Situation Recognition with Graph Neural Networks
Supplementary Material

Ruiyu Li[1], Makarand Tapaswi[2], Renjie Liao[2,4], Jiaya Jia[1,3], Raquel Urtasun[2,4,5], Sanja Fidler[2,5]

[1]The Chinese University of Hong Kong, [2]University of Toronto, [3]Youtu Lab, Tencent
[4]Uber Advanced Technologies Group, [5]Vector Institute

ryli@cse.cuhk.edu.hk, {makarand,rjliao,urtasun,fidler}@cs.toronto.edu, leojia9@gmail.com

We present additional analysis and results of our approach in the supplementary material. First, we analyze the verb prediction performance in Sec. 1. In Sec. 2, we present t-SNE [2] plots to visualize the verb and role embeddings. We present several examples of the influence of different roles on predicting the *verb-frame* correctly. This is visualized in Sec. 3 through propagation matrices similar to Fig. 7 of the main paper. Finally, in Sec. 4 we include several example predictions that our model makes.

1. Verb Prediction

We present the verb prediction accuracies for our fully-connected model on the development set in Fig. 1. The random performance is close to 0.2% (504 verbs). About 22% of all verbs are classified correctly over 50% of the time. These include `taxiing, erupting, flossing, microwaving`, *etc*. On the other hand, verbs such as `attaching, making, placing` can have very different image representations, and show prediction accuracies of less than 10%.

Our model helps improve the role-noun predictions by sharing information across all roles. Nevertheless, if the verb is predicted incorrectly, the whole situation is treated as incorrect. Thus, verb prediction performance plays a crucial role.

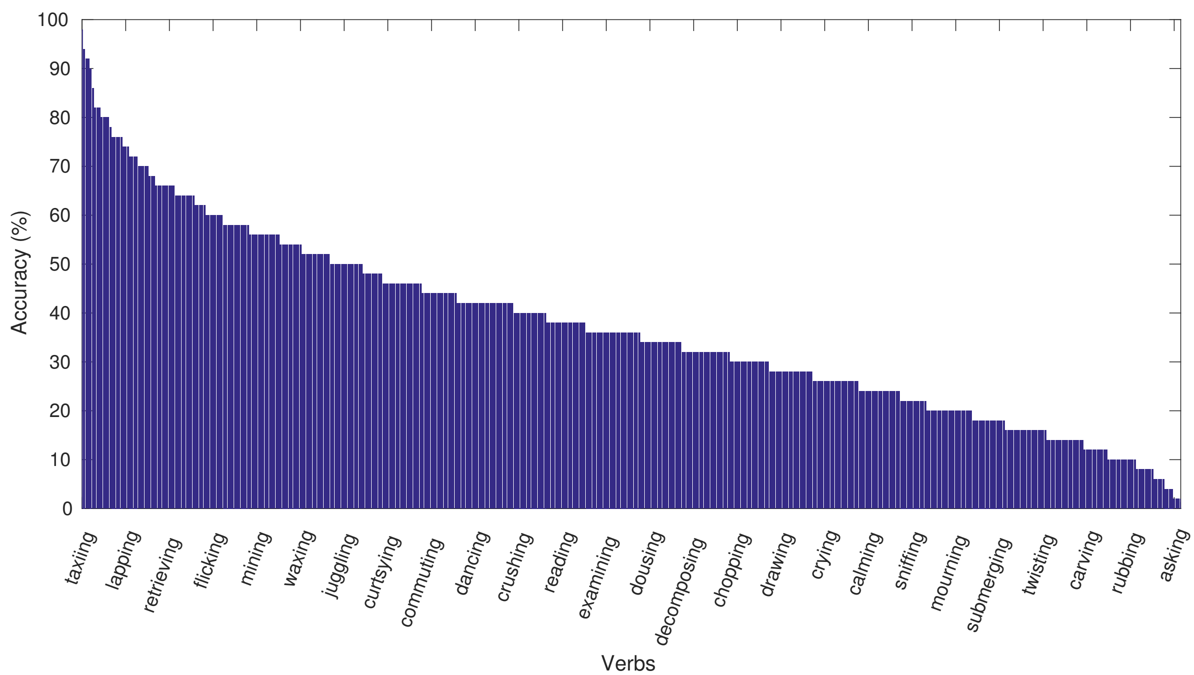


Figure 1. Verb prediction accuracy on the development set. Some verbs such as `taxiing` typically have a similar image (a plane on the tarmac), while verbs such as `rubbing` or `twisting` can have very different corresponding images.

Confusion between similar verbs. We analyze the confusion between similar verbs, that according to the metrics, leads to incorrect situation recognition. In the main paper, Fig. 8 presents a few examples where we are able to correctly predict the roles, but the situation is classified as wrong since the verb is incorrect.

The *imSitu* dataset consists of 504 verbs, and while we do have a complete 504×504 confusion matrix, visualizing the results is hard. As explained in the dataset [3], the verb frames were obtained using FrameNet. We notice that the 504 verbs from the *imSitu* dataset are grouped into 161 FrameNet verbs [1]. For example, several verbs such as `walking, climbing, skipping, prowling` and 26 others are clustered together to the FrameNet verb: `self_motion`. The clusters need not be large, and 73 of 161 clusters consist of just one verb.

We use this as a clustering, and present several confusion matrices for verb clusters in Fig. 2. All verb predictions that do not belong to the cluster are grouped as `others`. While, the others column does collect most of the predictions, there is significant confusion between similar verbs.

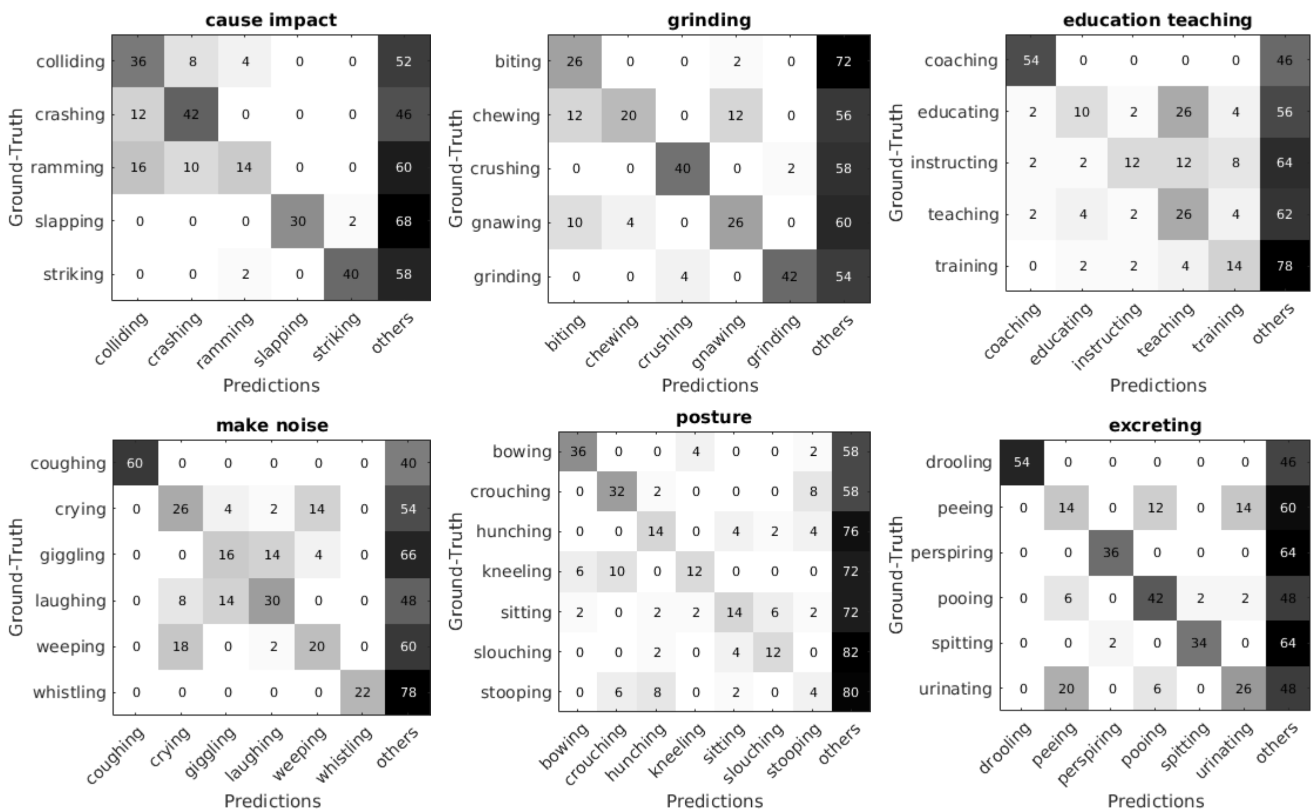


Figure 2. Confusion matrices for verb prediction. Each row indicates the expected ground-truth, and the columns are predictions (each row sums to 100%). As it is not possible to show all 504 verbs, we pick verb clusters based on their FrameNet labels (shown in the title). Confusion between remaining verbs not in the cluster is grouped in the last column as `others`. The examples show significant confusion between verbs which are hard to differentiate visually: `colliding-crashing-ramming`, or `crying-giggling-laughing-weeping`.

2. Verb and Role Embeddings

We initialize the hidden states of our role nodes (*c.f.* Eq. 2 of the main paper) with

$$h^0_{a_e} = g(W_{in}\phi_n(i) \odot W_e e \odot W_v \hat{v})\,, \tag{1}$$

where, W_v and W_e are verb and role embeddings respectively, and $e \in \mathbb{R}^{190}$ and $\hat{v} \in \mathbb{R}^{504}$ are one-hot vectors representing the noun for a specific role, and the predicted verb. $\phi_n(i)$ is the image representation using the noun-prediction CNN. Note that both verbs and roles are embedded to a $\mathbb{R}^{1024}$ space.

Verbs. The dataset consists of 504 verbs. We first show a plot depicting all verbs in Fig. 3. Owing to the number of verbs, this is quite hard to see, nevertheless, we can still observe clusters of similar verbs (*e.g.* `dusting-cleaning-`

Figure 3. 2D t-SNE representation of the all the learned verb embeddings. While the number of labels is quite large, it is still possible to see small clusters of verbs forming at the periphery of the figure. **top**: farming-harvesting, pouring-emptying-milking, slicing-chopping-peeling. **top-right**: carting-wheeling-heaving, pinching-poking. **right**: providing-giving, offering-begging-serving, reading-squinting-staring. **bottom-right**: betting-gambling, grieving-mourning, baptizing-praying. **bottom**: glowing-flaming, bubbling-overflowing, sniffing-smelling. **bottom-left**: landing-taxiing, dialing-calling-phoning-typing, boating-rowing. **left**: drinking-lapping, microwaving-baking, mining-climbing-descending. **top-left**: dusting-scrubbing-cleaning-wiping, drying-hanging, repairing-fixing-installing.

`scrubbing-wiping`, `recording-singing-performing`, *etc.*).

Additionally, we use the verb clustering afforded by the FrameNet verb associations, and select a set of 196 verbs from the 11 largest clusters (cluster size ≥ 8). We present their embeddings in Fig. 4. The learned embeddings not only discover the clustering, but are also able to associate across clusters. For example, (in the top-left corner), `applying` and

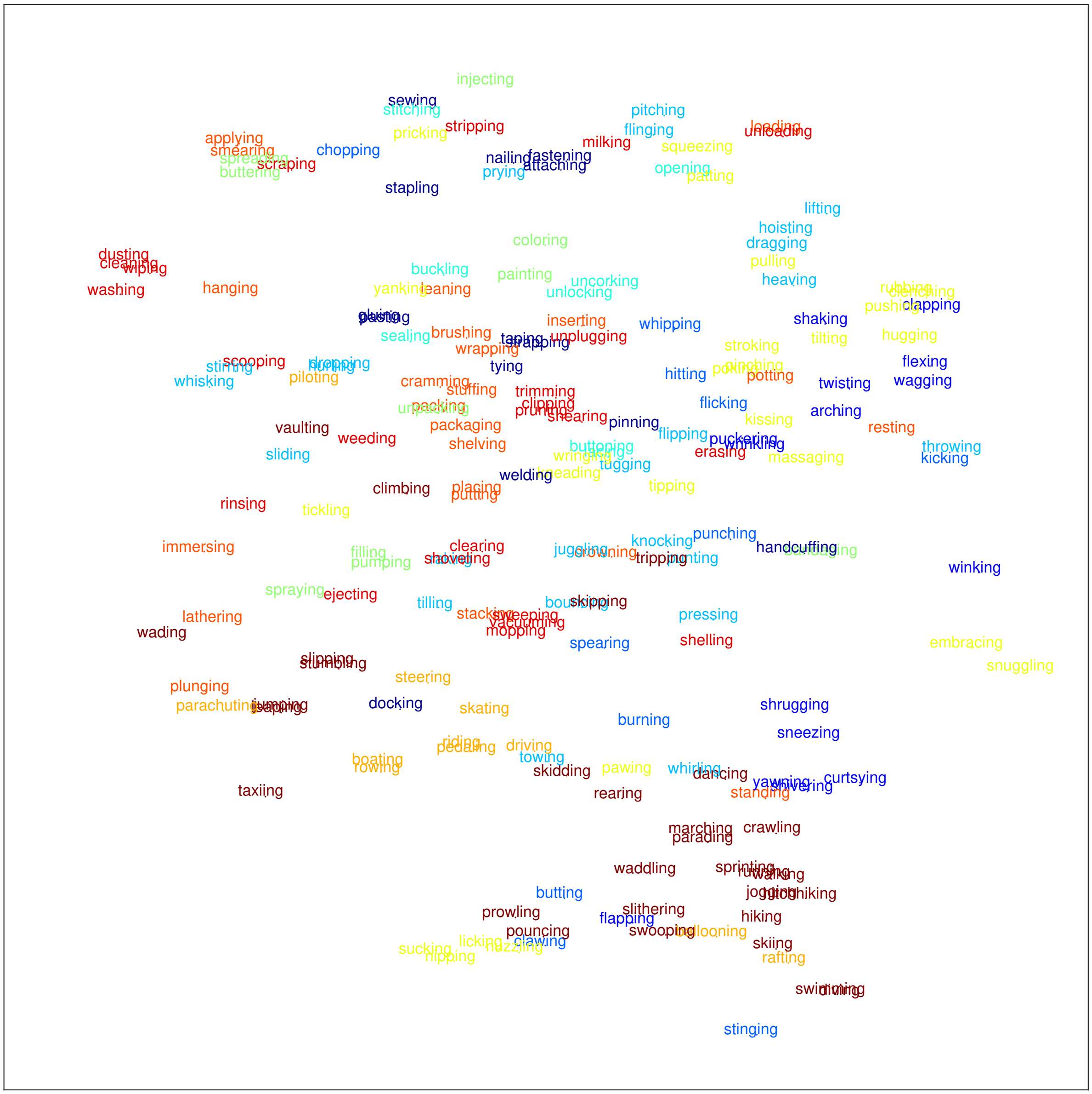


Figure 4. 2D t-SNE representation of the learned verb embeddings of the verbs belonging to 11 largest clusters (using FrameNet verb clustering). The clusters are: *attaching, body_movement, cause_harm, cause_motion, closure, filling, manipulation, operate_vehicle, placing, removing, self_motion*. Each cluster is assigned a unique color from the jet colormap. Our model is even able to learn to embed similar verbs across these FrameNet groupings. For example, it brings together `whirling` (FrameNet: cause_motion) and `dancing` (FN: self_motion); `raking` (FN: cause_motion) and `shoveling` (FN: removing); `packing` (FN: placing) and `unpacking` (FN: filling); `throwing` (FN: cause_motion) and `kicking` (FN: cause_harm); and many others.

`smearing` belong to the `Placing` FrameNet verb, while `spreading` and `buttering` correspond to `Filling` in FrameNet. Nevertheless, our model is able to learn that these verbs may have similar context (*e.g.* buttering bread), and brings their representations close.

Roles. The dataset comes with 190 roles, however, 139 of them are unique to one verb. For example, the roles `top` and `bottom` appear only once, in the frame for the verb `stacking`. Similarly, roles `shape` and `cloth` appear only when the verb is `folding`. We present two-dimensional t-SNE [2] representations of the learned role embeddings in Fig. 5. We associate same colors with role pairs that are associated with only one verb (there are only 12 such pairs, accounting for 24

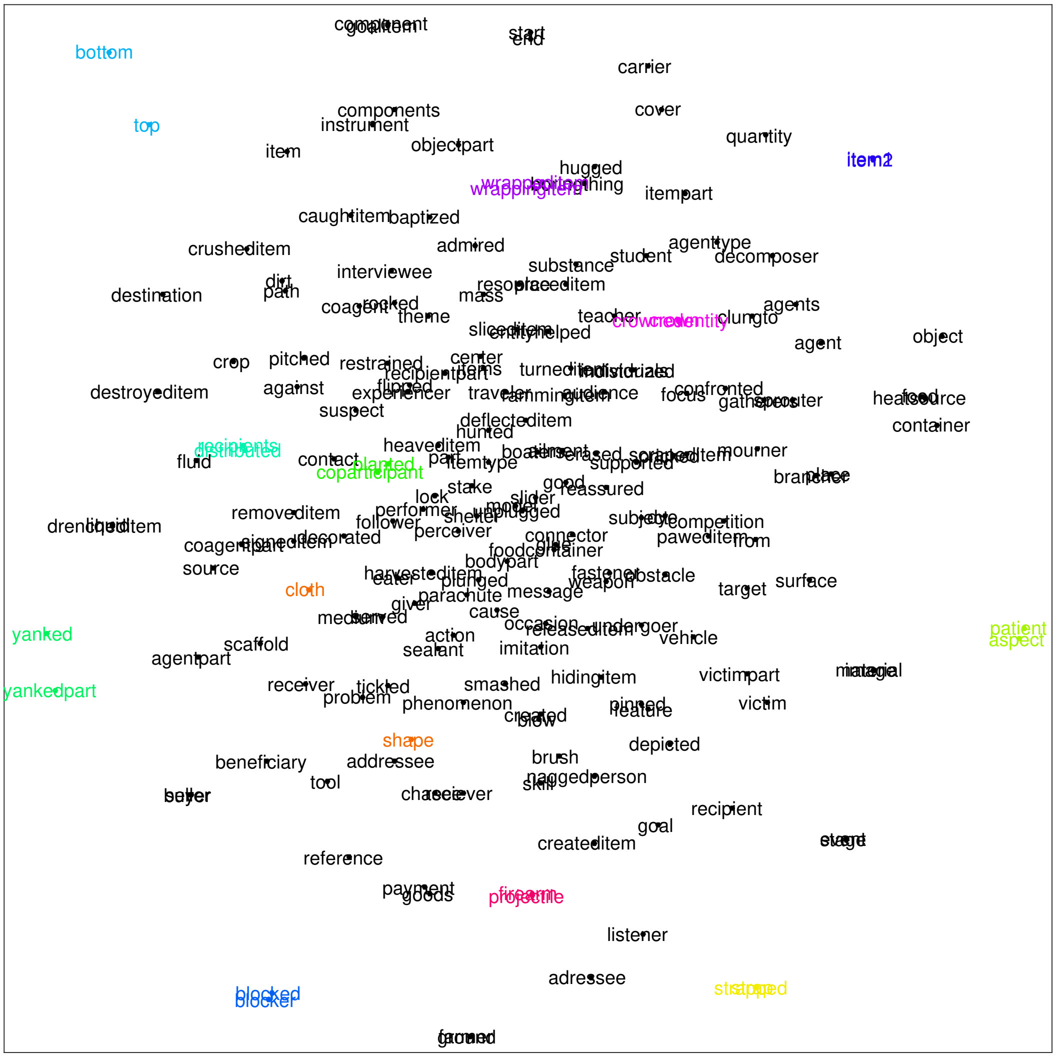


Figure 5. 2D t-SNE representation of the learned role embeddings. Note how semantic roles capturing similar themes are brought together. For example, `blocked-blocker`, or `recipients-distributed`, or `payment-goods`. Additionally, related semantic roles that apply across verbs are also brought together. For example, `components-instrument-object-part`, or `liquid-drencheditem`, `foodcontainer-glue-connector`. As most roles do not present a natural clustering, we are unable to color all roles, and they are shown in black. Colored roles are associated with one unique verb.

of 190 roles). All other roles are shown in black. In the Fig. 5, we see that the strongly related pairs that are unique to one verb (and colored) are very close to each other. Additionally, other semantic roles that are related, *e.g.* `food`, `heatsource`, `container` (right side of figure) are also close together.

3. Visualizing the propagation matrices.

We visualize the propagation matrix for 30 more verbs (extending Fig. 7 of the main paper). Note that, even though we choose the verbs randomly, we see that many verbs do have dominant roles that influence others. Each row consists of the matrix, and 4 randomly chosen images corresponding to the verb.

Our model propagates information between all roles, and we present the norm of the message sent by each role to the other in the propagation matrix. The verb and list of roles is displayed at the beginning of each row for simplicity. The rows and columns of the propagation matrix follow this ordering of roles.

Verb: ADJUSTING
Roles: agent, place, item, feature, tool

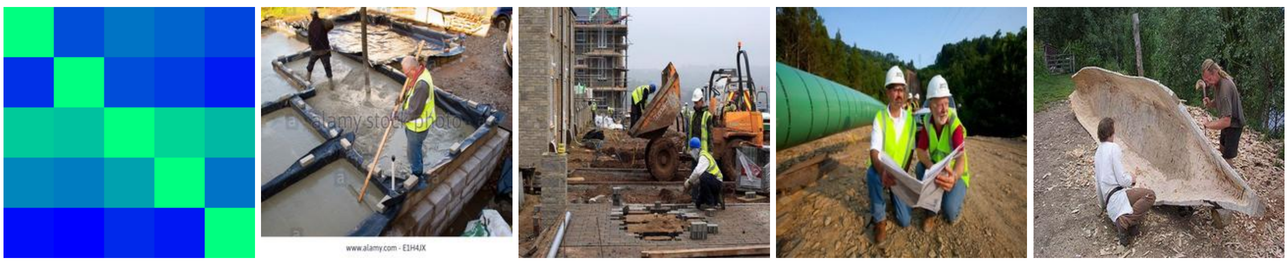
Verb: ASKING
Roles: agent, place, addressee

Verb: AUTOGRAPHING
Roles: agent, place, item, receiver

Verb: BROWSING
Roles: agent, place, goalitem

Verb: BRUSHING
Roles: agent, place, target, tool, substance

Verb: BUILDING
Roles: agent, place, goalitem, components, tool



Verb: BURNING
Roles: agent, place, target

Verb: CARRYING
Roles: agent, place, item, agentpart

Verb: CHECKING
Roles: agent, place, patient, aspect, tool

Verb: COMMUTING
Roles: traveler, place, vehicle

Verb: CRAFTING
Roles: agent, place, created, instrument

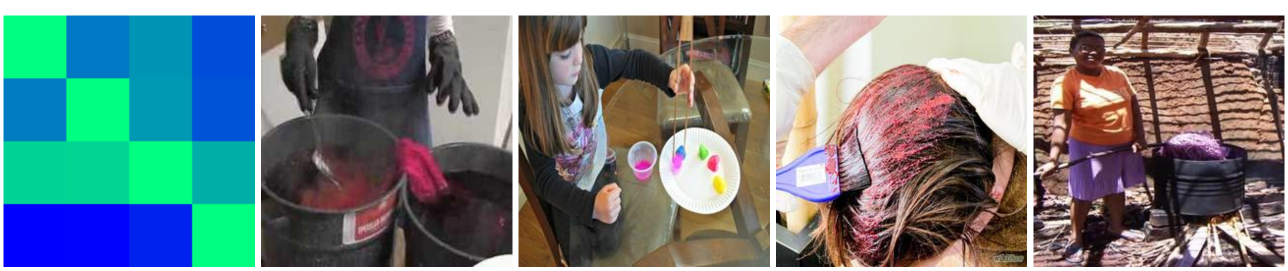
Verb: DECORATING
Roles: agent, place, decorated, item

Verb: DIPPING
Roles: agent, place, item, substance

Verb: DISTRACTING
Roles: agent, place, victim

Verb: DISTRIBUTING
Roles: agent, place, tool, distributed, recipients

Verb: DYEING
Roles: agent, place, dye, material



Verb: EXAMINING
Roles: agent, place, item, tool

Verb: FLICKING
Roles: agent, place, object, objectpart, agentpart

Verb: GIVING
Roles: agent, place, item, recipient

Verb: GLUING
Roles: agent, place, item, goal, connector

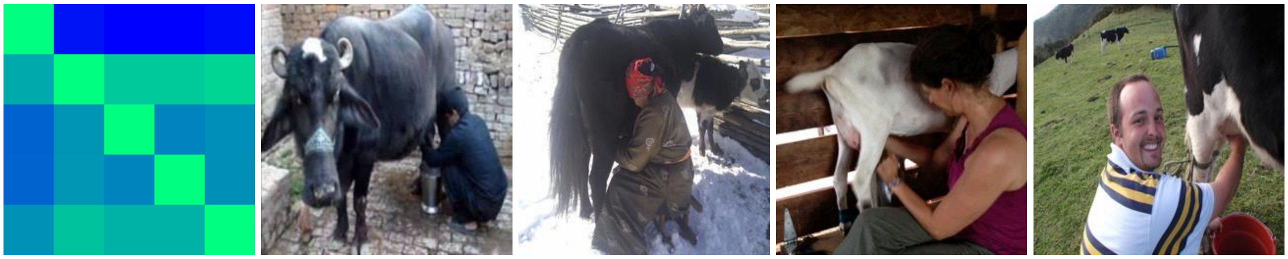
Verb: HUNCHING
Roles: agent, place, surface

Verb: INSTALLING
Roles: agent, place, component, destination, tool

Verb: KISSING
Roles: agent, place, coagent, coagentpart, agentpart

Verb: LAUNCHING
Roles: agent, place, item, source, destination

Verb: MILKING
Roles: agent, place, source, tool, destination



Verb: OFFERING
Roles: agent, place, item, beneficiary

Verb: PACKING
Roles: agent, place, item, container

Verb: PAWING
Roles: agent, place, paweditem, agentpart

Verb: PERFORMING
Roles: agent, place, event, stage, tool

Verb: PLUMMETING
Roles: agent, place, start, destination

4. Prediction Results

We round up the supplementary material with several more example predictions from our model. Fig. 6 shows predictions that are completely correct. Fig. 7 shows examples where we are able to predict the correct verb, but not all the role-noun pairs. Such examples are counted towards the *value* metric, but not *value-all*. Finally, Fig. 8 shows top-scoring (log-probability) examples where the verb is wrongly predicted, but is mostly plausible (the correct noun predictions are not captured by any metric). The role-noun pairs here are often correct.

References

[1] C. J. Fillmore, C. R. Johnson, and M. R. L. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250, 2003. 2

[2] L. J. P. van der Maaten and G. E. Hinton. Visualizing High-Dimensional Data using t-SNE. *Journal of Machine Learning Research*, 2008. 1, 4

[3] M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *CVPR*, 2016. 2

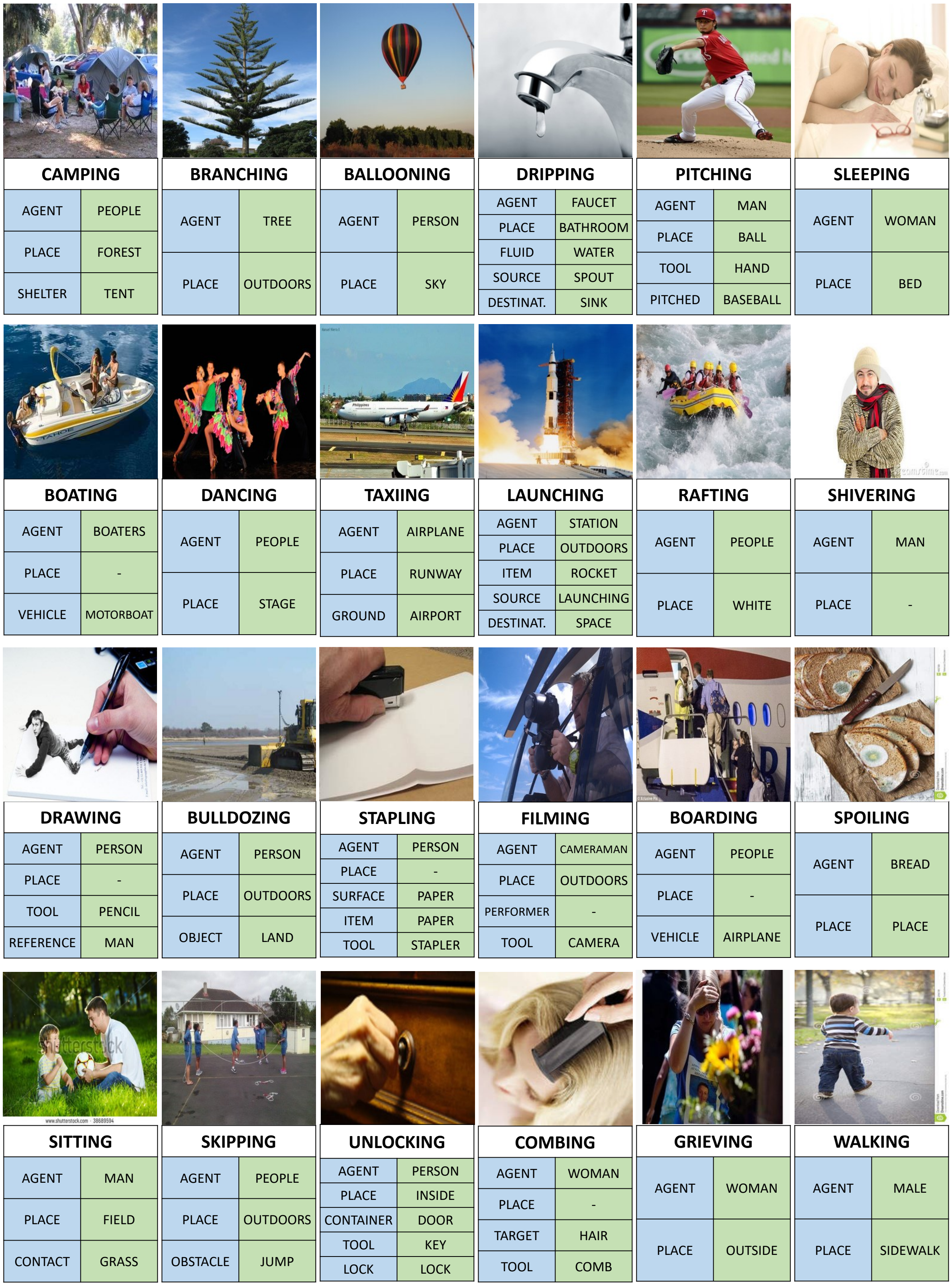


Figure 6. Images with top-1 predictions from the development set. For all samples, the predicted verb is correct, and is shown below the image in bold. Roles are marked with a blue background, and predicted nouns with green when correct, and red when wrong. We are able to correctly predict the situation (verb and all role-noun pairs) for all example images shown here.

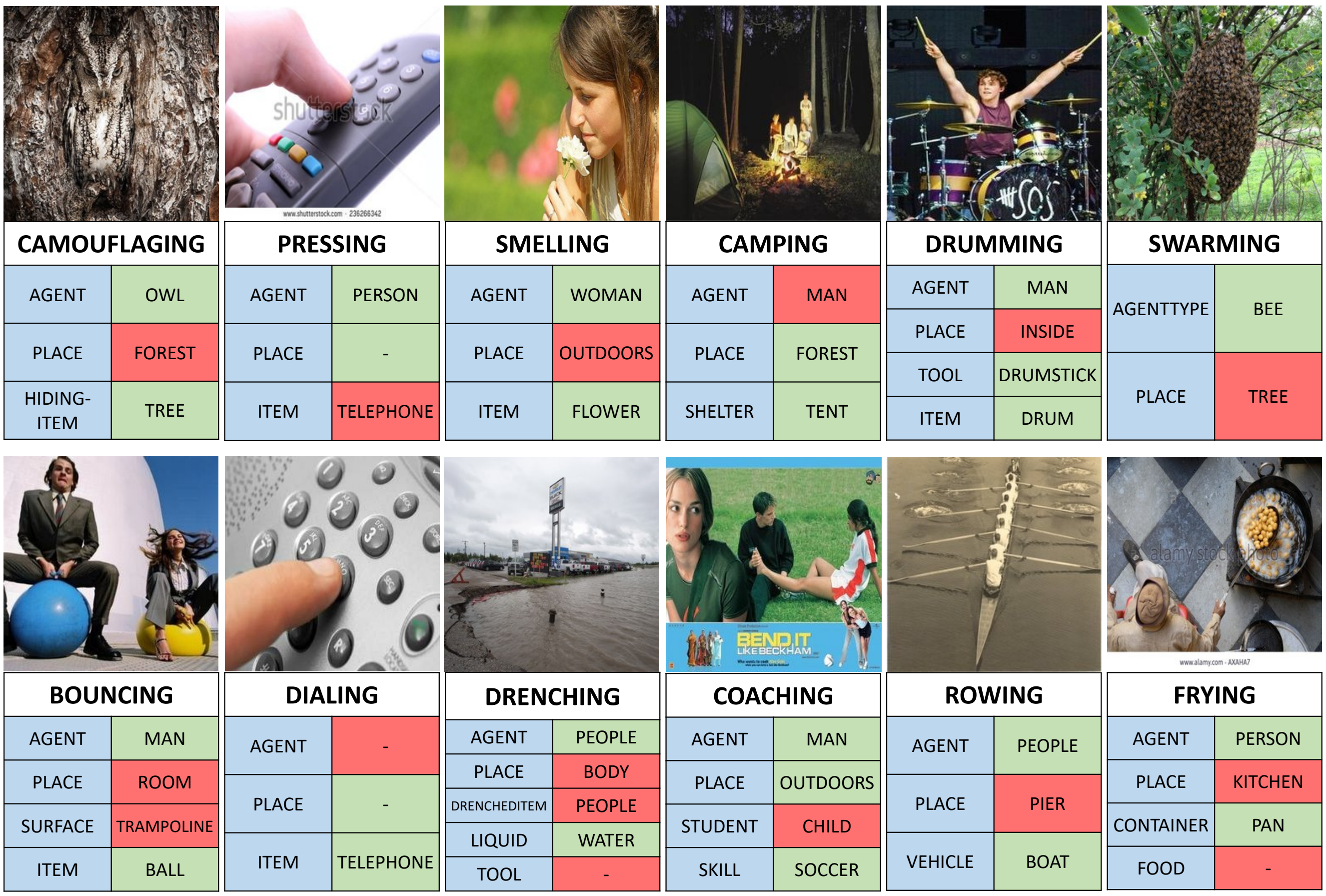


Figure 7. Images with top-1 predictions from the development set. For all samples, the predicted verb is correct, and is shown below the image in bold. Roles are marked with a blue background, and predicted nouns with green when correct, and red when wrong. We show examples with genuine errors in prediction (*e.g.* the `telephone` for the verb `pressing` is clearly a `remote control`). However, some examples are marked wrong due to the lack of matching ground-truth annotations (*e.g.* the woman `smelling` the flower is `outdoors` (GT: field)).

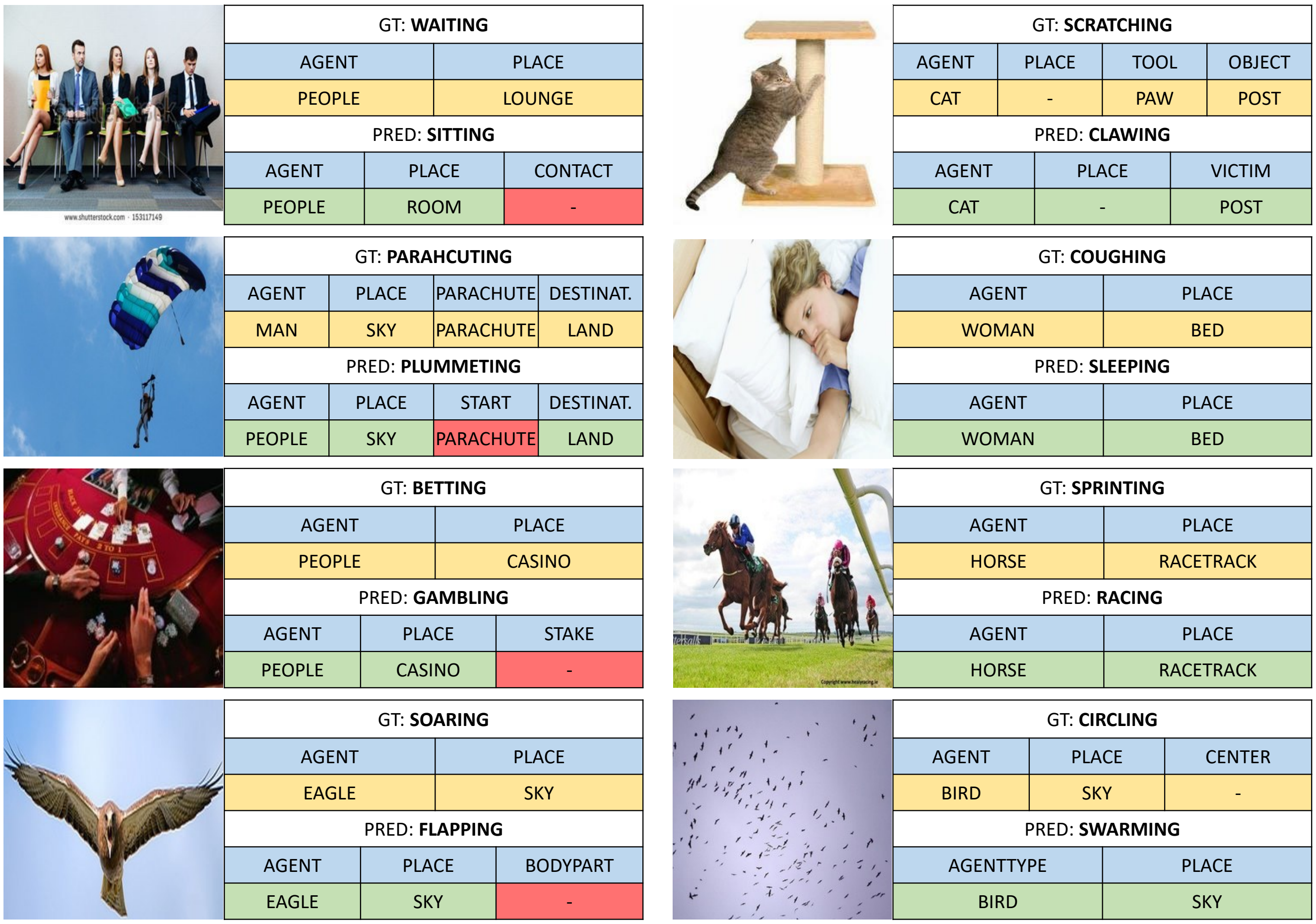


Figure 8. Images with ground-truth and top-1 predictions from the development set. Roles are marked with blue background. Ground-truth (GT) nouns with yellow, and predicted (PRED) nouns with green when correct, or red when wrong. Although the predicted verb is different from the ground-truth, it is very plausible. Some of the verbs refer to the same frame (*e.g.* `sprinting`, `racing`), and contain the same set of roles, which our model is able to correctly infer.