cloud with 1XA100 80GB 12 vCPU 125 GB RAM
- The entire training process was conducted using the 'DeepSpeed' framework developed by Microsoft.
- Training timings as follows
  - F1: 48 hrs
  - F2: 26 hrs
  - F3: 80 hrs
  - F4: 24 hrs



Dataset format for LlaVA model

Models		ROUGE	BLEU	COSINE
F1	Train	41.10	0.513	0.42
	Test	35.50	0.497	0.31
F2	Train	0.008	0.003	0.006
	Test	0.008	0.003	0.006
F3	Train	26.00	0.293	0.36
	Test	25.70	0.287	0.34
F4	Train	40.62	0.503	0.40
	Test	33.70	0.421	0.33

LlaVA (outcomes)



Week_01_page_005 - Train

Q: Is the output a classification or a regression?

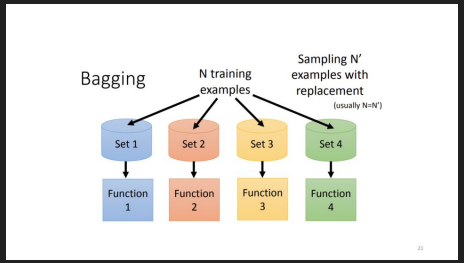
GT: The output is a regression predicting the shape.

F1: The output is a regression.

F2: The output is a regression.

F3: The output is a classification.

F4: It's a regression output.



Week_15_page_022 - Test

Q: How is sampling done in Bagging as per the slide?

GT: In Bagging, sampling is done by taking N' examples with replacement.

F1: Bagging sampling is done post-training and does not require retraining.

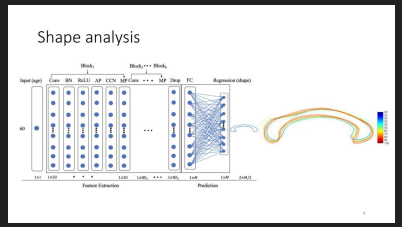
F2: Bagging sampling is done post-training and does not require retraining.

F3: Sampling in Bagging, as shown in the slide, is done by creating multiple bags or subsets of the original dataset. Each bag contains a random selection …. robust predictions.

F4: Sampling in Bagging, as shown on the slide, is done by selecting a subset of the data and processing it in training sets.

Ref: https://arxiv.org/abs/2304.08485

LlaVA (outcomes)



Week_01_page_005 - Train

Q: Can you explain the slide?

GT: And I will introduce one more of the previous that I developed out of these examples. That's exactly how manifold, which means that if we are keeping sampling some samples, from the source domain to the target domain. So now, you can see this image in the shoes. How monkey changes the interviewer working human being right. It's corresponding to those plan dot here. So, this is about the time frame of the many goals.

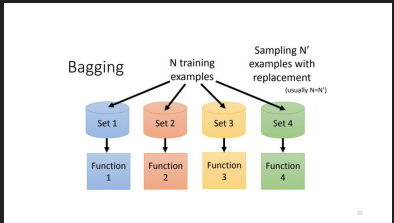
F1: And this is another example in the geodesic sampling on manifold. So, it's depicted by this kind of dots. Again, it's about human's brain structure. And here's more details of it.

F2: The slide presents a scatter plot with data points grouped into two clusters, with the objective function value for 'shape analysis'.

F3: The slide is a visual representation of a geodesic sampling on a manifold. It features a diagram with a group of people walking along a path, possibly on a sphere or a curved surface. The people are positioned at various points along the path, with some closer to the center and others near the edges. The slide also includes mathematical equations and symbols, which are likely related to the geodesic sampling process. These equations help to describe the relationship between the people's positions and the underlying mathematical structure of the manifold. Overall, the slide provides a visual and mathematical representation of the geodesic sampling concept.

F4: The slide discusses 'Geodesic Sampling on Manifold'. It visually represents data points on a manifold and explains how these points were sampled.

LlaVA (outcomes)



Week_15_page_022 - Test

Q: Can you explain the slide?

GT: This is the power of bagging, and here is the more about explanation of the process. For example, in this case, you have n training examples, you example with replacement, Set1, Set2, Set3, Set4. I mean, in each subset you will view train, new models. So here is a function of function 2, function 3, function 4.

F1: And this is about the samples. So, they are different than the training examples. Right? And this is a function for the regression problem. And this is the function for the classification problem. So, they are different. Right?

F2: The slide presents a flowchart of the 'Bagging' algorithm. It begins with labeling data and continues through training, sampling, and placing labels. The process involves setting parameters and performing operations like function and iteration.

F3: The slide presents a diagram that illustrates the concept of bagging, which is a technique used in machine learning to improve the performance of a model. Bagging involves training multiple models on different subsets of the same data, and then combining their predictions to produce a more accurate result. The diagram shows the process of bagging, with each model being trained on a different subset of the data. The outputs from these models are then combined to create a final prediction. The slide also includes a comparison between bagging and sampling, which is another technique used in machine learning. Sampling involves selecting a subset of the data to train the model on, whereas bagging involves training multiple models on different subsets of the data. The slide highlights the differences between these two techniques and their potential applications in improving the performance of machine learning models.

F4: The slide presents a decision tree with 'Bagging' as the primary node. It's further divided into two training examples: 'Set 1' and 'Set 2'. Each of these sets has multiple nodes, likely representing different training data. The ultimate goal is to 'Sample N'

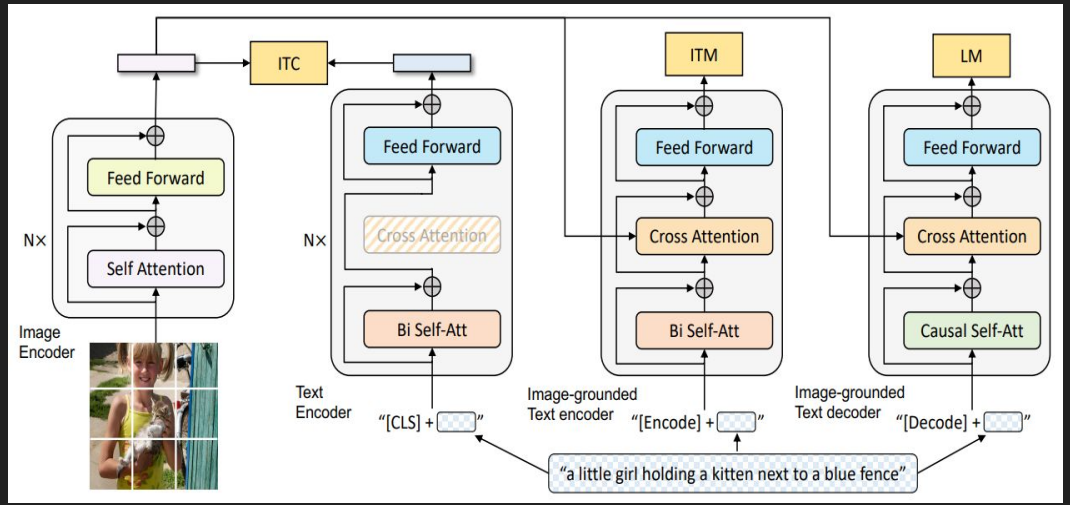
BLIP

BLIP (Bootstrapped Language Image Pretraining) model combines vision-language pre-training and fine-tuning stages to improve image-text understanding and generation.

Hardware:

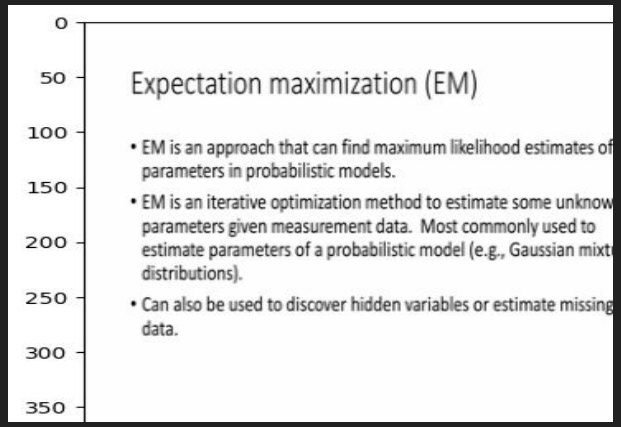
- Trained on INVIDIA Tesla V100 GPU
- The entire training process was conducted using the Google Colab Pro + environment.
- Training timings as follows
  60-70-epochs - 10-12 hours

Models		ROUGE	BLEU	COSINE
BLIP	Train	39.63	0.0	0.39809
	Test	39.681	0.0	0.37385
BLIP-2	Train	-	0.0	0.197
	Test	-	0.0	0.1809



https://paperswithcode.com/method/blip

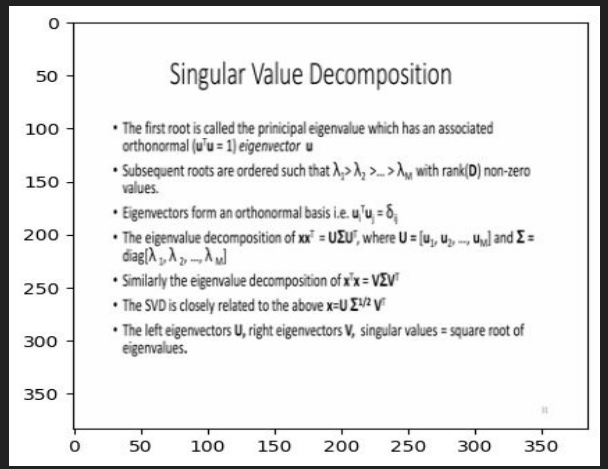
BLIP (outcomes)



Q: Is the output a classification or a regression?

Actual Answer: yes, em is an iterative optimization method.

Predicted Answer: iteratively combines probabilities with respect to each gaussian contributes to the expectation.

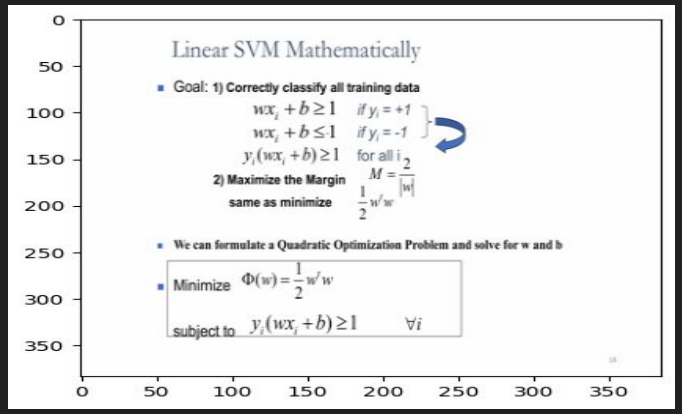


Q: what is the principal eigenvalue?

Actual Answer:the principal eigenvalue is the first and largest eigenvalue, associated with an orthonormal eigenvector.

Predicted Answer: it ensures that the algorithm identified observed features and often makes it more interpretable.

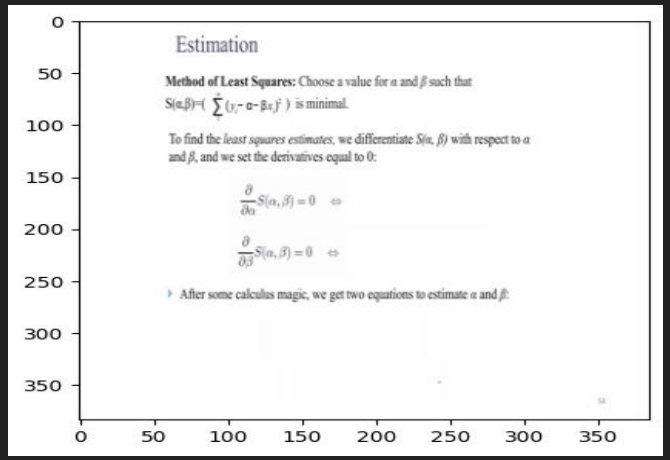
BLIP (outcomes)



Q: how does the svm decide the boundary?

Actual Answer: svm decides the boundary by finding a hyperplane that maximizes the margin between the two classes while correctly classifying all training data.

Predicted Answer: it ensures that svm finds the best possible probability of correctly classifying its large margin



Q: what is the goal of the method of least squares?

Actual Answer:the goal is to minimize the sum of squared differences for best fit.

Predicted Answer: it aims to minimize the sum of squared differences between observed and predicted values.

Custom Model (model, training, outcomes)

Model Summary:

Visual Model: DeiT.
Text Encoder: GPT-2 Large.
Text Decoder: GPT-2 Large with LM Head.

Feature Integration: Fusion layer combines visual and text embeddings.

Training Details

- Total Parameters: 1.64 billion (All trainable).
- Trainable Parameters: 103.68 million.
- Deployment: Model is running on GPU.
- GPU Memory Allocation: 6.4 GB.

- Initial Learning Rate: 1e-6, using Adam optimizer.
  Scheduler: ReduceLROnPlateau with patience of 5 epochs.
- Loss Function: CrossEntropyLoss, ignoring pad tokens.

Training Progress

Duration: 50 epochs over 13 hours, 48 minutes.
Trained on NVIDIA RTX 3090 24GB

Progression of Learning:

Start: Train Loss: 9.14, Val Loss: 8.19.
End: Train Loss: 6.76, Val Loss: 7.32.

Models		ROUGE	BLEU	COSINE
Custom Model	Train	10.12	0.0005	0.1392
	Test	09.31	0.0004	0.1279

Pix2Struct (model, training, outcomes)

Model Summary:

Type: Image-to-text for visual language understanding.

Pretraining: Parses masked web page screenshots into HTML.

Features: Variable-resolution input, integrated language and vision prompts.

Training Details

- Total Parameters: 1.34 billion (All trainable).
- Deployment: Model is running on GPU.
- GPU Memory Allocation: 2.6 GB.

- Learning Rate: Initial rate set to 1e-5.
- Optimizer: AdamW.
- Loss Function: CrossEntropyLoss, ignoring pad tokens.

Training Progress

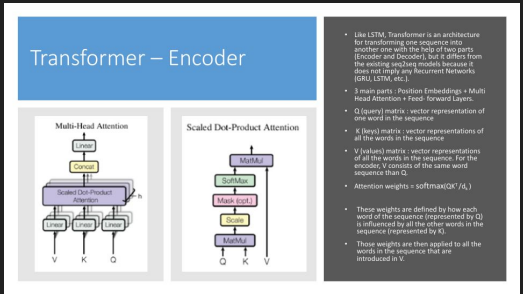
Duration: 5 epochs over 3 days, 16 hours, 32 minutes.
Trained on NVIDIA RTX 3090 24GB

Progression of Learning:

Epoch 1: Train Loss: 0.264, Val Loss: 2.256, Duration: 17:46:00.
Epoch 2: Train Loss: 0.237, Val Loss: 2.174, Duration: 1 day, 8:48:49.
Epoch 3: Train Loss: 0.227, Val Loss: 2.111, Duration: 2 days, 3:10:28.
Epoch 4: Train Loss: 0.220, Val Loss: 2.075, Duration: 3 days, 0:18:50.
Epoch 5: Train Loss: 0.214, Val Loss: 2.058, Duration: 3 days, 16:32:13.

Models		ROUGE	BLEU	COSINE
Pix2Struct	Train	35.25	0.0981	0.4002
	Test	35.13	0.0889	0.3994

LlaVA vs BLIP vs Custom Model vs Pix2Struct - Eval on Training Data



Week_14_page_021

Q: Is Transformer a type of Recurrent Network?

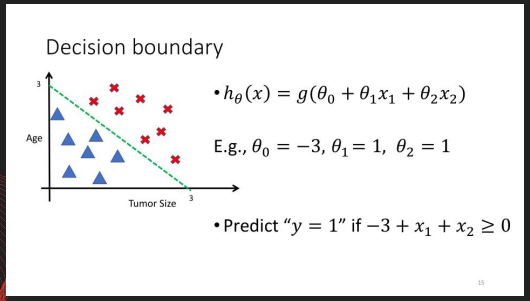
GT: No, Transformer is not a type of Recurrent Network.

LlaVA: No, Transformer is not a type of Recurrent Network.

BLIP: It emphasizes that it focuses on simplifying the tree, leading to more accurate and reliable predictions

CM: It's a form of Recurrent Neural Network. It can be used in any general pattern recognition task, or in image classification tasks. It can be used to find …

Pix2Struct: Yes, Transformer is a type of Recurrent Network.



Week_04_page_016

Q: What is the significance of the dotted line?

GT: The dotted line represents the decision boundary that classifies data points into two categories based on their features.

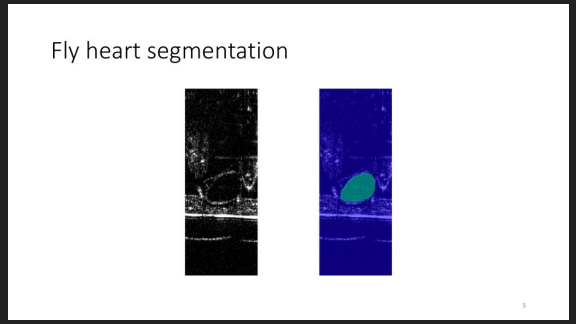
LlaVA: The dotted line represents a boundary or threshold in the data. It might indicate a certain point or category separation.

BLIP: It ensures that the algorithm makes fewer assumptions about the best fit between the predicted

CM: The purpose is to show that the "problem" with the JV is in the way the student is acting. The squares are not any kind of indication of the direction …

Pix2Struct: The dotted line represents the percentage of the tumor size that is predicted by the decision boundary.

LlaVA vs BLIP vs Custom Model vs Pix2Struct - Eval on Training Data



Week_01_page_006

Q: Can you explain the slide?

GT: And we also have applications. We rely see the he's the animals like here in just about fly heart segmentation. The purpose of here is to try to segment the fly heart area. And then you can see how this area keeps changes. Right?

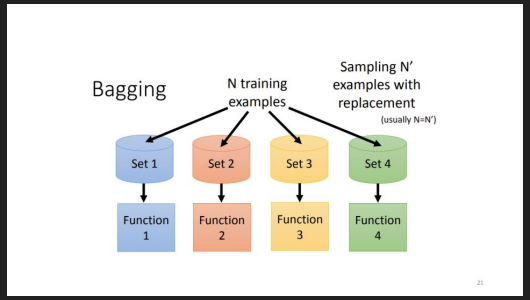
LlaVA: And this is about the fly heart segmentation. So, there are still some dead pixels in the left and right sides. Right? But overall, it's better than the previous one. No?

BLIP: The slide emphasizes the significance of iteratively, but doesn't specify it.

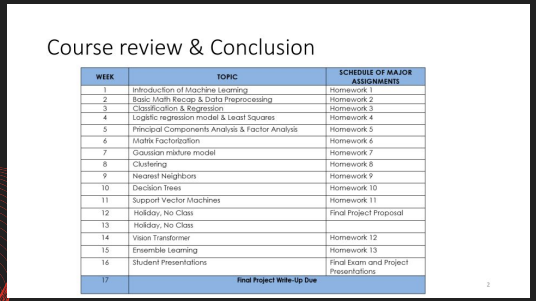
CM: The website explains how they can prevent the post-attack failure of the E.T. results. In the video, the US power grid is shown to be in a bad situation. The worm is described as being a specialized, highly complex and dangerous type of worm. The information is clearly not from the PowerPilot website. …

Pix2Struct: The slide shows fly heart segmentation, showing the relationship between the heart and the heart's diameter.

LlaVA vs BLIP vs Custom Model vs Pix2Struct - Eval on Test Data



Week_15_page_022



Week_16_page_003

Q: How is sampling done in Bagging as per the slide?

GT: In Bagging, sampling is done by taking N' examples with replacement.

LlaVA: Bagging sampling is done post-training and does not require retraining.

BLIP: It ensures that the algorithm predicts the probability of observing the entire sample space

CM: I had to go to the slide's page and type in the 'A' rates, then input the actual values, and then the average. I also find the 'A' data input is a little confusing, …

Pix2Struct: Sampling N' examples with replacement is done as per the slide.

Q: What was the topic of week 10?

GT: The topic of week 10 was decision trees.

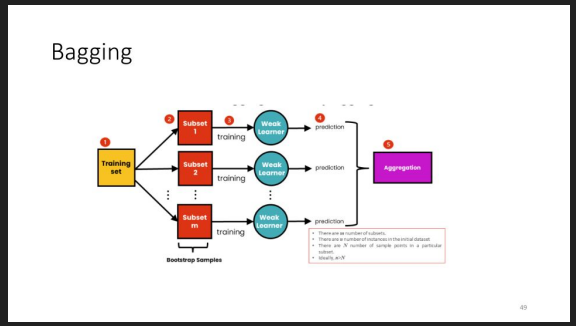
LlaVA: The topic of week 10 was the Homework 1.

BLIP: it emphasizes that iteratively combines probabilities for the final prediction, ensuring it generalize

CM: The topic is a big one. His advice is a big topic. There is a big topic, and the topic is everything. It was a really great week. …

Pix2Struct: The topic of week 10 was Decision Trees.

LlaVA vs BLIP vs Custom Model vs Pix2Struct - Eval on Test Data



Week_15_page_049

Q: Can you explain the slide?

GT: So, let's do it again from this chart again. So, this is the about bagging. Remember, you have a bunch of subsets during the parallel, each one, you have trained on weak learner. So eventually, you are aggregating all these predictions using major development.

LlaVA: So, this is about the bagging. So, we have this red, blue, and green. So, basically, from the red, you will have a new red. If it is red, you will not get anything. If it is blue, you will get a new blue. And if it is green, you will get a new green. And this is so-called the bagging.

BLIP: It emphasizes that the one - vs - all method is pivotal for multiclass problems

CM: The correct answer is: not a good one. What is the correct answer, and is it not the one that is pushed on the slide by the slide-on-Dot-Mover, of which the image in the left-hand slide-on-Dot-Mover slide-on-Dot-Mover…

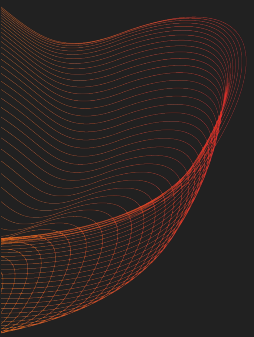
Pix2Struct: The slide explains the importance of aggregation in predicting the training set, and how it can be used.

Results

Best Model		ROUGE	BLEU	COSINE
LlaVA	Train	41.10	0.513	0.42
	Test	35.50	0.497	0.31
BLIP	Train	39.63	0.0	0.39809
	Test	39.681	0.0	0.37385
Custom Model	Train	10.12	0.0005	0.1392
	Test	09.31	0.0004	0.1279
Pix2Struct	Train	35.25	0.0981	0.4002
	Test	35.13	0.0889	0.3994

Model Weights

Model	Location
LlaVA	Huggingface
BLIP	Huggingface
Custom Model	Meta and Huggingface
Pix2Struct	Google



Future Work

1. Expanding the dataset size.
2. Extending the training duration.
3. Optimizing the hyperparameters like temperature, top-p, top-k etc.
4. Enhancing data cleaning process.
5. Securing additional resource for training.

Conclusion

- Upon evaluating the four models - LLaVA, BLIP, Custom Model and Pix2Struct - it's evident that LLaVA is leading in performance. By further optimizing its hyperparameters, we anticipate a notable enhancement in its effectiveness.

- Additionally, the performance metrics suggest that extended training durations could significantly improve the capabilities of all models.

- Lastly, the results clearly indicate the necessity for additional fine-tuning, particularly for the specific task of Visual Question Answering (VQA). This targeted approach will likely yield substantial improvements in our model's accuracy and reliability.

Thank you for your time and attention 🙂