Visual Question Answering

Nemanja Vujadinović & Tina Mihajlović

mentors: Vladimir Jovanović, Stefan Mojsilović et al.

What will we cover today?

- What is VQA?
- Why we “need” it?
- Methodology
- Dataset
- Training
- Results
- Demo
- You either win or learn (or both)
- Won a battle, but how to win the war?

What is VQA?

Q: Is this man the Olympics 2024 gold medalist?

VQA model

A:Yes! 🐐

Why we “need” this?

Assistive technology

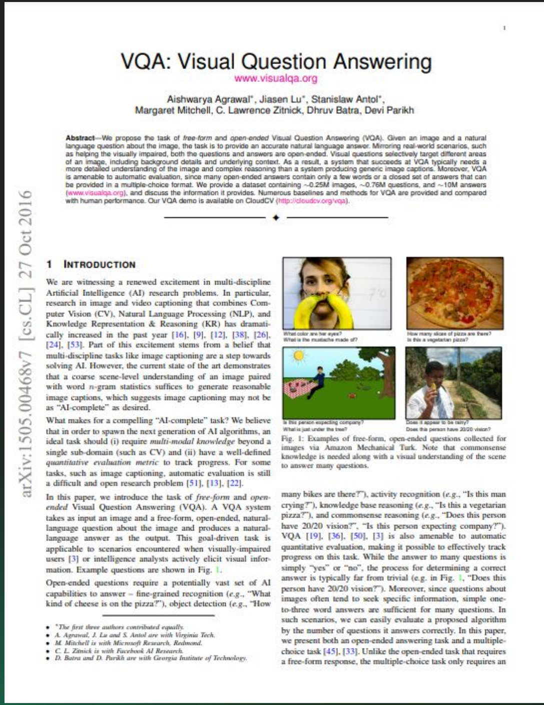
Interactive Assistants

Interactive Learning

Why we “need” this?

It’s super cool!

Methodology



VQA: Visual Question Answering

www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

Abstract—We propose the task of *free-form* and *open-ended* Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing ~0.25M images, ~0.76M questions, and ~10M answers (www.visualqa.org), and discuss the information it provides. Numerous baselines and methods for VQA are provided and compared with human performance. Our VQA demo is available on CloudCV (http://cloudcv.org/vqa).

arXiv:1505.00468v7 [cs.CL] 27 Oct 2016

1 Introduction

We are witnessing a renewed excitement in multi-discipline Artificial Intelligence (AI) research problems. In particular, research in image and video captioning that combines Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation & Reasoning (KR) has dramatically increased in the past year [16], [9], [12], [38], [26], [24], [53]. Part of this excitement stems from a belief that multi-discipline tasks like image captioning are a step towards solving AI. However, the current state of the art demonstrates that a coarse scene-level understanding of an image paired with word n-gram statistics suffices to generate reasonable image captions, which suggests image captioning may not be as "AI-complete" as desired.

What makes for a compelling "AI-complete" task? We believe that in order to spawn the next generation of AI algorithms, an ideal task should (i) require *multi-modal knowledge* beyond a single sub-domain (such as CV) and (ii) have a well-defined *quantitative evaluation metric* to track progress. For some tasks, such as image captioning, automatic evaluation is still a difficult and open research problem [51], [13], [22].

In this paper, we introduce the task of *free-form* and *open-ended* Visual Question Answering (VQA). A VQA system takes as input an image and a free-form, open-ended, natural-language question about the image and produces a natural-language answer as the output. This goal-driven task is applicable to scenarios encountered when visually-impaired users [3] or intelligence analysts actively elicit visual information. Example questions are shown in Fig. 1.

Open-ended questions require a potentially vast set of AI capabilities to answer – fine-grained recognition (*e.g.*, "What kind of cheese is on the pizza?"), object detection (*e.g.*, "How many bikes are there?"), activity recognition (*e.g.*, "Is this man crying?"), knowledge base reasoning (*e.g.*, "Is this a vegetarian pizza?"), and commonsense reasoning (*e.g.*, "Does this person have 20/20 vision?", "Is this person expecting company?").

VQA [19], [36], [50], [3] is also amenable to automatic quantitative evaluation, making it possible to effectively track progress on this task. While the answer to many questions is simply "yes" or "no", the process for determining a correct answer is typically far from trivial (e.g. in Fig. 1, "Does this person have 20/20 vision?"). Moreover, since questions about images often tend to seek specific information, simple one-to-three word answers are sufficient for many questions. In such scenarios, we can easily evaluate a proposed algorithm by the number of questions it answers correctly. In this paper, we present both an open-ended answering task and a multiple-choice task [45], [33]. Unlike the open-ended task that requires a free-form response, the multiple-choice task only requires an

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

- *The first three authors contributed equally.
- A. Agrawal, J. Lu and S. Antol are with Virginia Tech.
- M. Mitchell is with Microsoft Research, Redmond.
- C. L. Zitnick is with Facebook AI Research.
- D. Batra and D. Parikh are with Georgia Institute of Technology.

+

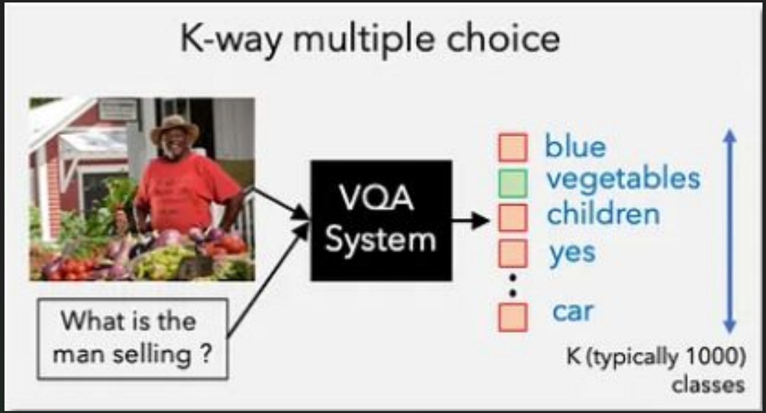
+

lesson

Start small :)

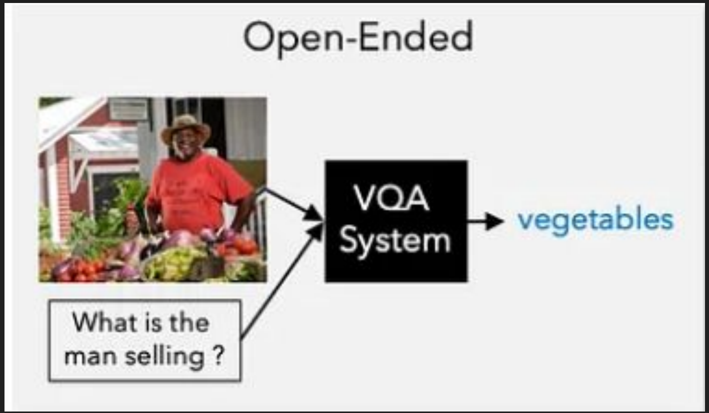
Types of VQA models

Discriminative



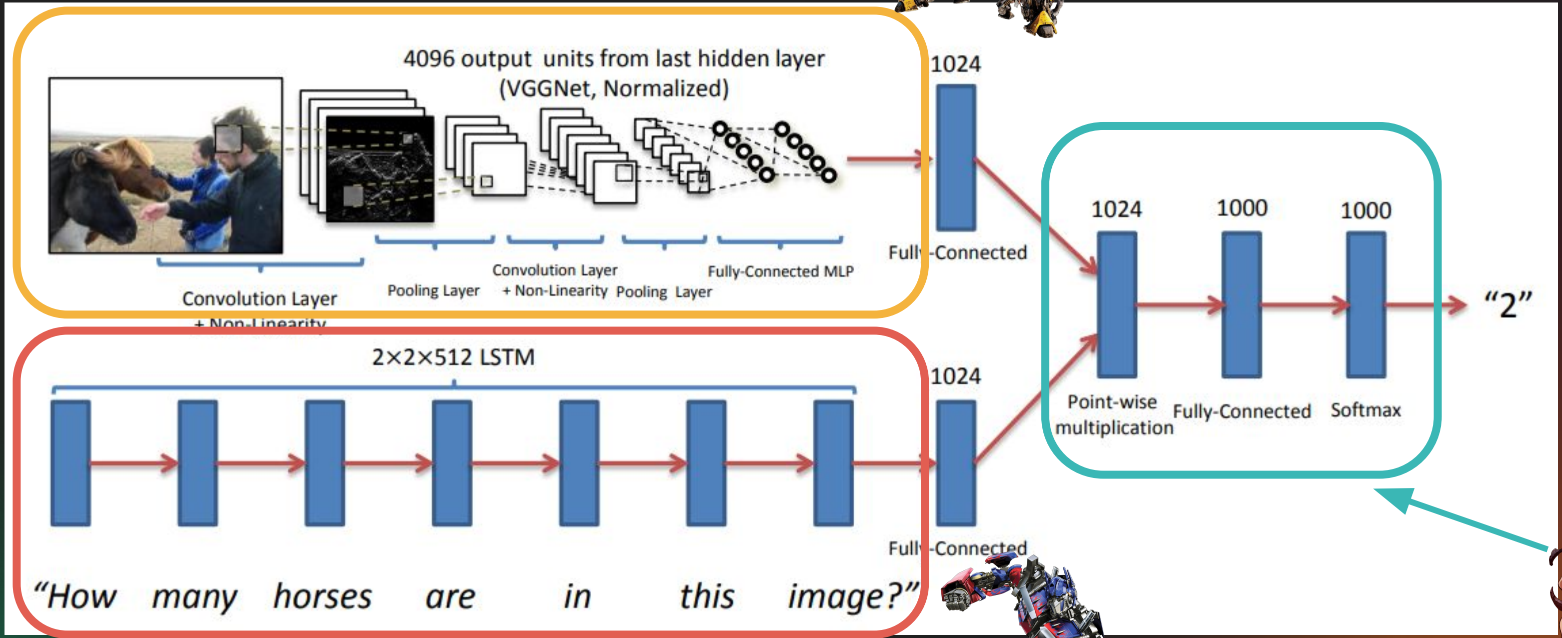
Pros: simple, efficient, consistent
Cons: restricted, dependent

Generative SOON



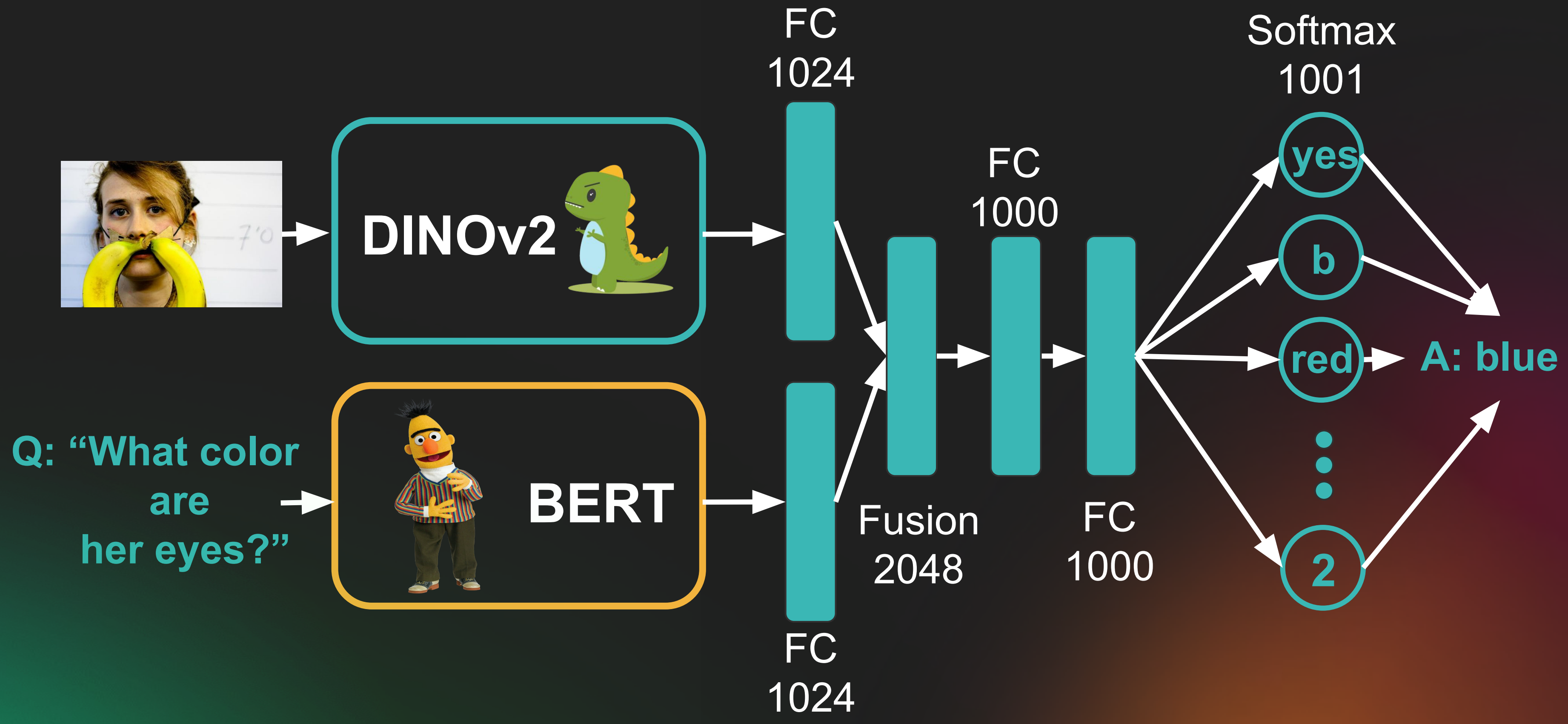
Pros: flexible, expressive
Cons: complex, non reliable

Original VQA paper



Goal: Beat the original paper (accuracy ~58%)

Our proposed architecture



Dataset

- VQA v2 dataset
- Splits
  - Training set - 82783 images, 447028 questions, 4470282 answers
  - Validation set - 40504 images, 218721 questions, 2187210 answers
  - Test set - 81434 images, 439744 questions, 4397440 answers

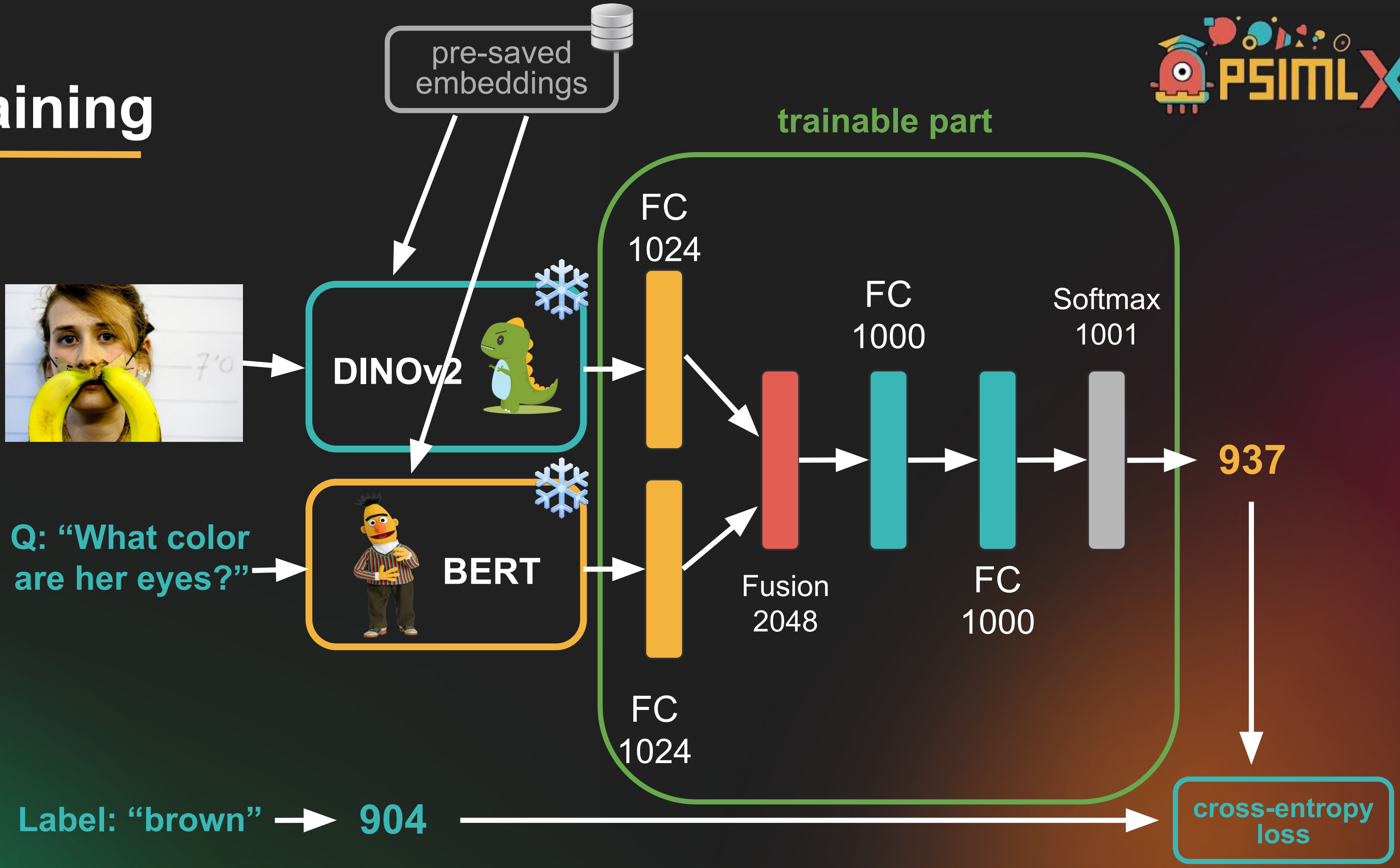
lesson

Know your data!

Dataset

- Selecting the best answer to question
  - Most frequent answer
  - Levenshtein distance
- 1000 most frequent answers
  - 87% of answers in training and validation sets
  - Other answers labeled as <unknown>

Training


pre-saved
embeddings
trainable part
DINOv2
Q: “What color
are her eyes?”
BERT
FC
1024
FC
1024
Fusion
2048
FC
1000
FC
1000
Softmax
1001
937
Label: “brown”
904
cross-entropy
loss

Training… it's a process

- Iteration 1*:
  - batch size: 2048
  - learning rate: 1e-3
- Results 1:
  - 70% accuracy on validation set
  - Model outputs gibberish answers??
- Debugging 1:
  - 5 hours, 5 mentors and inf tears later…
- Lesson 1:
  - Line by line debugging is your best friend
  - If the output doesn't make sense, your code doesn't

Training… it's a process

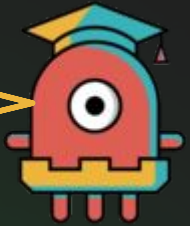
- Iteration 2:
  - batch size: 2048
  - learning rate: 1e-3
- Results 2:
  - 48% accuracy on validation set
  - Stagnates in the first 150 epochs
- Debugging 2:
  - Overfitting diagnosed (70%+ accuracy on training set)
- Lesson 2:
  - Use regularization methods (L1, L2, Dropout)
  - Use less complex network

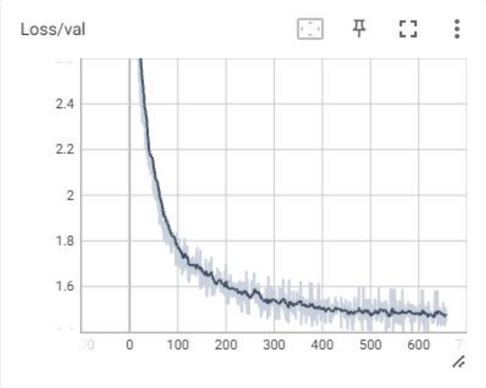
Final results

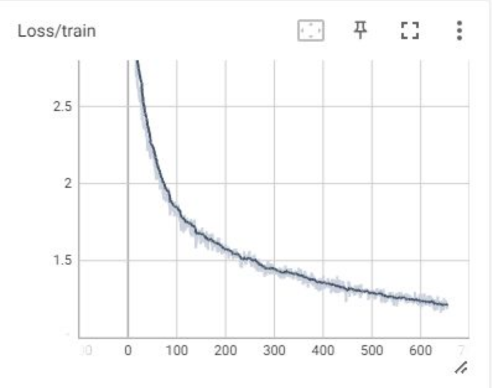
- Final iteration:
  - batch size: 3072
  - learning rate: 1e-5
- 51.07% top 1 answer accuracy (< 58%)
- 88.35% top 5 answer accuracy

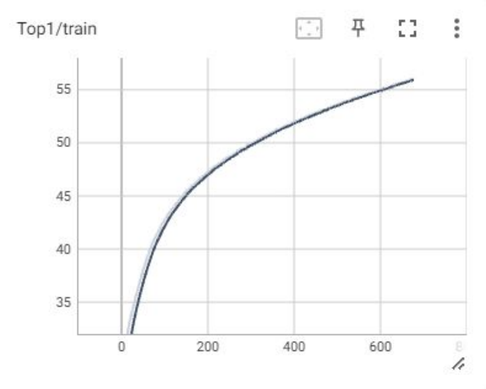
lesson

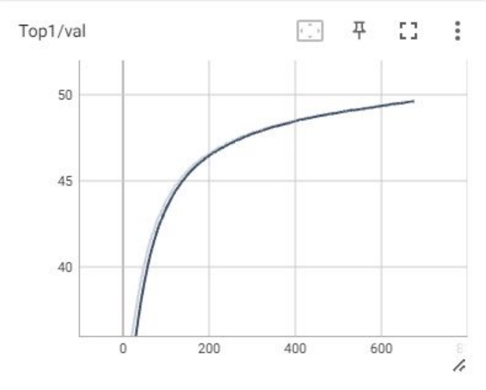
Trial and error == ML











Demo

New chat

User:

What color is the kid's hair?

3:49 am

VQA Model:

Blonde

3:49 am

Demo

VQA Model

Labels:	Blonde	Brown	Red	Yellow	Blue
Probs:	0.9118	0.07	0.009	0.014	0.001

ANSWER

Demo

New chat

User:

How many people are in the picture?

3:49 am

VQA Model:

2

3:49 am

Demo

VQA Model

Labels:	2	3	4	5	1
Probs:	0.3446	0.25	0.172	0.06	0.04

ANSWER

You either win or learn (or both)

- Start small, build from there
- Purposefully overfit the model on a single sample to confirm its correctness
- Know your data
- Debugging is your best friend
- Know your basics
- Sometimes it simply works… and sometimes it simply doesn’t
    - examine, hypothesize, implement 🔁

Won a battle… what's next?

- Up the current accuracy
    - Regularization
    - Experiment with different fusions (MFB, cross attention)
- Try encoding the answers with BERT (Danda suggestion)
- Classification -> Generation

Thank you for attention!

Questions? ✋?