gain understanding

SURPRISE

—

EVALUATION METRICS

(1)

Self-guided, curiosity driven agents which learn from the ground-up using intrinsic motivation

BS EVALUTIVE AI

(BS)

→ Build new metrics that correlate with human preferences / Perception

---

We focus too much on low-level automatic metrics, which often have little to do with the original project goals

—

how?
As we are better in recognition than generation + have all these big datasets, we can fall back to retrieval ~~training~~ to get interesting artefacts

How to get open ended creativity !!

[Evaluation Bluesky]

human-in-the-loop modelling.

Use functional programming for evaluation

a universal meter / toolbox /
Standardised evaluation process for every creative system creator to follow.

| BLUE SKY

HOW
Perceptual Loss metrics

Semantic Difference

for questionaiVe

Property model
Intrinsic Motivation
+
Embodiment
+
Experience /BS
||||||

How To?

first up simple

negative log - likelihood
loss functions / framework
↑
and customize the loss
function / framework.

⊙ Beyond maximizing
the likelihood &
the training data?
[BS]

HOW
① Use hierarchical
representations...
② Combine knowledge
with deep
graph
neural features

[How]
Diversity metrics
like INCEPTION SCORE

Green note:

HOW

• Relax parameters

Red note (top right):

) Eval Metrics:

Metrics define selection from search, so very imp to choose the right one.

Wrong metrics will generate bad outputs.

Yellow note (middle left):

HOW?

Advance the field by critically reviewing evaluation methods of research before publication

(evaluation strategies are often inconsistent w/ other literature, or measure the wrong things)

Red note (middle right):

GUIDE AND /OR CONSTRAIN THE GENERATIVE PROCESS

Yellow note (bottom left):

It can be the evidence that the procedure is good or bad.

Purple note (bottom right):

Evaluation is important to drive the research community to a clearer, and similar/ unified goal.

e.g. BLEU score in NLP helps advance the field very quickly, after the winning underspends where parameters to work on and improve.

BS/

COMPARE DIFFERENT LATEN SPACES CROSS-DOMAIN

WHY IS THIS **IMPORTANT?**

BETWEEN SUBJECTIVE & OBJECTIVE FUNCTIONS

TO ASSESS SYSTEMS

---

how to create a metric that is bias free?

Evaluation is important based on physiological (questionnaire

---

- I keep ruminating on the **concept** of comprehension in relation to surprise.

- Something easily understood is not a surprise, but when you finally comprehend something difficult, Eureka!

Quality after creation

Why? In order to create networks that we can evaluate as successful we must be able to at least a in some way, describe what creativity is and i think that's a very hard task because it forms so organically, im not sure if what we have is good enough atm

Should the user always be able to provide his/her own parameters or even an entire evaluation function?

(also related to personalization)