Supplementary Material for *SLIP: Self-supervision meets Language-Image Pre-training*

1 Full Scaling Results

We include the full results of our scaling experiments in Table 1, in which we simultaneously increase model size and training epochs. As measured by ImageNet classification accuracy under the three settings (zero-shot transfer, linear classification, and end-to-end finetuning), both large models and longer training generally improve performance.

The exception to this trend is the linear classification performance of SLIP ViT-L/16, which degrades slightly with longer training. This behavior also persists across the various other downstream benchmarks, where SLIP ViT-L/16 does worse on average when trained for 100 epochs than when trained for 25 epochs. We note that both the zero-shot transfer and end-to-end finetuning performance of SLIP ViT-L/16 improve with longer training, contrary to the behavior seen with linear classification. Thus we cannot declare this behavior to be a case of simple overfitting, as the representations are still improved for the other evaluation settings.

	0-shot			Linear			Finetuned		
Model	25	50	100	25	50	100	25	50	100
ViT-S/16	38.3	39.3	39.5	66.4	67.6	68.3	80.3	80.7	80.7
ViT-B/16	42.8	44.1	45.0	72.1	73.0	73.6	82.6	82.9	83.4
ViT-L/16	46.2	47.4	47.9	76.0	75.8	75.1	84.2	84.7	84.8

Table 1: Full scaling experiment results. SLIP pre-training scales well to larger models and more longer training as measured by zero-shot transfer, linear classification, and end-to-end finetuning, with the exception of linear classification performance using ViT-L.

2 Additional Linear Classification Benchmarks

In Table 2 we show linear classification results on all 26 downstream datasets (including ImageNet). With ViT-B and ViT-S, SLIP pre-training for 100 epochs does best. As with ImageNet, SLIP ViT-L also does worse on average when trained for 100 epochs than when trained for 25 epochs. The dataset average is 0.5 points lower for the 100 epoch model.

As expected, linear classification accuracy is much higher than zero-shot transfer accuracy (shown in Table 4 in the main paper). However, the gap between zero-shot and linear performance varies between datasets. On datasets which are straightforward vision tasks but poorly represented among the YFCC100M imagery, such as Patch Camelyon, MNIST, KITTI distance, and GTSRB, linear classification massively improves accuracy, often from a baseline of around chance performance. On datasets which share more overlap with YFCC100M, such as Food-101, Caltech-101, and Caltech-UCSD Birds 2011, we see significant improvements as well.

However, with HatefulMemes and Rendered SST2, two datasets which require OCR capabilities, the linear classification performance of all models is still around chance. These results suggest, perhaps unsurprisingly, that zero-shot transfer results are much more dependent on what visual and semantic concepts were seen during training than linear classification, since they do not enjoy the benefit of further training examples. We also note that relative rankings within each model size are also quite unstable where the best results alternate between the 25 and 100 epoch models. This is very similar to what we see in the zero-shot transfer evaluations, as discussed in Section 5 in the main paper.

		Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2	ImageNet	Average
ViT-S	CLIP	71.1	82.5	63.2	66.5	70.7	28.2	25.3	61.6	63.4	75.6	92.4	89.6	45.2	92.1	92.4	83.6	65.7	63.6	22.0	80.4	67.7	43.2	44.3	54.6	51.0	59.3	63.7
	SimCLR (25 ep)	64.3	77.9	51.5	27.0	69.0	14.3	16.1	61.0	46.6	59.5	34.5	83.4	47.0	87.4	90.4	83.0	52.6	60.3	13.3	82.4	70.4	44.0	43.5	55.7	53.4	58.1	55.6
	SLIP (25 ep)	77.4	80.7	63.5	**67.0**	**74.2**	39.3	**31.5**	70.7	68.9	84.5	**95.0**	**91.3**	**52.7**	95.7	94.3	90.5	**68.2**	65.8	22.4	81.9	76.9	50.8	**51.6**	**59.4**	**54.3**	66.4	68.3
	SLIP (100 ep)	**78.7**	**84.1**	**66.3**	66.0	73.9	**40.6**	30.7	**71.6**	**71.3**	**85.7**	94.8	90.7	50.4	**96.4**	**95.2**	**89.0**	**68.2**	**66.8**	**23.3**	**82.9**	**77.7**	**52.2**	50.5	56.3	53.4	**68.3**	**68.7**
ViT-B	CLIP	77.6	86.2	70.7	70.9	73.7	41.8	29.5	66.0	68.2	82.0	94.3	93.8	49.7	94.9	94.5	88.3	72.5	65.8	24.9	82.9	72.6	47.9	48.7	55.2	54.8	66.5	68.2
	SimCLR (25 ep)	73.0	82.6	63.3	44.7	71.9	32.0	26.3	69.6	62.4	76.8	85.6	91.6	49.5	92.9	93.2	89.8	65.6	65.0	16.0	84.9	73.0	50.1	52.6	57.5	53.9	64.0	64.9
	SLIP (25 ep)	83.0	87.7	**71.6**	70.9	76.3	47.4	**34.4**	73.9	73.1	88.1	**96.1**	**94.5**	53.5	97.4	**95.9**	**92.8**	75.5	**68.6**	25.1	**84.4**	80.4	55.0	**54.2**	56.8	55.0	72.1	71.7
	SLIP (100 ep)	**83.1**	**88.9**	71.5	**72.0**	**76.4**	**49.0**	33.9	**75.9**	**75.1**	**89.1**	92.4	**94.5**	**54.4**	**98.2**	95.5	92.0	**75.9**	67.9	**25.6**	83.0	**82.0**	**55.6**	53.9	**59.9**	**56.1**	**73.6**	**72.1**
ViT-L	CLIP	81.8	91.2	75.1	75.1	75.4	46.9	33.0	66.2	72.0	84.2	95.9	95.7	54.7	96.5	95.1	90.6	76.4	68.7	27.2	83.6	75.9	51.4	51.9	**59.7**	53.8	70.5	71.1
	SimCLR (25 ep)	73.6	89.7	68.3	32.5	73.5	18.9	17.1	66.0	55.6	69.8	70.1	90.1	48.9	93.1	90.7	86.7	56.0	61.8	17.2	**85.5**	69.8	51.3	48.6	57.2	54.3	66.7	62.0
	SLIP (25 ep)	**86.5**	**92.9**	**77.2**	**76.6**	**78.0**	**52.0**	**37.4**	75.2	78.3	**90.8**	**97.6**	**95.8**	56.8	**98.7**	96.6	**93.3**	**79.9**	68.9	**29.1**	84.8	83.0	**59.0**	**57.5**	57.0	54.5	**76.0**	**74.4**
	SLIP (100 ep)	84.1	91.0	74.5	72.5	76.7	51.1	37.2	**77.0**	**79.9**	90.5	97.5	**95.8**	**60.2**	**98.7**	**97.0**	93.2	75.6	**70.6**	28.2	85.2	**83.1**	56.7	55.0	55.4	**56.2**	75.1	73.8

Table 2: Linear classification evaluation with ViT S, B, and L on a variety of classification benchmarks. Best results in bold. SLIP outperforms CLIP and SimCLR on most of the tasks, frequently with a significant margin.

As expected, linear classification accuracy is higher than zero-shot transfer accuracy. However, the gap between zero-shot and linear performance varies between datasets. On the Patch Camelyon dataset, linear classification brings the performance of SLIP ViT-L from 50.6% when evaluated under zero-shot transfer to 85.2%, a huge improvement. With most other datasets, the gap between linear classification and zero-shot transfer performance is significant but less dramatic. And on HatefulMemes and Rendered SST2, two datasets which require OCR capabilities, the linear classification performance of all models is still close to chance. We note that relative rankings within each model size are still somewhat unstable, similar to the zero-shot transfer evaluations, where the best results alternate between the 25 and 100 epoch models.

3 Additional Implementation Details

Datasets. YFCC15M [23] [29] contains raw HTML captions and titles which we lightly preprocess before training. We unescape the HTML then remove HTML tags and urls with simple regex matching.

CC3M [25] is collected from an initial set of 5B candidate images, of which 99.9% are filtered out according to simple image and text heuristics for quality and content. Many of these filters are relaxed by CC12M [6] in order to collect a bigger and potentially noisier dataset. CC3M also hypernymizes proper nouns, numbers, and infrequent entities to make the dataset more amenable to training and evaluating image captioning systems, the original design for the dataset. In contrast, CC12M only replaces person names for privacy. Our versions of these datasets contain 3.1M and 11.0 M images respectively, due to asset removal.

Pre-training. During pre-training we use a cosine learning rate decay schedule with 1 epoch (~3500 iterations) of linear warmup when training on YFCC15M. When pre-training for 100 epochs we use 2 warmup epochs. On YFCC15M (14.6M images), we train for 25 epochs and on CC12M (11.0M images) we train for 35 epochs. This amounts to approximately the same number of iterations as 300 epochs on ImageNet-1K [24]. Due to the smaller size of CC3M (3.1M images), we train for 40 epochs to reduce overfitting. We trained on up to sixteen 8× V100-32GB servers, and to fit SLIP ViT-Large/16 in memory we accumulated gradients over two steps.

End-to-end Finetuning. We use a similar training recipe for finetuning all models on ImageNet based on the ImageNet finetuning recipe from BeiT [1] using AdamW and a batch size of 1024 with learning rate of 4e-3 and weight decay of 0.05, along with various data augmentations and regularization methods. As we increase model size we also increase regularization. For ViT-S we set drop path to 0 and layer decay to 0.65, for ViT-B we set drop path to 0.1 and layer decay to 0.65, and for ViT-L we set drop path to 0.

4 Algorithm Pseudocode

Algorithm 1 SLIP-SimCLR: PyTorch-like Pseudocode

```
# fi, ft: image, text encoders
# hi, ht: CLIP image, text projectors
# hs: SimCLR projector
# c: SimCLR loss scale
def forward(img, text):
    xi, x1, x2 = crop(img), aug(img), aug(img)
    yt = tokenize(text)

    wi, w1, w2 = fi(xi, x1, x2)
    wt = ft(yt)

    z1, z2 = hs(w1), hs(w2) # SSL embed: N x C2
    zi, zt = hi(wi), ht(wt) # CLIP embed: N x C1

    loss = c * simclr(z1, z2) + clip(zi, zt)
    return loss

# s: learnable log logit scale
def clip(zi, zt):
    zi, zt = normalize(zi, zt)
    label = range(N)
    logit = exp(s) * zi @ zt.T

    li = CrossEntropy(logit, label)
    lt = CrossEntropy(logit.T, label)

    loss = (li + lt) / 2
    return loss

# tau: softmax temperature
def simclr(z1, z2):
    z1, z2 = normalize(z1, z2)
    label = range(N)
    mask = eye(N) * 1e9

    logit = z1 @ z2.T
    logit1 = z1 @ z1.T - mask
    logit2 = z2 @ z2.T - mask

    logit1 = cat(logit, logit1)
    logit2 = cat(logit.T, logit2)

    l1 = CrossEntropy(logit1 / tau)
    l2 = CrossEntropy(logit2 / tau)

    loss = (l1 + l2) / 2
    return loss
```

Notes: `@` is the matrix multiplication operator. `k.T` is `k`'s transpose. `eye` constructs an identity matrix. `cat` concatenates two matrices.

5 Nearest neighbor visualizations

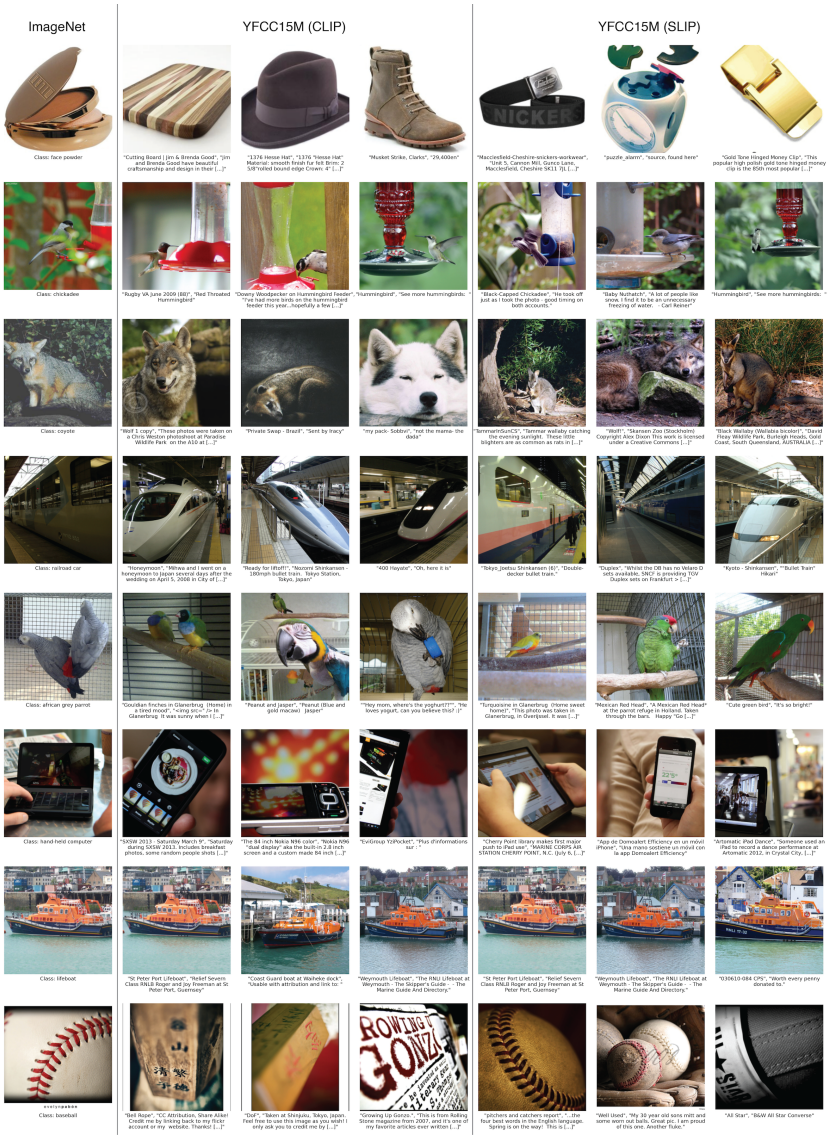

Figure 1: Visualization of 3 nearest neighbors for randomly sampled ImageNet validation images (left), from YFCC15M for CLIP (middle) and SLIP (right) ViT-B/16 trained on YFCC15M. For each model, the 1st to 3rd neighbor is shown from left to right.

6 Dataset Infosheet

Dataset	Metric	Chance performance	Description
Food-101 [4]	acc	1.0	101 categories of food dishes
CIFAR-10 [17]	acc	10.0	10 categories of animals and vehicles
CIFAR-100 [17]	acc	1.0	100 categories of animals, vehicles, plants, objects, scenes, people
CUB-200-2011 [31]	acc	0.8	200 species of mostly North American birds
SUN397 [32]	acc	2.2	397 categories of various indoor and outdoor scenes
Stanford Cars [16]	acc	0.8	196 categories of cars (make, model, and year)
FGVC Aircraft [19]	mean per class	1.0	102 categories of aircraft (manufacturer, family, and variant)
Describable Textures [8]	acc	2.1	47 categories of texture patches
Oxford Pets [21]	mean per class	2.7	37 breeds of cats and dogs
Caltech-101 [10]	mean per class	5.2	101 categories of objects
Oxford Flowers [20]	mean per class	1.5	102 species of common UK flowers
MNIST [18]	acc	10.0	10 categories of handwritten digits
FER-2013 [12]	acc	24.7	7 categories of human facial emotions
STL-10 [9]	acc	11.4	10 categories of animals and vehicles
EuroSat [13]	acc	10.0	10 categories of land from satellite imagery
RESISC45 [7]	acc	2.2	45 categories of land from satellite imagery and aerial photography
GTSRB [28]	acc	5.9	43 categories of German traffic signs
KITTI Distance [11]	acc	31.0	4 categories of traffic scenes with nearby cars in varying positions
Country211 [22] [29]	acc	0.5	211 countries represented by geo-tagged images
Patch Camelyon [30] [2]	acc	50.0	2 classes of metastatic or benign lymph node slide patches
UCF101 Frames [27]	acc	1.3	101 categories of human actions using the middle frame of each clip
Kinetics 700 Frames [5]	mean(acc1, acc5)	0.4	700 categories of human actions using the middle frame of each clip
Clevr Counts [14]	acc	12.9	8 categories of rendered scenes with varying numbers of objects
Hateful Memes [15]	ROC AUC	50.0	2 categories of hateful or not hateful image macros
Rendered SST2 [22] [26]	acc	50.1	2 classes of positive or negative movie reviews rendered as text
ImageNet [24]	acc	0.1	1000 categories of objects

Table 3: Info sheet for classification datasets. Chance performance is computed by assuming random predictions of the labels in proportion to their frequency in the test set.

7 Ethical considerations.

SLIP faces all of the same ethical considerations as CLIP, both in terms of the harmful applications it may enable, as well as the potential for amplifying and perpetuating problematic behavior in the real world. CLIP's ability to leverage noisy and minimally filtered data scraped from the open internet has already spurred researchers to begin collecting data in a more careless manner than previously possible for supervised learning [3]. A more cautious and responsible approach to selecting training data may alleviate the most egregious model behaviors.

8 Practical limitations.

SLIP computes embeddings of image views for both the self-supervised objective and the CLIP objective. This increases the activation count and memory footprint of the model during the forward pass, which results in slower training (30.5 hours for SLIP vs 22.3 hours for CLIP to train ViT-B/16 on 64 V100 GPUs).

After pre-training, SLIP incurs no additional cost since its vision backbone can be used by itself.

From the downstream zero-shot results, we note that pre-training on uncurated data alone appears to be an inefficient route to recognizing specific visual concepts, especially concepts unlikely to be widely shared on social media or the broader internet. Even with a massive amount of curated data, CLIP's zero-shot performance on many datasets is still far below what can easily be achieved by finetuning a smaller pre-trained model on a modest amount of labeled data. This can be addressed by finetuning CLIP for specific applications or including more pre-training data from the domain of interest.

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. ArXiv **abs/2106.08254** (2021)
2. Bejnordi, B.E., Veta, M., van Diest, P.J., van Ginneken, B., Karssemeijer, N., Litjens, G.J.S., van der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M.C.A., Geessink, O.G.F., Stathonikos, N., van Dijk, M.C., Bult, P., Beca, F., Beck, A.H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H., Heng, P.A., Hass, C., Bruni, E., Wong, Q., Halici, U., Öner, M.Ü., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A.I., Kraus, O.Z., Shaban, M., Rajpoot, N.M., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Phoulady, H.A., Kovalev, V.A., Kalinovsky, A., Liauchuk, V., Bueno, G., del Milagro Fernández-Carrobles, M., Serrano, I., Deniz, O., Racoceanu, D., Venâncio, R.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**, 2199–2210 (2017)
3. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes. ArXiv **abs/2110.01963** (2021)
4. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision (2014)
5. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. ArXiv **abs/1907.06987** (2019)
6. Changpinyo, S., Sharma, P.K., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3557–3567 (2021)
7. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**, 1865–1883 (2017)
8. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014)
9. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: AISTATS (2011)
10. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**, 594–611 (2006)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition pp. 3354–3361 (2012)
12. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A.C., Mirza, M., Hamner, B., Cukierski, W.J., Tang, Y., Thaler, D., Lee, D.H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R.T., Popescu, M.C., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Zhang, C., Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. Neural networks : the official journal of the International Neural Network Society **64**, 59–63 (2013)
13. Helber, P., Bischke, B., Dengel, A.R., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**, 2217–2226 (2019)

14. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1988–1997 (2017)
15. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. ArXiv **abs/2005.04790** (2020)
16. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013)
17. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)
18. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition (1998)
19. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. Tech. rep. (2013)
20. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing pp. 722–729 (2008)
21. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. 2012 IEEE Conference on Computer Vision and Pattern Recognition pp. 3498–3505 (2012)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
23. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**, 211–252 (2015)
25. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
26. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP (2013)
27. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. ArXiv **abs/1212.0402** (2012)
28. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The german traffic sign recognition benchmark: A multi-class classification competition. The 2011 International Joint Conference on Neural Networks pp. 1453–1460 (2011)
29. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K.S., Poland, D.N., Borth, D., Li, L.J.: Yfcc100m: the new data in multimedia research. Commun. ACM **59**, 64–73 (2016)
30. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology (Jun 2018)
31. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)

32. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 3485–3492 (2010)