Towards an Unequivocal Representation of Actions

Michael Wray
University of Bristol

Davide Moltisanti
University of Bristol

Dima Damen
University of Bristol

firstname.surname@bristol.ac.uk

Abstract

This work introduces verb-only representations for actions and interactions; the problem of describing similar motions (e.g. 'open door', 'open cupboard'), and distinguish differing ones (e.g. 'open door' vs 'open bottle') using verb-only labels. Current approaches for action recognition neglect legitimate semantic ambiguities and class overlaps between verbs (Fig. 1), relying on the objects to disambiguate interactions. We deviate from single-verb labels and introduce a mapping between observations and multiple verb labels – in order to create an Unequivocal Representation of Actions. The new representation benefits from increased vocabulary and a soft assignment to an enriched space of verb labels. We learn these representations as multi-output regression, using a two-stream fusion CNN. The proposed approach outperforms conventional single-verb labels (also known as majority voting) on three egocentric datasets for both recognition and retrieval.

1. Introduction

Consider a collection of verbs one uses to describe preparing morning coffee: *open, pick, put, turn, scoop, pour, fill, stir, close, etc.* Verbs represent important information about how we can interact with the world, yet – especially in the English language – are usually given context in the form of object(s) for disambiguation. The motion that is used to push a door is different to that of pushing a button and, as such, door and button are used to differentiate between the two motions (*i.e.* 'push-door' vs 'push-button'). This was recently highlighted in [16], where increased confusion has been reported by human annotators when given a singular verb label, compared to verb-noun labels. However, this leads to motions being tied towards objects when in fact the same motion could be applied to different objects, *i.e.* opening a cupboard is similar to opening a microwave or a fridge.

In this work we explore the idea of describing the action using a soft assignment over individually-ambiguous verb labels, yet keep it applicable for interactions with multiple objects. Take for example: {*open, hold, turn, rotate*}; by using multiple verbs, the motion is less ambiguous, yet is kept general to describe interactions with multiple objects, (*e.g.* jar, bottle, tap). Note that we are not attempting to discover the objects being used; rather *we seek a coherent representation of the action*, which can be used for recognition and retrieval tasks. We focus on the egocentric domain, as object interactions are frequent and successive within a common environment.

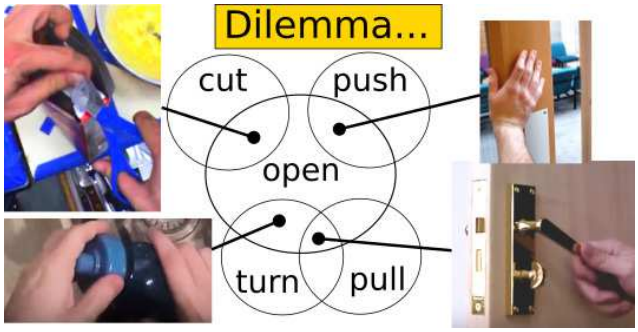

Figure 1. Using single-verbs results in class overlaps.

We propose benefits of using multiple verbs in Fig. 2. In Fig. 2(a), we query our predicted representations using the verbs 'turn-on/off', combined with one other verb ('rotate' vs 'press'). The proposed unequivocal representation can make the distinction between a tap closed by rotating [first row in blue] and one by pressing [second row in blue], whereas neither single-verb nor verb-noun labels can. Models trained using the proposed representation can learn an enriched space of verb labels. In Fig. 2(b), different interactions can be retrieved using a common sub-action.

While we note that multi-label representations have become increasingly common for object recognition [20, 21], using multiple verbs to describe an action is under-explored for video understanding. Previous datasets, egocentric [4, 22, 3] and non-egocentric [7, 9, 17, 19] are annotated with a pre-selected number of verbs and commonly evaluated with classes defined as verb-noun pairs. A few works attempt verb-only labels [18, 22], with both noting the difficulty and ambiguity of using single verb only labels. Khamis and Davis [8] do use multi-verb labels in action recognition. However, they use a small amount of verbs (10) which describe non-overlapping actions, whereas we focus on using

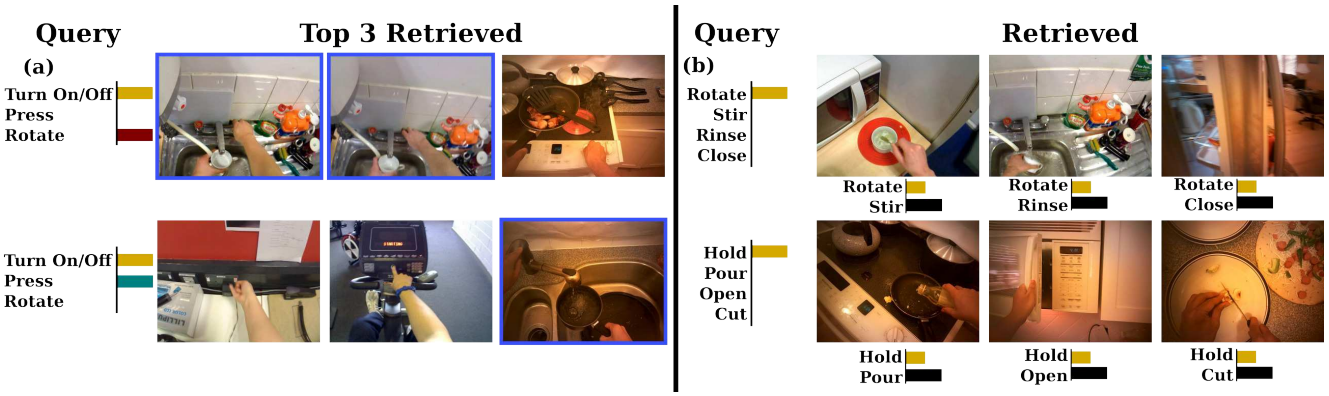

Figure 2. Benefits of using multi-verb labels. (a) The labelling method is able to distinguish between turning on/off a tap by rotating and pressing (highlighted in blue). (b) Verbs such as rotate and hold can be learned via context from other actions.

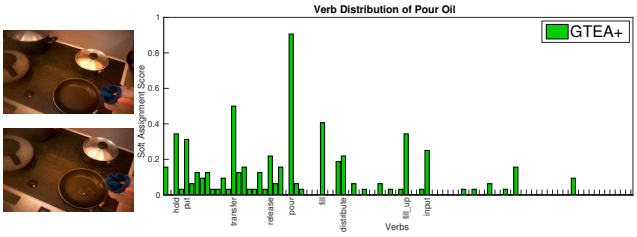

Figure 3. Example annotation for the Pour Oil class from GTEA+.

verbs which describe a single action.

We next present the proposed representation. It is crowd-sourced, as in [1, 6, 15], and evaluated using two-stream CNN [5]. We present results for classification and retrieval using three egocentric datasets.

2. The Unequivocal Representation of Actions

In this section, we define the proposed representation that assigns multiple verb-only labels to action segments, in order to reduce single-verb ambiguity. We use annotations as collected in [23] for the three public datasets [22, 3, 4]. The annotations were collected per class with multiple annotators choosing which verbs, out of a list of 90, were applicable for the video (see Fig. 3).

Definitions: When using a **Single-verb Label (SL)**, each video $x_i \in X$ has a corresponding label $\boldsymbol{y_i} \in Y$ where $\boldsymbol{y_i}$ is a one-hot vector, over verbs $V = \langle v_j \rangle$. To minimise class overlaps, a small set of semantically distinct verbs are typically used. We use majority voting to create **SL**.

Alternatively, a **Multi-verb Label (ML)**, $\boldsymbol{y_i} = \langle y_{i,j} \in \{0, 1\} \rangle$ is a binary vector over V. Multiple verbs can be used to describe the video, e.g. 'pour' and 'fill'. Hard assignment though can be problematic for sub-actions. Verbs such as 'hold' do not fully describe the object interaction yet cannot be ignored as irrelevant. We construct **ML** as a binary vector where each verb above a threshold of 50% is set to 1.

In introducing the **Soft Assigned Multi-Label (SAML)**, we wish to increase the size of V while accommodating sub-actions. Soft assignment offers a ranking of verb labels, ordered by the ability of the label, to sufficiently describe the ongoing interaction. In the Soft Assigned Multi-Label, each video x_i will have a label vector $\boldsymbol{y_i} = \langle y_{i,j} \in [0, 1] \rangle$ over V. For two verbs v_j, v_k, in SAML, $y_{i,j} > y_{i,k} > 0$ when the first verb is more commonly used to describe the action in x_i, while $y_{i,k}$ is still a valid/relevant label. We normalise the responses by the number of annotators to get the soft assignment score.

Note: Acquiring the proposed representations from semantics is potentially challenging. Some verbs will be related semantically; e.g. 'hold' and 'grasp' are synonyms. Others are related via context, e.g. 'pour' and 'fill' are linked depending on the viewpoint (*Is the bottle being poured?* or *Is the cup being filled?*). Finally, we have sub-actions, e.g. the user must 'hold' the bottle to be able to 'pour' its contents. While these relationships can be explicitly stated, they are *interestingly* not available in lexical databases and hard to discover from public corpora. We study two commonly used sources of semantic information, WordNet [11] and Word2Vec [10] embeddings, showing their limitations:

		WordNet	Word2Vec
synonyms	(e.g. 'hold'-'grasp')	√	×
context-related	(e.g. 'fill'-'pour')	×	×
sub-actions	(e.g. 'hold'-'pour')	×	×

Learning: For each of the three labelling approaches (Fig 4), we wish to learn a function, $\phi : \mathcal{W} \rightarrow \mathbb{R}^D$ which maps a video representation $\mathcal{W}$ onto labels with $D = |V|$. For brevity we define $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \phi(x_i)$, $\hat{y}_{i,j}$ as the predicted value for verb v_j of video x_i and $y_{i,j}$ as the corresponding ground truth. Typically, the single label (SL) is learned using a cross entropy loss of the softmax scores. To learn the multi-label (ML) we use a sigmoid binary cross entropy loss as commonly used in multi-label classification [12].

In the Soft Assigned Multi-Label (SAML) representation, each element in $\boldsymbol{y_i}$ can take any value in the range

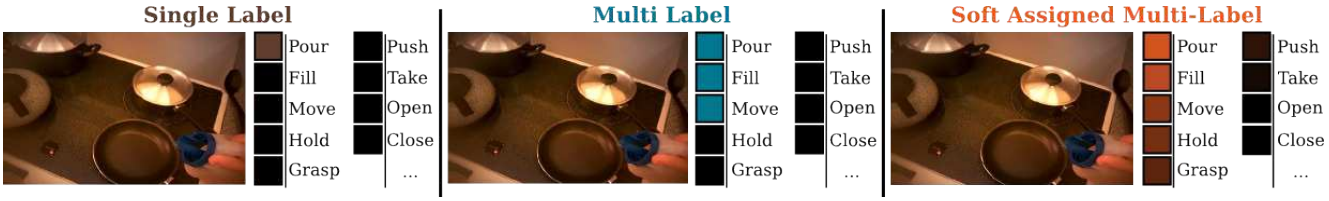

Figure 4. In single verb labelling, only one verb can be chosen, often excluding valid labels. Multi-Label increases the vocabulary size and allows multiple verbs with equal importance to describe the same action. Soft Assigned Multi-Label increases the pool of verbs even further and uses soft assignment for each.

$[0, 1]$. We formulate this as a multi-task learning problem as defined in [14], solved as a multi-output regression without any independence assumptions. We again use the sigmoid binary cross entropy loss. We consciously avoid a ranking loss as it only learns a relative order and does not attempt to approximate the representation.

Prediction and Evaluation: We can use $\phi = \{\phi_{SL}, \phi_{ML}, \phi_{SAML}\}$ to predict the labels for a previously unseen input x_i. We next present two ways to evaluate ϕ for predicting SL, ML and SAML.

We can evaluate ϕ in its ability to find 'relevant' verbs. Given a threshold α for what verbs are deemed relevant $V_i^\alpha = \{v_j : y_{i,j} \geq \alpha,\ \forall v_j \in V\}$, The top k predicted verb labels would then be $\hat{V}_i^\alpha = \{\hat{y}_{i,j} : \hat{y}_{i,j} \in top_k(\hat{\boldsymbol{y}}_{\boldsymbol{i}}) \wedge k = |V_i^\alpha|\}$. The accuracy can now be calculated as a percentage of the overlap between the predicted and ground truth verbs:

$$A(\alpha|\phi) = \frac{1}{|X|} \sum_i \frac{|V_i^\alpha \cap \hat{V}_i^\alpha|}{|V_i^\alpha|} \tag{1}$$

Note that $A(\alpha|\phi_{SL})$, for any α, matches traditional classification accuracy, making this metric comparable to V-N.

Additionally, we can treat the ϕ as an embedding function such that $\boldsymbol{y_i}$ and $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ represents the ground-truth and predicted embeddings of video x_i respectively. Any verb v_j can be located as a vertex in the label space which corresponds to the one-hot vector with v_j set to 1 and the rest to 0, which we refer to as $\mathbf{v}_j$. We can thus define video-to-text retrieval as the ranking of verbs from closest to furthest based on the L_2 distance ($||\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \mathbf{v_j}||$), which we can compare to the true ranking ($||\boldsymbol{y_i} - \mathbf{v_j}||$). Similarly, we can define text-to-video retrieval from a given label $\mathbf{v}_j$ as the ranking of all embedded predictions $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ using the same distance metric. Importantly, we can construct more interesting text-to-video query vectors that involve multiple verbs to describe videos. Assume $\mathbf{u}_i^n$ is a binary vector with n verbs being set to 1, and the rest to zero. We accordingly perform text-to-video retrieval on these multi-verb queries and compare the various labelling methods. In order to evaluate both text-to-video as well as video-to-text retrieval, we use mean Averaged Precision (mAP) over all queries as used in [13].

	SL	ML	SAML	V-N
BEOID	78.1	**93.0**	87.8	**93.5**
CMU	59.2	74.1	73.5	**76.0**
GTEA+	59.2	**71.9**	67.8	61.2

Table 1. Action recognition accuracy (%) results compared to verb-noun(V-N) classes using Eq. 1.

3. Experiments and Results

We evaluate on the three annotated egocentric datasets: BEOID [2] (732 video segments), CMU [3] (404 video segments) and GTEA+ [4] (1001 video segments) using the annotations defined in Sec. 2. All three datasets include videos of daily activities indoors, recorded using a head mounted camera. For each dataset, 5 cross-fold validation was used where we equally distribute videos from each class.

Implementation Details: We trained for 100 epochs and tested using the two stream fusion CNN model from [5], pre-trained on UCF101 [19]. The number of nodes in the last layer equals the number of verbs $|V| = 90$.

Action Recognition Accuracy: We motivate our work by stating that multi-verb labels allow removing the ambiguity of the performed action compared to single-verb labels. We accordingly evaluate all datasets using the original ground-truth verb-noun classes, and report results in Table 1 as V-N. We train V-N using the same loss function as SL, but use the dataset's verb-noun class labels. We show that indeed adding verbs decreases the ambiguity and produces comparable results to verb-noun classes on BEOID (-0.5%), slight drop on CMU (-2%) and outperforms Verb-Noun classes on the largest of the three datasets, GTEA+ (+10%). Moreover, the reported accuracies on ML and SAML are significantly higher than the ambiguous single-verb labels on all datasets, despite the increase in the number of verbs as potential output labels. We note that ϕ_{ML} consistently achieves the highest accuracy compared to ϕ_{SL} and ϕ_{SAML} suggesting that it is more suited for recognition.

Video-to-Text Retrieval Results: We next evaluate ϕ for video-to-text retrieval. Table 2 compares mAP of retrieval using each labelling method. ϕ_{SAML} has a solid performance consistently, even when compared to SL in retrieving a single verb. Note that ϕ_{ML} has a good performance

	BEOID			CMU			GTEA+		
	SL	ML	SAML	SL	ML	SAML	SL	ML	SAML
Single-Verb only	**0.85**	0.65	0.83	**0.71**	0.55	0.68	**0.73**	0.46	0.62
$\alpha \geq 0.5$ ranking	0.46	**0.95**	0.89	0.40	**0.82**	0.74	0.42	**0.79**	0.76
$\alpha \geq 0.3$ ranking	0.34	0.64	**0.92**	0.36	0.68	**0.81**	0.32	0.60	**0.75**
Avg. mAP	0.55	0.75	**0.88**	0.49	0.68	**0.75**	0.49	0.62	**0.71**

Table 2. Video-to-Text retrieval results on the three datasets using mAP. SAML performs consistently well over different ground truth.

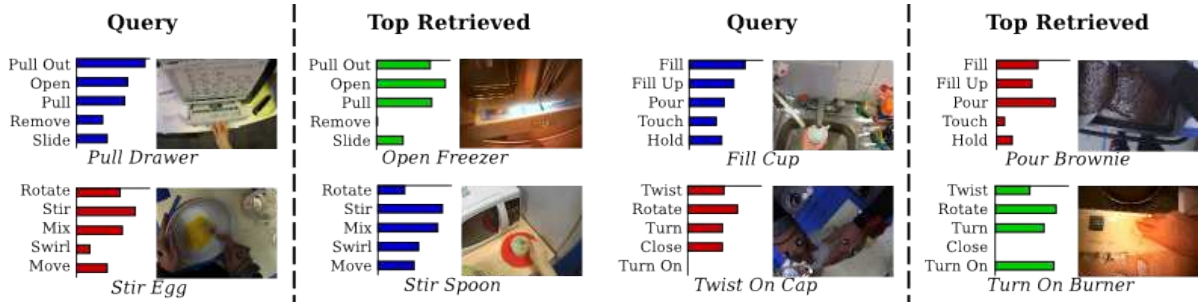

Figure 5. Examples of cross dataset retrieval of videos using either videos or text. Blue: BEOID, Red: CMU and Green: GTEA+.

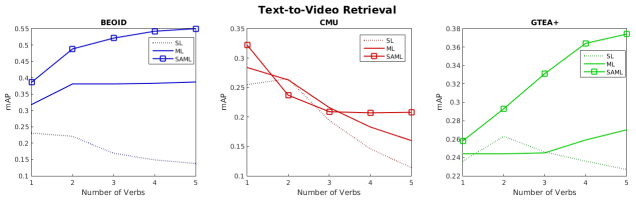

Figure 6. Results of text-to-video retrieval across all three datasets using mAP and a varying number of verbs in the query.

only at $\alpha \geq 0.5$ as this matches the labels on which it has been trained. This is emphasised when we look at the average mAP scores of each ϕ across all datasets. SAML achieves the highest average mAP across all three datasets.

Text-to-Video Retrieval Results: Figure 6 reports mAP results for text-to-video retrieval. In these results, we query the predicted representations using a binary vector $\mathbf{u}_i^n$ with n verbs, and $n = \{1 \cdots 5\}$. Note that the verbs chosen are those that co-occur in the dataset to avoid antonyms like 'open' with 'close'. All possible combinations of n co-occuring verbs have been tested. As expected SL and ML perform best with either one or two verbs used as input with the mAP staying steady or dropping. SAML increases its mAP for BEOID and GTEA+ as the number of verbs increases – outperforming SL and ML. This suggests that with the full vocabulary, the method is able to better learn the multi-verb representations for both the main verbs used to describe an action as well as any sub-actions.

Cross-Dataset Retrieval: In this section we perform video-to-video retrieval, using SAML labels, across datasets. We first predict the representation of a video using ϕ_{SAML}, then use this representation to find the top-retrieved video from a different dataset. For example, when querying using the video 'stir egg' from CMU, the top-retrieved video is of 'stir spoon' from BEOID. Interestingly, 'pull drawer' from BEOID retrieves 'open freezer' in GTEA+, as they both include the same motion, one being the drawer of a printer and the other the drawer of a freezer. In the last example, ϕ_{SAML} relates 'twist-on cap' from CMU to 'turn-on burner' from GTEA+ as both perform similar motions.

We earlier presented sample text-to-video cross-dataset retrievals in Fig. 2. While we don't report cross-dataset quantitative results, we believe examples in Fig. 5 show the potential of the proposed representations beyond a single dataset.

4. Conclusion and Future Work

In this paper, we present the case for using multi-verb labels for action videos, and propose the Soft Assigned Multi-verb labels. Compared to single verb-only labels, this offers an unambiguous representation of the interaction, embracing class overlaps. On the other hand, when compared to verb-noun labels, this representation generalises to multiple and unseen objects whilst still performing similarly comparably for recognition.

The representations, learned using a two-stream fusion CNN, are able to predict the correct verb labels – outperforming single-verb labels, for both recognition and retrieval. This representation can be useful of zero-shot or few-shot learning, predicting novel combination of verb labels. We will embark on assessing this next.

References

[1] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[2] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas. You-do, I-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014.

[3] F. De La Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. *Robotics Institute*, 2008.

[4] A. Fathi, Y. Li, and J. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012.

[5] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

[6] S. Gella, M. Lapata, and F. Keller. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *NAACL*, 2016.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[8] S. Khamis and L. S. Davis. Walking and talking: A bilinear approach to multi-label action recognition. In *CVPRW*, 2015.

[9] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[11] G. Miller. Wordnet: a lexical database for english. *CACM*, 1995.

[12] J. Nam, J. Kim, E. L. Mencía, I. Gurevych, and J. Fürnkranz. Large-scale multi-label text classification-revisiting neural networks. In *ECML PKDD*, 2014.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[14] P. Rai, A. Kumar, and H. Daume. Simultaneously leveraging output and task structures for multiple-output regression. In *NIPS*, 2012.

[15] M. R. Ronchi and P. Perona. Describing common human visual actions in images. In *BMVC*, 2015.

[16] G. A. Sigurdsson, O. Russakovsky, and A. Gupta. What actions are needed for understanding human actions in videos? In *ICCV*, 2017.

[17] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[18] S. Singh, C. Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016.

[19] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *Technical Report CRCV*, 2012.

[20] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, 2016.

[21] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. CNN: single-label to multi-label. *CoRR*, 2014.

[22] M. Wray, D. Moltisanti, W. Mayol-Cuevas, and D. Damen. Sembed: Semantic embedding of egocentric action videos. In *ECCVW*, 2016.

[23] M. Wray, D. Moltisanti, W. Mayol-Cuevas, and D. Damen. Improving classification by improving labelling: Introducing probabilistic multi-label object interaction recognition. *arXiv preprint arXiv:1703.08338*, 2017.