A Data Collection and Analysis

The annotators worked with the following protocol.

1. **Meta:** The caption talks about when, where or how the picture is taken. This category have more fine grained classes, namely, *Where*, *When*, and *How*. Meta is chosen when the caption content mentions time, location, and/or means by which the image was taken or created.

   Examples of *How* include captions that talk about background or mention "portrait", "screenshot", "closeup" or "a view of".

2. **Visible:** The caption is Visible just by looking at the picture. Caption content describes an event or state of being as captured in the image such as "a man is walking".

3. **Action:** The image describes an action. The image depicts an action that is also described in the text.

4. **Subjective:** The caption is a matter of opinion. The interpretation of caption content can be subjective. This includes captions that contain predicates of personal taste such as "the weather is beautiful", "the pie is yummy".

5. **Story:** Text and image work together like story and illustration. This includes captions that serve as instructions, and captions that insert background information or an account of events.

6. **Irrelevant:** Text does not seem to be written by a person in relation to this specific image.

7. **Other**

We have carefully analysed the pairs that are in the *Irrelevant* and *Other* category. Figure 1 presents a number of these examples. Often times, the caption in these pairs are only partially true about the image. The following are a few examples of the comments that annotators left for us:

1. bathroom not bedroom
2. no soccer players
3. there is no skier
4. football team should be volleyball team
5. caption is gibberish
6. a chair in the woods not bridge
7. might not be dead
8. not sure if I can distinctly say visible because of the pronoun
9. no queen but her baton

Top unigram and bigram features of our baseline Naive Bayes SVM for predicting the labels are listed in Table 2.

		Subjective	Action	Story	Meta
Ground Truth	Visible	3.96%	16.71%	8.08%	22.49%
	Subjective		0.72%	2.96%	1.25%
	Action			2.72%	9.13%
	Story				2.89%
		Subjective	Action	Story	Meta
Model	Visible	1.01%	10.62%	9.67%	54.55%
	Subjective		0.00%	1.49%	0.76%
	Action			2.12%	7.96%
	Story				8.06%

Table 1: Rates of co-occurrences of different labels in groundtruth and the model outputs.

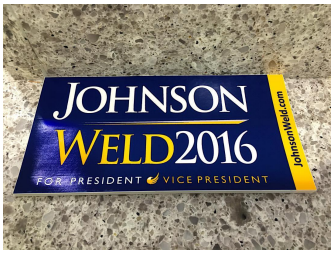
Visible	icon	background	letter
Subjective	perfect	unique	royal
Story	also	find	inspired
Meta	west construction	view night	view inside

Table 2: Top Naive Bayes SVM unigram and bigram features for predicting the labels.

Other: text

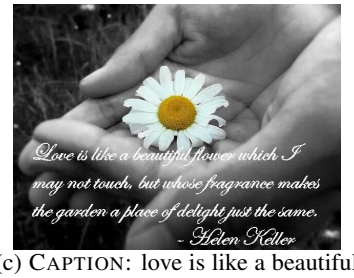
(a) CAPTION: a gardener may water the plant daily but fruits grow only in the season.

Other: text



(b) CAPTION: a sign on the wall.

Other: text



(c) CAPTION: love is like a beautiful flower ...

Other

(d) CAPTION: loft style bedroom with beamed ceiling and gothic window.

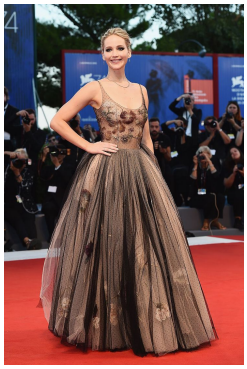
Subjective, Story, Irrelevant, Other

(e) CAPTION: soccer players walking off the field

Irrelevant, Other

(f) CAPTION: a freestyle skier charges down a beautiful snow covered mountain in slow motion.

Other: gibberish



(g) CAPTION: actor in retail at the mother!

Other: gibberish

(h) CAPTION: put all the extraordinary things that are not ordinary, simple things to do everything on that simple.

Story, Irrelevant, Other

(i) CAPTION: the queen's baton carried by people (there is no queen but is that her baton?)

Visible, Irrelevant, Other

(j) CAPTION: sunny woodland morning along river.

Subjective, Other: gibberish

(k) CAPTION: dead mole on the ground

Meta, Irrelevant, Other

(l) CAPTION: a bridge in the woods

Figure 1: Examples of image–caption pairs in the *Other* and *Irrelevant* category.