

Learning Action Changes by Measuring Verb-Adverb Textual Relationships

Supplementary material

Davide Moltisanti, Frank Keller, Hakan Bilen, Laura Sevilla-Lara
The University of Edinburgh, United Kingdom

{davide.moltisanti, frank.keller, h.bilen, l.sevilla}@ed.ac.uk

1. Adverbs in Recipes - Details

Parsing Captions We use SpaCy [3] to parse captions in HowTo100M [2]. We start filtering captions containing a verb in one of the following tenses/forms: ‘VB’: base verb (e.g. “take”), ‘VBP’: present tense (e.g. “take”), ‘VBZ’: present tense 3rd person singular (e.g. “takes”), ‘VBG’: gerund, present participle (e.g. “taking”). We discard verbs in past tenses to avoid parsing adjectives as adverbs (e.g. “coarsely ground”). We then look among the syntactic children of each verb to find adverbs attached to the verb. We manually cluster verbs and adverbs with a similar meaning. We then filter out: i) verbs and adverbs co-occurring less than 100 times; ii) adverbs related to location (e.g. “diagonally”), feelings (e.g. “happily”), instants/periods (e.g. “immediately, continually”), adverbs that are subjective (e.g. “beautifully”) or too generic (e.g. “normally”); iii) videos shorter than 5 seconds and longer than 1 minute.

Annotation After filtering we collected 11,271 video clips, which we annotated via Amazon Mechanical Turk (AMT). For each video we asked annotators to confirm if the action was visible and performed as indicated by the adverb. We also asked additional questions regarding video editing. Specifically, annotators were asked to check if: i) the speed of the video was altered (i.e. slowed-down or sped-up); ii) the video contains jump-cuts (i.e. parts of the action are skipped); iii) the video contains static segments (e.g. still frames with text). We collected these extra annotations for potential future studies. Each video was labelled by 3 annotators. We employed a total of 5 annotators for edge cases where people did not reach a consensus regarding the main questions (action is visible and is performed as indicated by the adverb). We kept videos where the majority confirmed that both the action and the adverb effect are visible, which resulted in 7,003 videos. We showed annotators several examples to illustrate the task.

Verb and Adverb Distributions We plot the adverb and verb distributions respectively in Figure 1 and 2. Like the existing datasets we reviewed in the paper, AIR exhibits a

Parameters	Our Model	Action Modifiers [1]
Attention model (same)	344,960	344,960
Features encoder (MLP)	267,786	-
Adverb parameters	-	2,621,440
Total	612,746	2,966,400

Table 1. Comparing number of parameters in our model and Action Modifiers [1], calculated for 10 adverbs. Attention parameters calculated with default settings: 4 heads, input features dimension equal to 1024 and Q, K, V dimensions equal to 512.

long tail with a heavy class imbalance. Figure 3 depicts the co-occurrence matrix of existing (verb, adverb) pairs in the dataset. The matrix is naturally sparse as not all adverbs apply to all verbs. Some pairs appear more frequently (e.g. “chop finely/coarsely”) compared to others (e.g. “drip gently, mash slowly”). This is expected as some actions and the ways they can be performed are more common than others.

2. Adverbs in Recipes - Video

We prepared a video to show a few samples from our new Adverbs in Recipes dataset, which you can watch at <https://youtu.be/YPNw35vtyu8>. Note how videos are well trimmed and do not contain unrelated content thanks to our better trimming method (see paper for more details). Actions are well visible and importantly are performed as indicated by the adverb, thanks to our manual review.

3. Comparing Models Capacity

Table 1 compares the number of parameters in our model and Action Modifiers [1]. We note that our model outperforms Action Modifiers with an order of magnitude fewer parameters (612,746 vs 2,966,400). In particular, our model scales much better according to the number of adverbs. In fact, Action Modifiers learns weights ($E \times E$) for each adverb. In the experiments $E = 512$, thus assuming the number of adverbs A is 10, in Table 1 we have $(E \times E) \times A = (512 \times 512) \times 10 = 2,621,440$. In contrast, only the last layer of our MLP changes according to

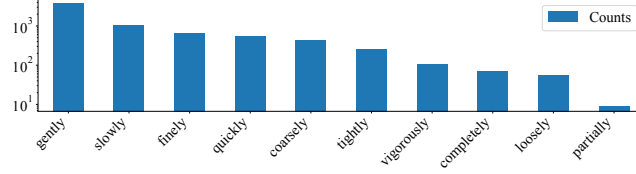


Figure 1. Adverbs in Recipes: adverb distribution (log scale).

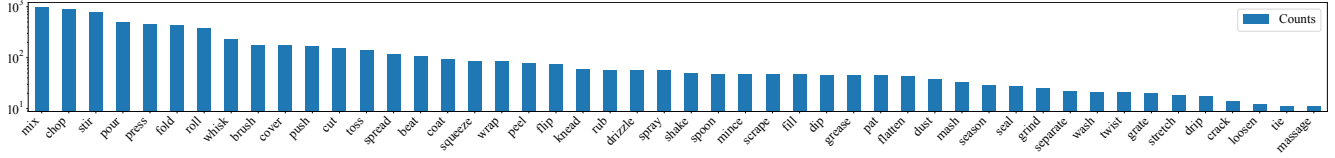


Figure 2. Adverbs in Recipes: verb distribution (log scale).

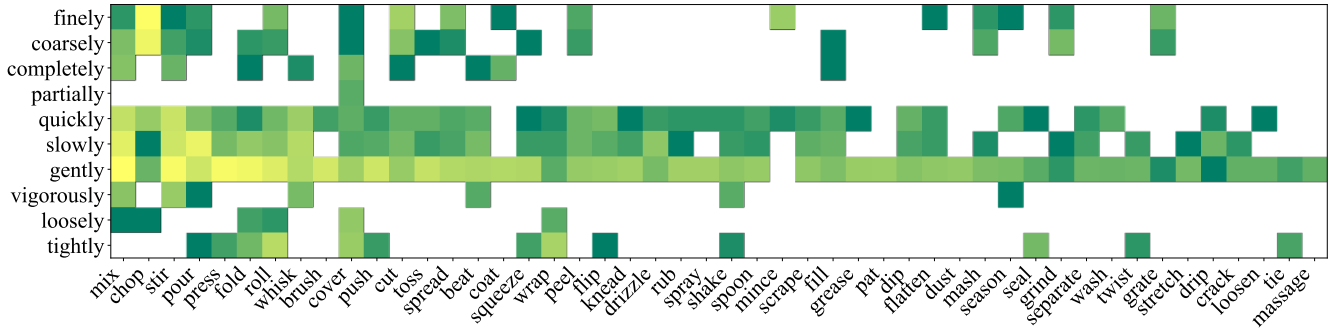


Figure 3. Adverbs in Recipes: verb-adverb co-occurrences. Darker (green)/Brighter (yellow) correspond to less/more frequent pairs. A missing square indicates that the pair does not appear in the dataset.

Model	mAP W	mAP M	Acc-A
Act Mod [1]	0.394 ± 0.023	0.140 ± 0.026	0.843 ± 0.013
MLP + Act Mod [1]	0.407 ± 0.044	0.151 ± 0.033	0.842 ± 0.012
CLS	0.676 ± 0.001	0.317 ± 0.007	0.847 ± 0.001
REG-fixed δ	0.455 ± 0.004	0.153 ± 0.018	0.835 ± 0.000
REG	0.662 ± 0.006	0.289 ± 0.010	0.863 ± 0.002

Table 2. Results variance on AIR. Numbers indicate mean \pm std.

the number of adverbs A . The input to the MLP is a vector of dimension 1024. We have 3 hidden layers of dimension 512, whereas the last layer has dimension A . Counting both weights and biases, our shallow MLP requires only 267, 786 parameters (assuming $A = 10$). The fact that we obtain state-of-the-art results with a much smaller capacity confirms that the key in our better performance lies in a better training strategy.

4. Results Variance

The size of the evaluated datasets is relatively small for deep learning methods. In order to assess the variance of the results we run experiments on AIR two more times, gathering a total of three runs including results from the paper, following the standard setting where models are trained with

antonyms and are tested using the action labels. Table 2 reports mean \pm standard deviation of the three evaluation metrics. Our method is more stable than Action Modifiers [1]. Most importantly, the ranking of the methods and the improvement of our method remains the same.

References

- [1] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *CVPR*, 2020. 1, 2
- [2] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 1
- [3] Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O’Leary McCann, jim geovedi, Jim O’Regan, Maxim Samsonov, Duygu Altinok, György Orosz, Daniël de Kok, Søren Lind Kristiansen, Raphaël Bournhonesque, Madeesh Kannan, Lj Miranda, Peter Baumgartner, Edward, Explosion Bot, Richard Hudson, Roman, Leander Fiedler, Raphael Mitsch, Ryn Daniels, Grégory Howard, Wannaphong Phatthiyaphaibun, Yohei Tamura, and Sam Bozek. explosion/spaCy: v3.4.3: Extended Typer support and bug fixes, Nov. 2022. 1