SYNTACTICALLY GUIDED GENERATIVE EMBEDDINGS FOR ZERO-SHOT SKELETON ACTION RECOGNITION

by

Pranay Gupta, Divyanshu Sharma, Ravi Kiran Sarvadevabhatla

in

IEEE International Conference on Image Processing (ICIP)
: 1
-5

Report No: IIIT/TR/2021/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
September 2021

SYNTACTICALLY GUIDED GENERATIVE EMBEDDINGS FOR ZERO-SHOT SKELETON ACTION RECOGNITION

Pranay Gupta, Divyanshu Sharma, Ravi Kiran Sarvadevabhatla

Center for Visual Information Technology
IIIT Hyderabad, Hyderabad 500032, INDIA.
ravi.kiran@iiit.ac.in
https://github.com/skelemoa/synse-zsl

ABSTRACT

We introduce SynSE, a novel syntactically guided generative approach for Zero-Shot Learning (ZSL). Our end-to-end approach learns progressively refined generative embedding spaces constrained within and across the involved modalities (visual, language). The inter-modal constraints are defined between action sequence embedding and embeddings of Parts of Speech (PoS) tagged words in the corresponding action description. We deploy SynSE for the task of skeleton-based action sequence recognition. Our design choices enable SynSE to generalize compositionally, i.e., recognize sequences whose action descriptions contain words not encountered during training. We also extend our approach to the more challenging Generalized Zero-Shot Learning (GZSL) problem via a confidence-based gating mechanism. We are the first to present zero-shot skeleton action recognition results on the large-scale NTU-60 and NTU-120 skeleton action datasets with multiple splits. Our results demonstrate SynSE's state of the art performance in both ZSL and GZSL settings compared to strong baselines on the NTU-60 and NTU-120 datasets.

Index Terms— ZSL, skeleton action recognition, VAE, deep learning, language and vision

1. INTRODUCTION

Advances in human action recognition have been predominantly driven by the abundance of online RGB videos. However, with the advent of accurate depth sensing technologies (e.g. Microsoft Kinect, Intel Real Sense), action recognition from 3D human skeleton data has also gained traction. Skeleton representations can be advantageous since they are compact, robustly separate the action subject (human) from background and enable privacy-preserving action capture.

The introduction of large scale skeleton action datasets such as NTU-60 [1] and NTU-120 [2], have allowed researchers to develop high-performance approaches for skeleton action recognition [3, 4, 5, 6]. However, these approaches are resource intensive, prone to overfitting and fail to generalize on classes outside the training set. Therefore, there is a strong motivation for Zero-Shot Learning (ZSL) approaches in an attempt to readily generalize across actions outside the training set.

In ZSL, visual representations and corresponding labels for *seen* classes are assumed to be available. During test time, the model is evaluated using data from *unseen* classes which are not present during training. Typically, side information (e.g. class attributes) is leveraged to transfer knowledge from the seen to unseen classes. As a popular approach, ZSL approaches employ a shared embedding strategy wherein the visual (image or video) features and semantic attribute features of the corresponding class labels are projected into a common embedding space [7, 8, 9, 10, 11, 12, 13]. Generative ZSL approaches present an alternative strategy wherein unseen samples [14] or features from unseen samples [11] are generated using Generative Adversarial Networks(GANs). Owing to the instability in training GANs, Variational Auto-Encoders(VAEs) [15, 16, 17] have also been used for feature generation.

ZSL has been previously explored for skeleton action recognition. In the only available work [18] (arXiv), embedding based methods [8, 19] are used to align visual feature embedding of skeleton action sequence with the text embedding of the descriptive action phrase (e.g 'take off jacket', 'put on glasses'). The visual features are represented by the final layer features of a skeleton action recognition model and the action phrase embedding is typically obtained by pooling the individual embeddings of words comprising the phrase. However, this approach does not enable alignment of visual embedding with respect to the individual contributors of phrase semantics - the verb ('action') and the noun(s) ('participating entities'). This inability is a major shortcoming since it does not enable generalization, i.e., being able to map the test action sequences to a description containing novel combinations of verbs and nouns, some of which might be from training action descriptions themselves.

To address these shortcomings, we propose an approach wherein the visual embedding is aligned based on the Parts of Speech (PoS) tags (verb, noun) of the phrasal words. Instead of directly mapping the visual and PoS-wise embeddings in a discriminative setting [20], we use group (per-PoS, visual) specific generative models with cross-group latent objective [17] for improved ZSL performance (Section 2). We also extend our approach to the Generalized Zero-Shot Learning (GZSL) problem, a more challenging and realistic variant of ZSL wherein good performance is required from seen *and* unseen classes. We do so by incorporating a confidence-based gating mechanism. (Section 2.5). Our approach enables state-of-the-art performance for ZSL and GZSL compared to strong baselines on the NTU-60 and the much larger NTU-120 dataset. (Section 4).

The source code and pre-trained models can be accessed at https://github.com/skelemoa/synse-zsl.

2. SYNSE

2.1. Problem definition

Let $D_{tr} = \{(x_s^{tr}, y_s^{tr})\}$ denote the set of N_{tr} training samples where x_s^{tr} denotes visual feature embedding of a skeleton action se-

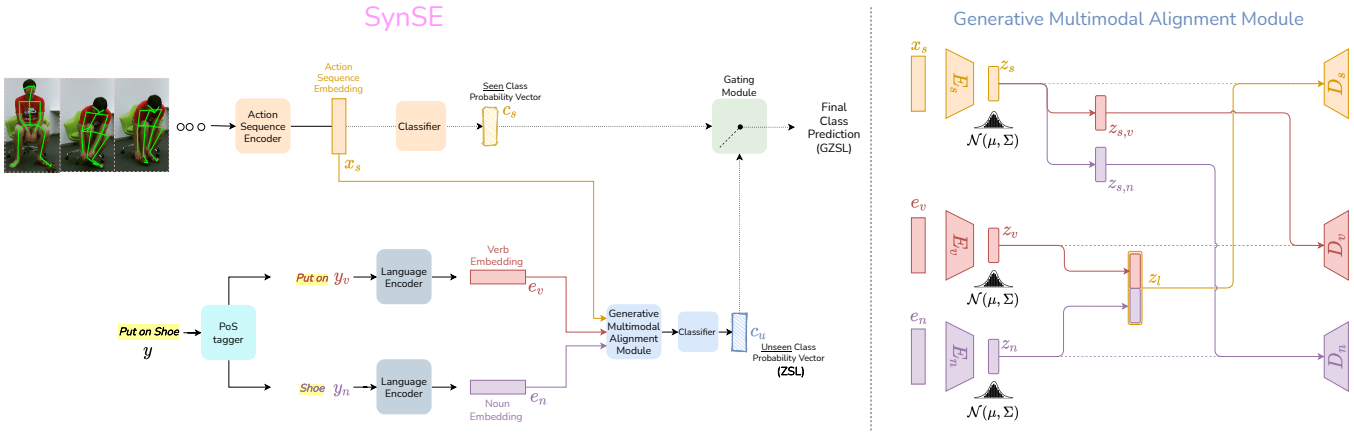

Fig. 1. Architectural diagram for our approach (SynSE). (left) The dotted path represents the process flow for the GZSL setting while the solid arrows represent the flow for ZSL. The Generative Multimodal Alignment Module is detailed on right side. It contains modality VAEs, where Part-of-Speech (PoS) specific latent generative embeddings z_v (verb), z_n (noun) are jointly aligned with segments ($z_{s,v}$, $z_{s,n}$) of latent generative skeleton embedding z_s via cross-modal alignment - refer Section 2 for more details. *Note that the RGB images have been included for reference. Only the skeleton sequence is provided as input to the network.*

quence, $y_s^{tr} \in Y_s$ is the corresponding member from the label set of seen classes. On similar lines, $D_u = \{(x_s^u, y_s^u)\}$ denotes the set of test samples with the subscript u standing for unseen. Suppose $\hat{y}$ represents the test time class prediction. For ZSL, we have $\hat{y} \in Y_u, Y_s \cap Y_u = \emptyset$ while for GZSL, we have $\hat{y} \in Y_u \cup Y_s, Y_s \cap Y_u = \emptyset$. For simplicity, we drop the subscript for seen, unseen and refer to the class names as y and the visual feature embedding as x_s.

2.2. Learning modality-wise latent generative spaces

A crucial requirement for a ZSL approach is the ability to correctly map novel inputs. For this, we employ a Variational Auto Encoder (VAE) [21] as the base architecture to learn the generative space of latent representations. A VAE is trained by maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)||p_\theta(z)) \quad (1)$$

Here, the first term on the right hand side is the reconstruction error and the second term is the Kullback-Leibler divergence between likelihood $p_\theta(z)$ and the prior $q_\phi(z|x)$. β is a hyperparameter which acts as a trade-off between the two error terms. A popular choice for the prior is the multivariate Gaussian distribution, $q_\phi(z|x) = \mathcal{N}(\mu, \Sigma)$. The VAE maps the input x initially to representations for μ, Σ and eventually to the randomized latent representation z via the reparameterization trick [21].

The first stage in our approach involves learning individual latent generative latent spaces for visual and linguistic representations. This is achieved by using a VAE for each space. To enable semantically aware compositional generalization, the text description for class label y is tokenized into constituent Part-of-Speech (PoS) specific sets - y_v for verb and y_n for noun. The tokens are encoded using a natural language encoder module to obtain the corresponding PoS-wise embeddings e_v and e_n (see Figure 1). Since our approach employs independent VAEs for skeleton (s) and linguistic (v, n) representations, the overall cost function for a single sample can be written as:

$$\mathcal{L}_{VAE} = \sum_{m \in \{s,v,n\}} \mathbb{E}_{q_\phi(z_m|x_m)}[\log p_\theta(x_m|z_m)] - \beta D_{KL}(q_\phi(z_m|x_m)||p_\theta(z_m)) \quad (2)$$

2.3. Cross-modal alignment

The VAEs optimize latent representations for individual modalities. To achieve alignment between the skeleton sequence and linguistic latent representations, a cross-modal reconstruction objective is formulated [17]. First, the latent embeddings from the PoS embeddings (z_v, z_n) are concatenated (see z_l in Figure 1) and the result x_l is used to reconstruct the visual representation via the skeleton representation VAE's posterior decoder D_s. Next, the skeleton sequence latent embedding (z_s) is uniformly mapped to as many embeddings ($z_{s,v}$, $z_{s,n}$) as the number of PoS tags. Complementary to the processing of z_l, each of the split embedding is used to reconstruct the corresponding PoS token embedding (e_n, e_v) via the corresponding PoS token embedding's decoder (D_v, D_n). Overall, the cross-modal reconstruction objective for a training sample is formulated as:

$$\mathcal{L}_{CMR} = |x_s - D_s(z_l)|_2 + \sum_{m \in \{v,n\}} |e_m - D_i(z_{s,m})|_2 \quad (3)$$

Finally, the VAE loss and the cross-modal reconstruction loss are optimized together as:

$$\mathcal{L} = \mathcal{L}_{VAE} + \alpha \mathcal{L}_{CMR} \quad (4)$$

where α is a trade-off weight factor.

2.4. Zero-shot Classification using Latent Embedding

The PoS tag embeddings of each unseen class are respectively transformed by the PoS encoders (E_v, E_n) and used to obtain samples from the latent generative space (z_l - see Figure 1). A softmax classifier $f_u : z_l \rightarrow Y_u$ is trained to classify these latent samples into the unseen classes.

Method	NTU-60		NTU-120	
	55/5 split	48/12 split	110/10 split	96/24 split
Jasani [18] (preprint)	65.53	-	-	-
ReViSE [12]	53.91	17.49	55.04	32.38
JPoSE [20]	64.82	28.75	51.93	32.44
CADA-VAE [17]	**76.84**	28.96	59.53	35.77
SynSE (ours)	75.81	**33.30**	**62.69**	**38.70**

Table 1. ZSL accuracy (%) on the NTU-60 and NTU-120 datasets.

The cross-modal VAE setup described earlier aims to align the visual features with the language features in the common latent generative space. In other words, z_s and z_l are optimized to be interchangeable. Taking advantage of this, during inference, the unseen class skeleton sequence representation x_s is first obtained. Supplying x_s to the visual VAE encoder (E_s) enables us to obtain the mean visual latent embedding (μ_s) of the sequence[1]. The corresponding class prediction is obtained using μ_s and the classifier f_u mentioned previously.

2.5. Gating Module for GZSL

For a given skeleton sequence representation x_s, the probability distribution c_s over seen classes is obtained from the skeleton action recognition model $f_s : x_s \rightarrow Y_s$ from which the action sequence embedding has been sourced all along. The unseen class classifier $f_u : E_s(x_s) \rightarrow Y_u$, is a part of our ZSL approach described in the previous section which provides the unseen class probabilities c_u. The probability distribution over *all* the classes can be written as:

$$p(y|x) = c_s p^{gate}(s; c_s, c_u) + c_u p^{gate}(u; c_s, c_u) \quad (5)$$

Further, we use a gating model (due to its superior performance for GZSL in other domains) to first decide whether the sample belongs to a seen class or an unseen class [22]. For this, the seen and unseen class probabilities are used as features to train a probabilistic binary classifier $p^{gate}(s; c_s, c_u)$ [22]. The resulting outputs are used to determine the probability distribution over all ($Y_s \cup Y_u$) classes (Equation 5).

2.6. Implementation Details

Visual and Textual features: The visual features x_s, are realised using the 256 dimensional penultimate layer feature from 4s-ShiftGCN [3], a state-of-the-art deep network for skeleton action recognition. To maintain the zero-shot assumption, we train 4s-ShiftGCN only on the seen classes. We use the Sentence BERT model [23] to obtain 1024-dimensional PoS-wise word embeddings. Before splitting into verbs and nouns, the class names are modified to fill the missing PoS tag, e.g. 'reading' is changed to 'reading book', 'drop' to 'drop object', 'headache' to 'have headache'. For actions where adding the missing tag (usually a noun) would be unreasonable (e.g. 'jump up', 'stand up'), the average of all noun embeddings is used as a placeholder.

Architectural Details: We have a single dense layer as the encoder (E_s) and decoder (D_s), which map the input features (x_s, e_v, e_n) to the latent space (z_s, z_v, z_n) and vice versa. x_s is 256-dimensional and e_v, e_n are 1024-dimensional. The size of the latent dimension is based on the number of unseen classes. For small number (5) of unseen classes, the skeleton latent dimension is set as 100 and the latent dimension for the PoS tags is 50. For larger number of unseen classes, the latent dimensions are doubled to 200 and 100 for the skeleton latent dimension and PoS tags respectively. The ZSL classifier has a single dense layer which takes latent features as input and returns the softmax probabilities for unseen classes.

Training Details: The VAEs within the Generative Multimodal Alignment Module are optimized using the Adam optimizer with a learning rate of $1e^{-4}$ and a batch size of 64. The VAEs are trained using a cyclic annealing schedule [24] for multiple cycles to mitigate the vanishing KL divergence problem. The β hyperparameter for the KL divergence is turned on after 1000 epochs, starting with 0 and is increased with a rate of 0.0021 per epoch in each cycle. Similarly, the α parameter for the cross modal reconstruction is turned on after 1400 epochs for experiments on NTU-60 and 1500 epochs on NTU-120 and is kept constant with a value of 1. One cycle is completed in 1700 epochs for NTU-60 and 1900 epochs for NTU-120 dataset. The zero-shot classifier (Section 2.4) is also optimized using Adam with a learning rate for $1e^{-3}$. 500 features per unseen class are generated and the classifier is trained for 300 epochs.

The input to the gating model (Section 2.5) is the concatenation of the top k softmax probabilities from the outputs of the seen and unseen classifiers. We set k equal to the number of unseen classes and we temperature scale [25] the seen classifier probabilities as well. The gating model is implemented as a binary logistic regression classifier and optimized using LBFGS solver from the scikit-learn library with the default aggressiveness hyperparameter ($C = 1$). For training the gating model, we set aside a few samples from the training set and refer to them as the gating train set. Similarly, we set aside a few samples from validation set (gating validation set). We train the gating model using the gating train set and determine the hyperparameters (temperature coefficient, threshold) using the gating validation set. The gating module is configured for use in 'hard' gating mode wherein $p^{gate}(s; c_s, c_u)$ and $p^{gate}(u; c_u, c_u)$ in Equation 5 take binary values [22].

3. EXPERIMENTS

3.1. Datasets

NTU-60 [1]: This is a large-scale dataset curated for 3D human action analysis. It contains 56,880 samples belonging to 60 action classes, with 40 different subjects captured from 80 distinct camera viewpoints. The action sequences of skeleton representations are in the form of 3D coordinates for 25 human body joints. We create two splits for ZSL evaluation - a 55/5 split with 55 seen classes, 5 randomly chosen unseen classes and a more challenging 48/12 split.

NTU-120 [2]: NTU-120 builds upon NTU-60 and contains 60 additional fine-grained action classes. It contains a total of 114,480 samples spread across 120 actions performed by 106 different subjects captured from 155 different camera viewpoints. Analogous to the NTU-60 ZSL evaluation setup, we create two splits - 110 (seen)/10 (unseen) and 96 (seen)/24 (unseen).

3.2. Experimental Details

We perform ZSL and GZSL experiments on the NTU-60 and NTU-120 datasets on the described splits. Since no previous works for skeleton ZSL exist, we modify representative state-of-the-art approaches from other problem domains and implement from scratch.

[1]Note that $z_s = \mu_s + \Sigma_s \odot \mathcal{E}$, where $\mathcal{E} = \mathcal{N}(0, I)$ by the VAE reparameterization trick.

Method	NTU-60 (55/5) random split			NTU-60 (48/12) random split			NTU-120 (110/10) random split			NTU-120 (96/24) random split		
	s	u	h	s	u	h	s	u	h	s	u	h
ReViSE [12]	74.23	34.73	29.22	62.36	20.77	31.16	48.69	44.84	46.68	49.66	25.06	33.31
JPoSE [20]	64.44	50.29	56.49	60.49	20.62	30.75	47.66	46.40	47.05	38.62	22.79	28.67
CADA-VAE [17]	69.38	61.79	**65.37**	51.32	27.03	35.41	47.16	49.78	48.44	41.11	34.14	37.31
SynSE	61.27	56.93	59.02	52.21	27.85	**36.33**	52.51	57.60	**54.94**	56.39	32.25	**41.04**
SynSE (+ softgating)	65.17	59.51	62.21	69.23	21.74	33.09	74.76	37.68	50.10	72.54	21.09	32.67
SynSE (- temp. scaling)	74.45	37.46	49.84	45.74	25.87	33.05	67.87	38.05	48.77	66.97	25.55	36.99
SynSE (+ CADA-VAE's GZSL)	82.70	0	0	87.63	0	0	80.43	0	0	82.46	0	0

Table 2. GZSL Accuracy (%) for seen (s) classes, unseen (u) classes and their harmonic mean (h) on NTU-60 and NTU-120 datasets

Component	Default in SynSE	Ablation	Accuracy
Language Embedding	Sentence-BERT [23]	Word2Vec [26]	60.76
Visual Features	4s-ShiftGCN [3]	MS-G3D [4]	68.80
Latent Dimension	100	50	73.83
Latent Dimension	100	200	74.67
Latent features	500	250	73.89
Latent features	500	1000	73.82
	original		**75.81**

Table 3. SynSE ZSL accuracy (%) on the NTU-60 dataset for various ablations (55/5 split).

CADA-VAE [17] learns a generative latent space under a cross aligned and distribution aligned objective. Since we found the distribution alignment objective to induce instability in training, we omit it during optimization. ReViSE [12] aims to align the latent embeddings realised via autoencoders using a Maximum Mean Discrepancy criterion. JPose [20] attempts to learn PoS aware embeddings of word2vec representations for video retrieval tasks. It learns a series of progressively refined embeddings under inter/intra modal constraints in a discriminative setting. For fair comparison, the visual features and PoS embeddings are the same as ones used in our approach (Section 2.6).

4. RESULTS

4.1. ZSL results

Table 1 shows the ZSL results of the various approaches on the NTU-60 and NTU-120 datasets. For the $55/5$ split of NTU-60, the VAE-based generative approaches significantly outperform the discriminative embedding based approaches. SynSE's performance is comparable to that of CADA-VAE. Predictably, results on the more challenging $48/12$ split show that having a larger number of unseen classes impacts performance across the board. However, SynSE offers significant improvement over other baseline approaches, including CADA-VAE. On the larger NTU-120 dataset, SynSE outperforms other methods on both the splits.

4.2. GZSL results

Since we use a gating-based strategy for GZSL in SynSE (Section 2.5), we compare against other baselines by incorporating the same strategy. Specifically, the seen class classifier is kept the same while the specific baseline approach provides the corresponding unseen class probabilities. Following standard convention for GZSL, we report the average seen class accuracy (s), the average unseen class accuracy (u) and their harmonic mean (h). Table 2 shows the results for datasets and the associated pre-defined splits. Similar to the trend in ZSL for the $55/5$ split of NTU-60, SynSE performs poorer compared to CADA-VAE on the harmonic scale for the 55/5 NTU-60 split. However, it outperforms other approaches on the harmonic scale for other splits of NTU-60 and NTU-120. We also compare our hard gating strategy [27] with the soft gating based strategy [22]. The results in Table 2 show that soft gating is biased towards seen classes, resulting in poor harmonic accuracy. Additionally, Table 2 also shows the significant performance hit when temperature scaling (Sec. 2.6) is removed [25].

To further demonstrate the effectiveness of our GZSL strategy (i.e. gating model), we explored an alternative based on the approach used for CADA-VAE [17], which does not involve gating. As Table 2 shows, the resulting setup ends up too heavily skewed for seen classes and is unable to classify the unseen classes.

4.3. Ablations

We perform ablation experiments on the 55/5 split of the NTU-60 dataset to analyse the importance of the building blocks of our alignment module and design choices affecting its ZSL performance. As the results show (Table 3), Sentence-BERT is a superior choice to Word2Vec [26] for embedding PoS-tagged words. Similarly, 4s-ShiftGCN provides better visual embeddings compared to another state-of-the-art skeleton action recognition model MS-G3D [4]. We further ablate on the architectural choices by varying the size of the latent dimension. As shown in Table 3, we see that both an increase and decrease in the size of the latent embedding causes reduction in ZSL performance. In order to validate our choice of 500 latent features per class, we experiment with varying number of latent features with results as shown in Table 3.

5. CONCLUSION

In this work, we have presented SynSE, a compositional approach for infusing latent visual representations of skeleton-based human actions with syntactic information derived from corresponding textual descriptions. We present the first set of zero-shot skeleton action recognition results on the large-scale NTU-60 and NTU-120 datasets. Our experiments show that SynSE outperforms strong baselines for ZSL and the more challenging GZSL setup. Going forward, we would like to explore the viability of SynSE for zero-shot RGB video action recognition.

6. REFERENCES

[1] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019. 1, 3

[2] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 3

[3] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 183–192. 1, 3, 4

[4] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 4

[5] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[6] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12026–12035. 1

[7] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid, "Label-embedding for image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015. 1

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129. 1

[9] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2927–2936. 1

[10] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean, "Zero-shot learning by convex combination of semantic embeddings," *arXiv preprint arXiv:1312.5650*, 2013. 1

[11] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5542–5551. 1

[12] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov, "Learning robust visual-semantic embeddings," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3571–3580. 1, 3, 4

[13] Tanmoy Mukherjee, Makoto Yamada, and Timothy M Hospedales, "Deep matching autoencoders," *arXiv preprint arXiv:1711.06047*, 2017. 1

[14] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58. 1

[15] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai, "Generalized zero-shot learning via synthesized examples," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4281–4289. 1

[16] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2188–2196. 1

[17] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata, "Generalized zero-shot learning via aligned variational autoencoders," *red*, vol. 2, pp. D2, 2019. 1, 2, 3, 4

[18] Bhavan Jasani and Afshaan Mazagonwalla, "Skeleton based zero shot action recognition in joint pose-language semantic space," *arXiv preprint arXiv:1911.11344*, 2019. 1, 3

[19] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208. 1

[20] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 4

[21] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 2

[22] Yuval Atzmon and Gal Chechik, "Adaptive confidence smoothing for generalized zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11671–11680. 3, 4

[23] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019. 3, 4

[24] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin, "Cyclical annealing schedule: A simple approach to mitigating kl vanishing," *arXiv preprint arXiv:1903.10145*, 2019. 3

[25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. 3, 4

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013. 4

[27] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng, "Zero-shot learning through cross-modal transfer," in *Advances in neural information processing systems*, 2013, pp. 935–943. 4