Are you doing what I say?
On modalities alignment in ALFRED

Ting-Rui Chiang*  Yi-Ting Yeh*  Ta-Chung Chi  Yau-Shian Wang
Motivation

- There are benchmarks for instruction following.
  - Input: instructions, observation of the world
  - Output: interactions with the environment.
  - Goal: following the instructions.
- Intuition: A model should focus on the instruction it is doing.
- Research question:
  How well do models align the instructions with its interactions?
Task/Dataset: ALFRED

Goal instruction
- Warm a plate and place it on the table.

Step-by-step instructions
- Turn to the right and move towards the range, then turn to the right and move to the dishwasher in front of the window.
- Pick up the gray patterned plate from the counter to the left of the toaster and in front of the window.
- Turn to the left and then move to and face the range.
- ...

Output: action sequence
- turn right, turn right...
- pick [object], ...
- turn left, go forward...
- ...
Basic Model Architecture

Warm

a

plate

and

place

...
Measuring the Alignment

The instruction that action \(i\) corresponds to.

\[
B = \frac{1}{L_s} \sum_{i=1}^{L_s} \mathbb{1}[f(v_i) = f_M(v_i)]
\]

The instruction that model focus on.
- Attention
- Gradient

Instruction 1
1. turn right
2. go forward
3. go forward
4. ...
5. pick(knife)

Instruction 2
6. turn left
7. ...
8. go forward
9. go forward
10. ...
11. ...
12. ...

## Alignment Score

|       | Train |          | Seen |          | Unseen |          |
|-------|-------|----------|------|----------|--------|----------|
|       | Attn  | Grad     | Attn | Grad     | Attn   | Grad     |
| Random|       | 0.290    |      | 0.294    |        | 0.328    |
| Seq2Seq| 0.590 | 0.594    | 0.589| 0.593    | 0.573  | 0.577    |
| MOCA  | 0.337 | 0.366    | 0.341| 0.361    | 0.380  | 0.384    |
Improving the Model with a Program Counter

- Program counter $c$: the instruction to execute
  - Initialize with $\theta$
  - Monotonically incremental - predicted by model at each step.
    \[ c^{(t+1)} = c^{(t)} + \sigma(f_c(h^{(t)}) \]
  - Construct an attention mask
    \[ m_j^{(t)} = \exp\left\{ -\lambda \left| p_{j,\text{instr}}^{(t)} - c^{(t)} \right| \right\} \]

| $m_j^{(t)}$ | 1 | 1 | 1 | 1 | ~0 | ~0 | ~0 | ~0 | ~0 |
|---|---|---|---|---|---|---|---|---|---|
| $p_{\text{instr}}$ | 1 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 |
| Turn | around | and | goto | ... | Pick | up | the | ... |
Auxiliary Loss

| Instruction 1 | Instruction 2 | Instruction 3 | Aux. Loss (L2) |
|---------------|---------------|---------------|----------------|
| 1 turn right  | 6 turn left   | 8 go forward  | \( c_1 = 1 \) |
| 2 go forward  | 5 pick(knife)| 9 go forward  | \( c_2 = 1 \) |
| 3 go forward  |               |               | \( c_3 = 1 \) |
| \( \ldots \)  |               | \( \ldots \)  | \( c_5 = 2 \) |
| 5 pick(knife)|               |               | \( c_6 = 2 \) |
|               |               | 8 go forward  | \( c_8 = 3 \) |
|               |               | 9 go forward  | \( c_9 = 3 \) |
|               |               | \( \ldots \)  |                 |
## Experimental Results: Alignment Score

| Method                | Train Attn | Train Grad | Seen Attn | Seen Grad | Unseen Attn | Unseen Grad |
|-----------------------|------------|------------|-----------|-----------|-------------|-------------|
| Random                | 0.290      |            | 0.294     |           | 0.328       |             |
| Seq2Seq               | 0.590      | 0.594      | 0.589     | 0.593     | 0.573       | 0.577       |
| MOCA                  | 0.443      | 0.382      | 0.450     | 0.384     | 0.436       | 0.348       |
| MOCA + PC w/o loss    | 0.448      | 0.364      | 0.429     | 0.345     | 0.424       | 0.336       |
| MOCA + PC             | **0.813**  | **0.735**  | **0.777** | **0.705** | **0.724**   | **0.646**   |
## Experimental Results: Success Rate

|                | Seen                  | Unseen                |
|----------------|-----------------------|-----------------------|
|                | Task Success Rate     | Goal-Cond             | Task Success Rate | Goal-Cond |
| MOCA           | 19.2                  | 28.5                  | 3.8               | 13.4      |
| MOCA + PC w/o loss | 16.6                  | 25.7                  | 1.7               | 11.7      |
| MOCA + PC      | 19.5                  | 28.9                  | 3.9               | 13.3      |
Contributions

- We identify previous models’ incapability of aligning the modalities.
- We propose a method to improve the alignment.
Please check our paper for more details!