S9824

Surpassing State-of-the-Art VQA with Deep Learning Optimization Techniques under Limited GPU Resources

Quang D. Tran
Head of AI, AIOZ Pte Ltd

Erman Tjiputra
CEO, AIOZ Pte Ltd

AIOZ


AIOZ

INTRODUCTION

Photo credit:

Vietnamtourism

Concept credit:

Devi Parikh

Georgia Tech

*The **kids** are* *watching* *an **old master*** *writing letters.*

It is Tet holiday in Vietnam with warm and fragrant floral atmosphere.

The kids are very attentive and eager to wait for the old master drawing the traditional words.

Q: How many people are there?

A: **5**

Q: What is the old man doing?

A: **Writing**

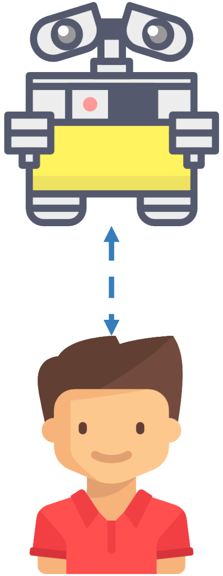
Q: Where is it?

A: **On street**

Human: What a nice picture! What event is this?
AI: *It is Tet holiday in Vietnam.*
You can see lots of flowers and the atmosphere is pretty warm.

Human: Wow, that’s great. What are they doing?
AI: *The kids are watching an old master drawing the traditional letters.*

Human: Awesome, what are the kids wearing?
AI: *It is Ao Dai. a Vietnamese traditional clothes.*
…

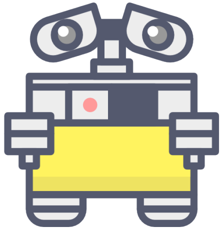


AIOZ

Vision

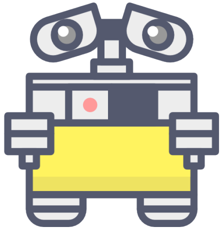
AI See

AIOZ



AI Understand

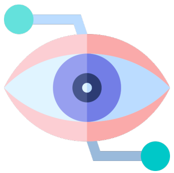
AIOZ



AI Reasoning

Words & Pictures

- Vision → Visual stream → Pictures
- Language → Text/Speech → Words



- *Pictures are everywhere*
- *Words are how we communicate*

Measuring & demonstrating AI capabilities

- *Image Understanding*
- *Language Understanding*

Words & Pictures

- Beyond visual recognition
- Language is compositional

"Two steeds are racing against two brave little dogs."

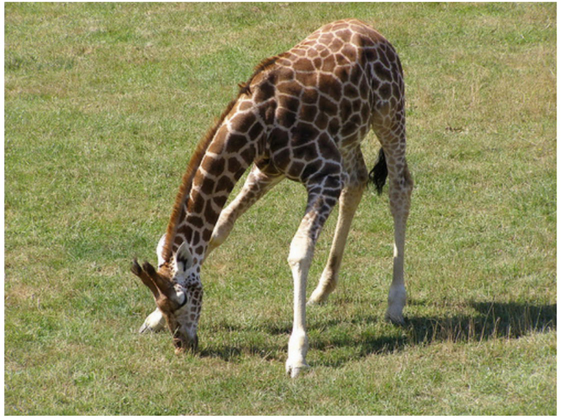
AIOZ

Image Captioning

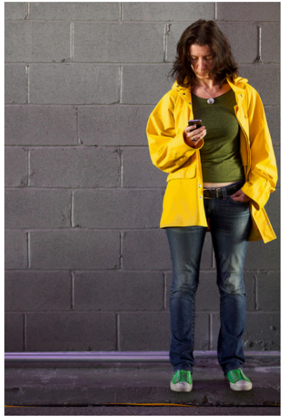
- Image captions tend to be **generic**
- Coarse understanding of image + simple language models can suffice
- **Passive**

a living room with a couch and a tv
logprob: -7.28

a baseball player swinging a bat at a ball
logprob: -4.84



a giraffe standing in a field of grass
logprob: -7.43

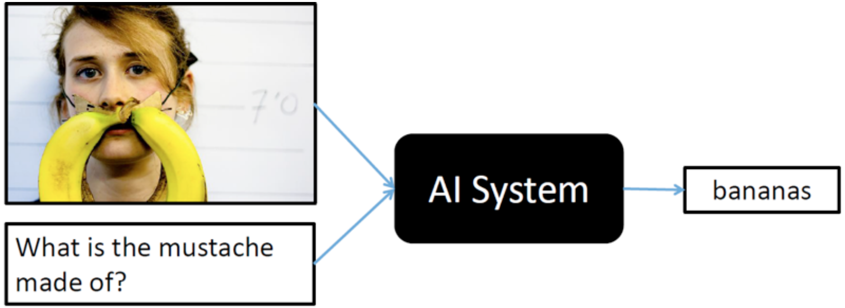


a man is holding a tennis racket in his hand
logprob: -8.90

a clock on a pole in front of a building
logprob: -8.14

Credit by: Karpathy (Stanford)

Introduction:
Visual Quesition Answering (VQA)



- Input = {Image/Video, Question}
- Output = Answer
- Question: asking on the detail of corresponding image
- Question types: Yes/No, Counting, Multi-Choices, Others.
- **Dataset:**
  - VQA-1.0, VQA-2.0, TDIUC, DAQUAR, Visual Genome, Visual-7W, Flickr-30, etc.

Visual Question Answering

"When a person understands a story, [they] can demonstrate [their] understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding."

- Wendy Lehnert (PhD, 1977)

Effective use of vast amounts of visual data

Improving Human Computer Interaction

Challenging multi-modal AI research problem

Visual Question Answering

Credit by: https://visualqa.org

- Details of the image
- Common sense + knowledge base
- Task-driven
- Holy-grail of semantic image understanding

Introduction:
Visual Quesition Answering (VQA)

- VQA on Image uses a image-question pair with answer label as an example → **Supervised Learning**

- Each answer is belonged to a predefined list → **A classifier task**

- Features are extracted from both image & question to determine answer → An **intersection** of Computer Vision & NLP

AIOZ

Introduction:
Visual Quesition Answering (VQA)

VQA question type

Yes/No Question

Counting Question

Free Style Question

VQA Challenge

Dataset: VQA 1.0 - 2.0

Yash Goyal, et al. *Making the V in VQA Matter…, CVPR 2017*

Aishwarya Agrawal, et al. *VQA: Visual Question Answering, ICCV 2015*

>0.25 million images

~1.1 million questions

~11 million answers

Dataset	Input	All	Yes/No	Number	Other
Real	Question	40.81	67.60	25.77	21.22
	Question + Caption*	57.47	78.97	39.68	44.41
	Question + Image	83.30	95.77	83.39	72.67
Abstract	Question	43.27	66.65	28.52	23.66
	Question + Caption*	54.34	74.70	41.19	40.18
	Question + Image	87.49	95.96	95.04	75.33

Human Performance

VQA Challenge: Leaderboard

VQA Challenge 2018

Organized by: VQA Team

77

Overview | Evaluation | Phases | Submit | My Submissions | Leaderboard | Discussions

Please select from following phases

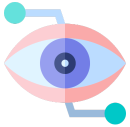
Phase: test2018, **Split:** test-standard

Rank	Participant Team	yes/no	number	other	overall	Last Submission at
1	**AIOZ**	87.96	54.99	63.28	72.61	4 months ago
2	**HDU-UCAS-USYD**	87.97	52.51	63.58	72.49	8 months ago
3	**MSRA-MSM**	87.17	55.19	62.56	71.96	1 month ago
4	**casia_iva**	86.98	51.05	62.31	71.31	9 months ago
5	**Tohoku CV Lab**	87.29	53.25	61.13	71.12	10 months ago

VQA
General Solution & Targets

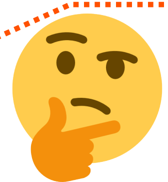
Modern approach for VQA task usually includes 4 main steps:

1. Feature Extraction
2. Joint Semantic Representation
3. Attention Mechanism
4. VQA Classifier

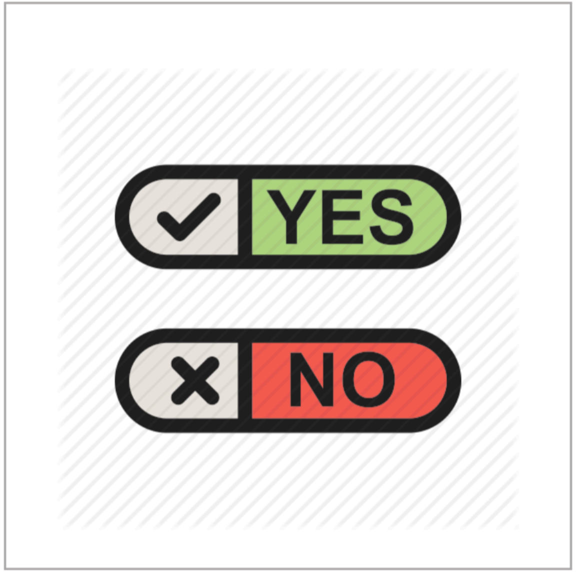



Resources Optimization

Accuracy → Ensemble models



YES/NO QUESTION DEALER



- Bias reduction.
- Attention mechanism improvement.

COUNTING QUESTION DEALER

- Counter module.
- Attention mechanism improvement.

FREE STYLE QUESTION DEALER

- Transfer learning.
- Attention mechanism improvement.

VQA Challenges

Question Identification and Model Combination

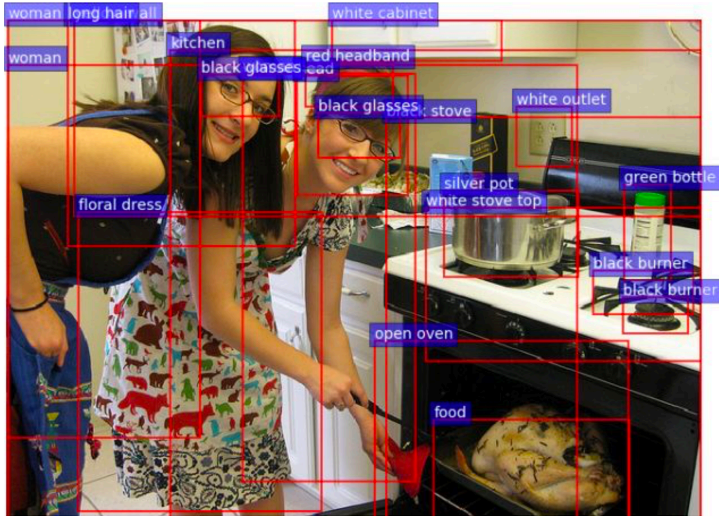
- Ensemble.
- Voting question type.

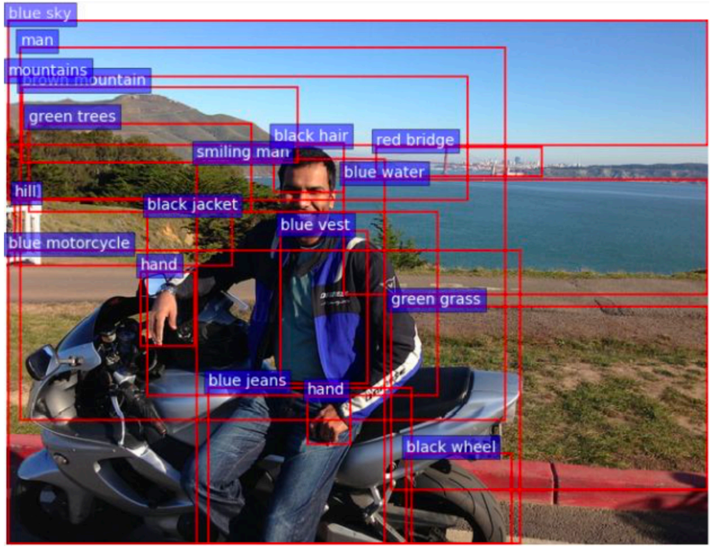
VQA Decomposition

VQA Feature Extraction
Visual & Question Embedding

- **Visual Feature:** Apply Bottom-Up attention
  - Use Faster RCNN to get candidate objects & their bounding boxes.
  - Use ResNet-101 to extract features to get final vector $V = \{V_1, V_2, \ldots, V_K\}$ with **K** is number of proposals.
    In this step, we find out that **K**, number of object proposals , plays an important role in increasing overall performance.

- **Question Feature:** Inherit from GloVe.





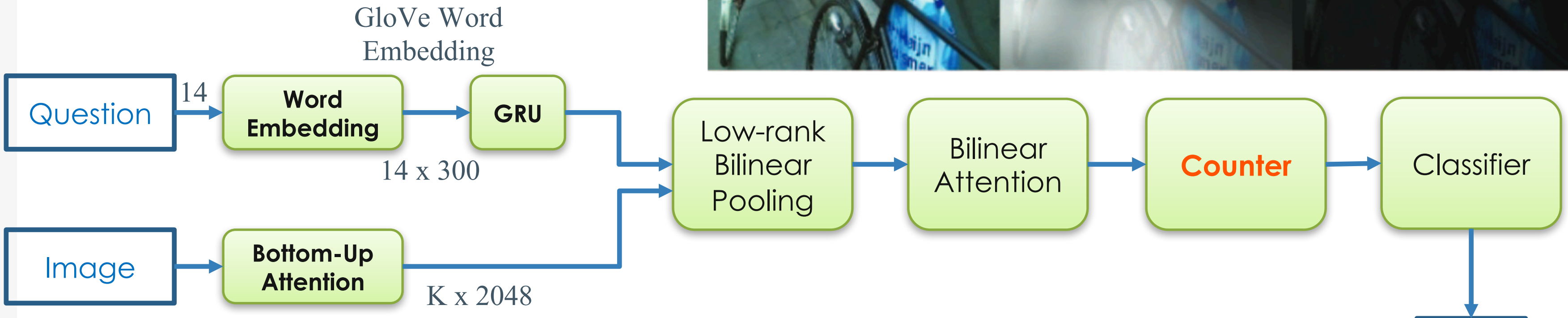
Reference: *Bottom-Up and Top-Down Attention, CVPR 2018*

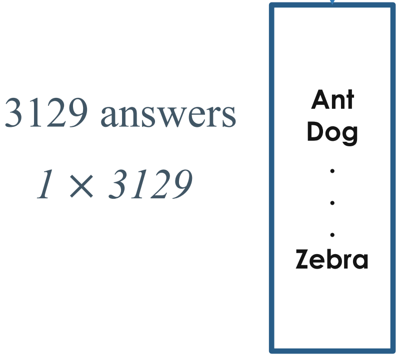
VQA Feature Extraction
Visual & Question Embedding

- K proposals = 50 is proved to be better in increasing performance.
- K value affects the number of bounding boxes that we store → reducing **K** would help decrease resource consumption and training time.

References model	Criteria		
	Over-all	memory (MiB)	time train(h)
30 boxes	68.65	5782	4.65
40 boxes	68.94	6382	5.74
50 boxes	***69.12***	7134	6.06
60 boxes	69.07	8656	6.5
70 boxes	69.03	10594	6.91
100 boxes	69.11	11308	6.93

VQA
Attention Mechanism





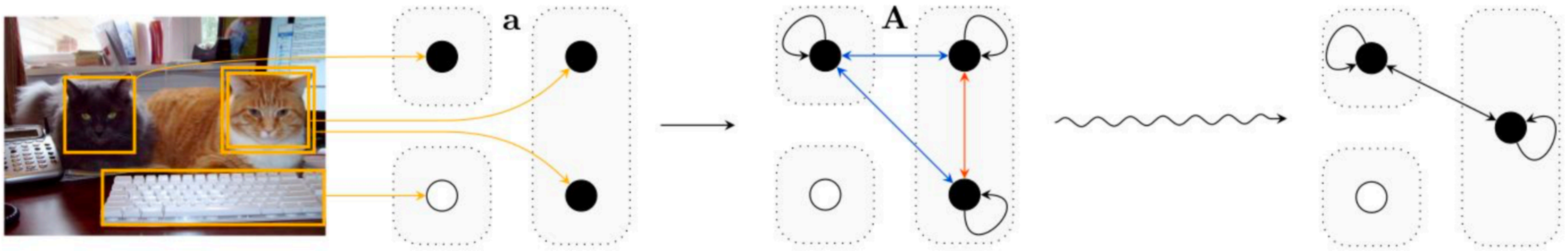
Bilinear Attention Network (BAN)

- Inspired from Co-attention mechanism [2]
- Find bilinear attention distribution
  → consider interaction among 2 groups of input channels
- High resource consumption: **using 4 GPUs**

[1] Jiasen Lu, et al., *Hierarchical Question-Image Co-Attention*, NIPS 2016
[2] Jin-Hwa Kim, et al. *Bilinear Attention Network,* NIPS 2018

VQA
Counting Module

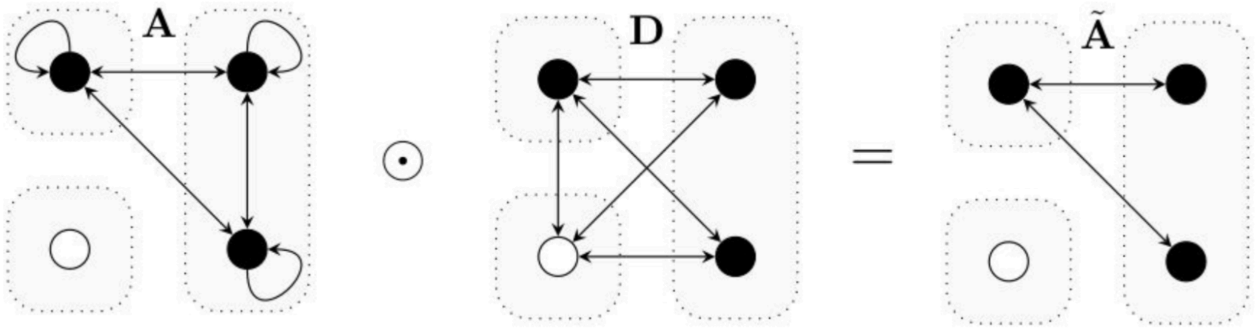
- Turn attention map **(a)** into attention graph ($A = a^T a$) to represent relation between objects.
- Objects have high attention score (black circle) will have connected edge.
- To get count matrix, we eliminate intra-object edges (red edges) and inter-object edges (blue edge)
  → The number of remaining vertices is the count result.



VQA

Counting Module

- To guarantee the objects are fully overlapping or fully distinct we add the normalization function for attention graph **A** and distance matrix **D** before removing intra-object edges and inter-object edges.
- The normalization function: $f(x) = x^{2(1-x)}$
- This function increase the value if it higher than 0.5 and decrease value if it lower than 0.5. The main objective is to widen the distance between low value and high value to make fully distinct or fully overlapping.



VQA
Counting Module

Reference models	VQA score			
	Over-all	Yes/No	Num-bers	Oth-ers
stack-att + counter	68.09	83.14	51.62	58.97
BAN + counter	69.8	85.19	53.38	60.37
BAN + counter + normalize	**69.92**	85.31	**54.06**	60.35

Evaluation Results with proposal counting module

VQA Model Optimization
Activation & Dropout

- Classifier task in VQA is designed to be simple.
  However, it is one of the most important module to improve overall performance.
  → We find out that optimize the only-one activation function in classifier task is important.

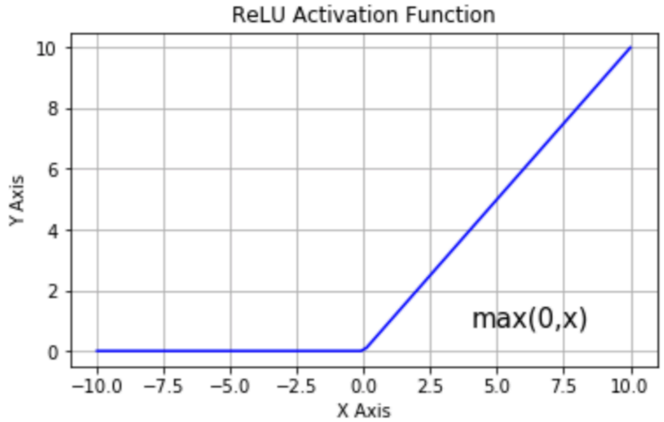
  Thus, we recommend:
  - Change ReLU activation function by another one (e.g., Swish).
  - Change Dropout value to local optimal of the corresponding activation function.

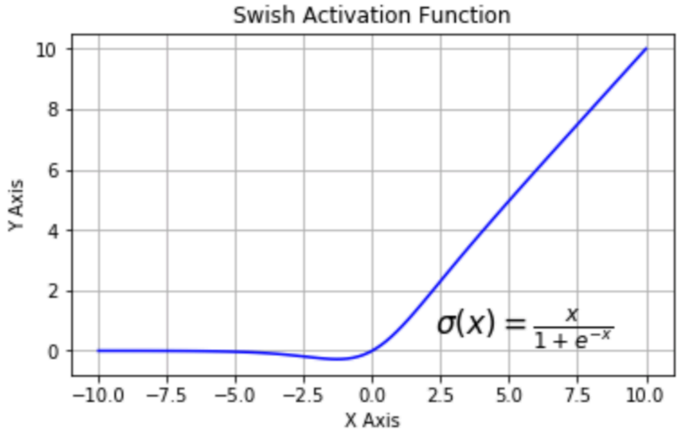
Pros:

- Resolve vanishing gradient problem.
- Provide sparsity in representation.
- Simple to implement.

Cons: No derivative at zero point.

VQA Classifer





Activation Function

$$ReLU(x) = \begin{cases} x \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases} = max(0, x)$$

changes all negative values into zero, this function helps to reduce the effect of weak features.

$$Swish = x.Sigmoid(x) = \frac{x}{1 + e^{-x}}$$

retains a certain amount of information when x < 0 to avoid losing too much information.

Reference models	VQA score in test-dev			
	Over-all	Yes/No	Num-bers	Oth-ers
Base BAN	69.59	85.00	53.25	60.11
Delayed updates	69.69	85.06	53.91	60.13
Baseline + Swish + drop 0.45	69.8	85.19	53.38	60.37
Baseline + Swish + drop 0.45 + norm count	**69.92**	**85.31**	**54.06**	**60.35**

Ensemble Method

Ensemble Method

Ensemble - Voting

Idea: Try to meet agreement of all models in predicting answer.

Model 1	Model 2	Model 3	Model 4	Model 5	Final rating
5	4	5	4	4	4

Ensemble Method

Proposal

- **Step 1:** Train member models for ensembling
- **Step 2:** Get prediction answer with each member model
- **Step 3:** Predict question type based on A-Q map learnt from data
- **Step 4:** Re-voting answer
- **Step 5:** return final ensemble model

Nums of ens	VQA score in test-dev				test-standard
	Overall	Yes/No	Numbers	Others	Overall
No	69.92	85.31	54.06	60.35	70.28
5	70.99	86.47	54.44	61.55	71.40
10	**71.2**	**86.57**	**55.07**	**61.72**	**71.53**

Algorithm 1: Question type voting method

Input :
memPreds: answers with confidence scores on input of each member model used for ensemble
mapping: answer-to-question-type mapping
Output:
qTypes: list of question types for questions in input data

```
qTypes ← Null
for answerPos in sizeOf(memPreds[0]) do
    voteList ← Null
    for modelPos in memPreds[1] do
        voteList ←
          memPreds[modelPos][answerPos]
    end
    freqDict ←
      countDuplicateAnswer(voteList)
    sumVoteDict ← sumVote(freqDict)
    qTypes ←
      typeVote(sumVoteDict, mapping)
end
return qTypes
```

Ensemble Method
Pros & Cons of Voting

Pros:

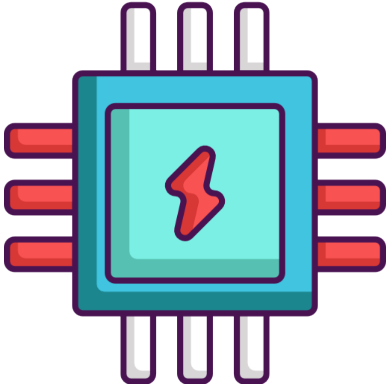
- Simple & easy to implement
- No architecture restriction
  → Identify question-type without training a classification model
- Reduce bias
- Maximize the performance of each model trained for specific question type

Cons:

- Useless when the number of voting is equal
- No emphasis in any specific good models

Resource Consumption Optimization

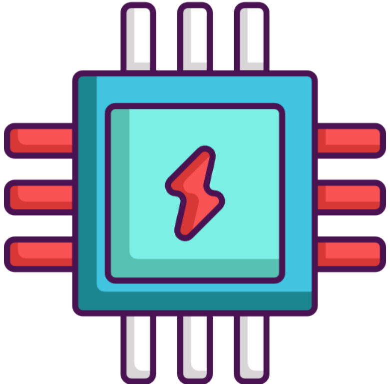
Resource Consumption Optimization



Processing Power

Computing Speed

Resource Consumption Optimization



- Fast half precision floating point (FP-16) for Deep Learning Training
- Delayed Updates (Gradient accumulating)

Resource Consumption Optimization
Mixed Precision Training

- ML models are usually trained in FP-32.
  - FP-64 (Double precision): expensive but high accuracy.
  - FP-32 (Single precision): less expensive also less accuracy.
  - FP-16 (Half precision): cheap but low accuracy.

- ML rule of thumb:
  - Balance of **speed** & **accuracy**.

- **Expectation:**
  "running with FP-16 while having comparable accuracy to FP-32"

AIOZ

Resource Consumption Optimization
Mixed Precision Training

Solution

- Baidu Research & NVIDIA has successfully trained FP-16 with accuracy comparable to FP-32, 2x speed up and reduced 1.5 times memory consumed.
- Reference: Paulius et al., *Mixed Precision Training*, ICLR 2018.

Pros

- Speed up training progress
- Training with larger model

Model	test-dev score	memory (MiB)	time train(h)
FP32 (1)	69.20	10250	5.78
Mixed precision(1)	69.22	9366	4.30
FP32 (2)	69.44	8352	16.89
Mixed precision(2)	69.50	6174	14.80
FP32 (3)	69.60	11012	6.62
Mixed precision(3)	69.57	8932	4.92

Resource Consumption Optimization
Delayed Updates

- Reference: Myle et al., *Scaling Neural Machine Translation*, ACL 2018

- We divide entire data into mini-batches. Do forward (compute outputs) and backward (compute gradients based on loss), then updating parameters (learning) on each mini-batch.

Model	**VQA score in test-dev**			
	Over-all	**Yes/ No**	**Num-bers**	**Oth-ers**
Batch 256	70.04	85.42	54.04	60.52
Batch 32, freq 8	69.91	85.19	52.5	60.82
Batch 32, freq 16, 2x lr	69.93	85.49	53.86	60.27
Batch 32, freq 16, 3x lr	69.58	85.19	52.9	60.03
Batch 16, freq 16	69.82	85.3	53.22	60.35
Batch 16, freq 32, 2x lr	69.79	85.21	53.23	60.36

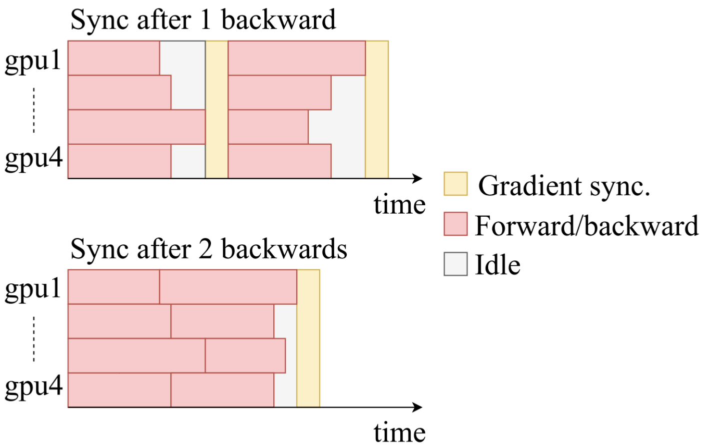
Evaluation results of delayed updates technique.

AIOZ

Resource Consumption Optimization
Delayed Updates

- **Problem**: When training a ML model based on a kind of gradient descent optimizers
  → Batch size must be considered carefully.
  - **Large batch size** → Fast Training
    → **Large memory usage**

- **Solution**: **Delayed updates (gradient accumulating)** is a technique that aims to deal with this limitation of memory usage.
  - As usual: *1 forward, backward - 1 update*
  - Delayed updates: ***N forward-backwards - 1 update***



- **Result**: With delayed updates, model is trained with a batch size equals to N times of itself. In example:
  - *With batch size 32, we can simulate running model with batch size 256 by setting $N = 8$ $(256 / 32 = 8)$.*

AIOZ

Knowledge Distillation

Knowledge Distillation

Introduction

The best results are usually achieved with:

- **Ensemble Models:** Use multiple learning algorithms to obtain better predictive performance.
- **Large Networks:** Use complicated and deep model to obtain better performance.

However: **Time** and **cost** of running inference in these machine learning model **are high** which make learning is hard to apply on embedded system.

Solution: Knowledge distillation learning which **distill latent knowledge** of these models into a **lighten** model and **minimize** the **shrink** of performance.

Knowledge Distillation
Latent Knowledge

- A classification function is **a labelling** function which map the representation of input to output.

- Theses representation help determine which class the given input is belonged based on computed **distribution over classes**.

- The distribution over classes (or the logits before this distribution) may contain **latent knowledge** extracted from input representation.

Knowledge Distillation
Latent Knowledge

- Ensemble models or large networks, which contain great latent knowledge, are called **teacher models**.
- Lighten model is called **student model**.
- Latent knowledge in teacher models effects performance of the student model through **loss function** and **logits** of teachers.
- **Soft logits** can improve convergence speed of student.
- **Temperature** hyper-parameter **T** can soften teachers' logits.

Knowledge Distillation

1. Train parent network from scratch and **get logits**.
2. Train student with **dual goal**:
   - **Predicting** the correct labels
   - **Matching** the output **distribution** of the teacher through distillation loss function.

Knowledge Distillation

The **softened targets** of the student and the teacher are the **probabilities** over classes computed by converting pre-softmax logits use the equation belows with **temperature T**:

$$Q_i^{\tau} = \frac{exp(l_i/T)}{\sum exp(l_i/T}$$

Knowledge Distillation

Loss function:

$$\mathcal{L}_{KD} = \alpha T^2 \mathcal{L}_{CE}(Q_S^\tau, Q_T^\tau) + (1-\alpha)\mathcal{L}_{CE}(Q_S, y_{true})$$

where $\mathcal{L}_{CE}$: **cross entropy** loss

Q_T^τ, Q_S^τ : the **softened targets** of the teacher and the student using the same temperature parameter T (T > 1).

α : a **control hyper-parameter** which effects directly to two components of the loss.

Knowledge Distillation
Open Applications

- In **Compact Networks**: Knowledge distillation help **distill** knowledge got from **large** architecture into a **lighten** model.
- In **Visual Question Answering (VQA)**: Knowledge distillation help **distill** knowledge got from trilinear modelling into bilinear modelling.
  - **Trilinear modelling**: high performed model however can not be used as inference in Free-form answer VQA.
  - **Bilinear modelling**: lighten and can be used for both training and testing. Performance is lower when comparing with trilinear modelling.

Summary

What did we discuss?

Components of a VQA Framework

Tatics for VQA Accuracy Improvement

Overcoming Limited Hardware Resources

Compact VQA for real-life deployment

Demo

VQA – Identity Recognition & Its Applications


Name & Gender
LE MANH TRUONG
Nationality
VNM
Document
Passport
Date of Birth
06/06/1996
Issued in
VNM
Expiry Date
17/03/2028
ID
C4780528
Confirm

Future Application

VQA – Potential Real-life Applications


AIOZ

Thank you for your listening!